Machine Learning Theory
Lecture 13

Nicholas Harvey

March 11, 2022

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. In these notes, we consider methods for solving $\inf_{x \in \mathcal{X}} f(x)$. In general, this infimum is not achieved, but for simplicity we will assume throughout that it is achieved by at a point x^*. (Otherwise, one could assume that x^* is an approximate minimizer, since our algorithms provide only approximate solutions anyways.)

1 Lipschitz functions

1.1 The basic setting

We begin with the most basic setting, in which f is L-Lipschitz with respect to the Euclidean norm. Since f is convex, we have $\partial f(x) \neq \emptyset$ for all $x \in \mathbb{R}^n$. The algorithm is shown in Algorithm 1.

Algorithm 1 Gradient descent for minimizing a convex, 1-Lipschitz function over $\mathbb{R}^n$.

```
1: procedure GradientDescent($x_1 \in \mathbb{R}^n,\ T \in \mathbb{N}$)
2:     Let $\eta = 1/\sqrt{T}$
3:     for $i \leftarrow 1, ..., T-1$ do
4:         $x_{i+1} \leftarrow x_i - \eta g_i$, where $g_i = \nabla f(x_i)$ if $f$ is differentiable, and otherwise $g_i$ is any subgradient in $\partial f(x_i)$.
5:     return $\sum_{i=1}^{T} x_i / T$
```

Theorem 1.1. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and 1-Lipschitz (with respect to $\|\cdot\|_2$). Fix an optimal solution $x^* \in \operatorname{argmin}_x f(x)$ and a starting point $x_1 \in \mathbb{R}^n$. Define $\eta = \frac{1}{\sqrt{T}}$. Suppose that $\|x_1 - x^*\|_2 \leq 1$. Then

$$f\left(\frac{1}{T}\sum_{i=1}^{T} x_i\right) - f(x^*) \ \leq \ \frac{1}{\sqrt{T}}.$$

Proof. We bound the error on the i^{th} iteration as follows:

$$\begin{aligned} f(x_i) - f(x^*) \ &\leq \ \langle\, g_i,\, x_i - x^* \,\rangle \qquad \text{(by the subgradient inequality (3.2))} \\ &= \ \frac{1}{\eta}\langle\, x_i - x_{i+1},\, x_i - x^* \,\rangle \qquad \text{(by the gradient step in line 4)} \end{aligned}$$

$$= \frac{1}{2\eta}\Big(\|x_i - x_{i+1}\|_2^2 + \|x_i - x^*\|_2^2 - \|x_{i+1} - x^*\|_2^2 \Big) \qquad \text{(by the cosine law (3.1)).}$$

To analyze the average error, sum the previous displayed equation over i. The last two terms telescope, yielding

$$\begin{aligned}
\sum_{i=1}^{T} \big(f(x_i) - f(x^*)\big) &\leq \frac{1}{2\eta}\Bigg(\Big(\sum_{i=1}^{T} \|x_i - x_{i+1}\|_2^2 \Big) + \|x_1 - x^*\|_2^2 - \|x_{T+1} - x^*\|_2^2 \Bigg) \\
&\leq \frac{1}{2\eta}\Big(\sum_{i=1}^{T} \|\eta g_i\|_2^2 + \|x_1 - x^*\|_2^2 \Big) \quad \text{(by the gradient step in line 4)} \\
&\leq \frac{\eta T}{2} + \frac{1}{2\eta} \qquad \text{(by (3.4) and assumption on } x_1\text{)}
\end{aligned}$$

Dividing by T and using Jensen's inequality (Lemma 3.8) and the definition of η gives

$$f\left(\sum_{i=1}^{T} \frac{x_i}{T} \right) - f(x^*) \;\leq\; \sum_{i=1}^{T} \frac{1}{T}\big(f(x_i) - f(x^*)\big) \;\leq\; \frac{\eta}{2} + \frac{1}{2T\eta} \;=\; \frac{1}{\sqrt{T}},$$

as required. □

Remark. Theorem 1.1 achieves the optimal rate for any algorithm that only accesses f using a subgradient oracle [2, Theorem 3.13].

Remark. Thinking ahead to future topics, let us observe a troubling aspect of this algorithm. We may think of $\mathbb{R}^n$ as an abstract vector space $\mathcal{V}$. The gradient $\nabla f(x_i)$ then lives in the dual space $\mathcal{V}^*$, whereas the iterates x_i lie in the primal space $\mathcal{V}$. Nevertheless, the algorithm performs arithmetic between these objects lying in different spaces. If we think of gradients as row vectors, then we are implicitly using the transpose operation to map from the dual to the primal space.

General reduction from arbitrary scale & Lipschitz value. The analysis present above assumes that the given function f is 1-Lipschitz. How shall we handle a function that is L-Lipschitz? It also has a certain "scale assumption" $\|x_1 - x^*\|_2 \leq 1$. How could we handle a general scale, say $\|x_1 - x^*\|_2 \leq R$? In Section 1.7 we will discuss a general reduction that can handle such scenarios.

Theorem 1.2. Suppose that we have a theorem giving a convergence rate guarantee $c(T)$ for gradient descent assuming f is 1-Lipschitz and assuming the "scale" $\|x_1 - x^*\|_2 \leq 1$. Suppose h is an L-Lipschitz function whose "scale" is bounded by R. Then there is a black-box reduction from h to f, showing that gradient descent on h achieves convergence rate $RL \cdot c(T)$.

1.2 The constrained setting

In this section we consider the problem $\min_{x \in \mathcal{X}} f(x)$ where $\mathcal{X}$ is a closed, convex set. Again, f is assumed to be convex and 1-Lipschitz.

The ordinary gradient descent algorithm does not ensure that the iterates remain in $\mathcal{X}$. In this section we modify the algorithm to project back onto $\mathcal{X}$. The algorithm now takes a gradient step from the iterate x_i to compute a new point y_{i+1}, then projects onto $\mathcal{X}$ to obtain the new iterate x_{i+1}.

Algorithm 2 Projected gradient descent for minimizing convex, 1-Lipschitz functions over a convex set.

```
1: procedure ProjectedGradientDescent($\mathcal{X} \subseteq \mathbb{R}^n$, $x_1 \in \mathcal{X}$, $T \in \mathbb{N}$)
2:     Let $\eta = 1/\sqrt{T}$
3:     for $i \leftarrow 1, ..., T-1$ do
4:         $y_{i+1} \leftarrow x_i - \eta g_i$, where $g_i \in \partial f(x_i)$.
5:         $x_{i+1} \leftarrow \Pi_{\mathcal{X}}(y_{i+1})$
6:     return $\sum_{i=1}^{T} x_i/T$
```

The algorithm, shown in Algorithm 2, is a slight modification of Algorithm 1. The theorem is a slight modification of Theorem 1.1. The only changes are highlighted below.

Theorem 1.3. Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set. Suppose that $f : \mathcal{X} \to \mathbb{R}$ is convex and 1-Lipschitz (with respect to $\|\cdot\|_2$). Fix an optimal solution $x^* \in \operatorname{argmin}_x f(x)$ and a starting point $x_1 \in \mathcal{X}$. Define $\eta = \frac{1}{\sqrt{T}}$. Suppose that $\|x_1 - x^*\|_2 \leq 1$. Then

$$f\left(\frac{1}{T}\sum_{i=1}^{T} x_i\right) - f(x^*) \;\leq\; \frac{1}{\sqrt{T}}.$$

Proof. We bound the error on the i^{th} iteration as follows:

$$\begin{aligned}
f(x_i) - f(x^*) &\leq \langle\, g_i,\, x_i - x^* \,\rangle \qquad \text{(by the subgradient inequality (3.2))}\\
&= \frac{1}{\eta}\langle\, x_i - y_{i+1},\, x_i - x^* \,\rangle \qquad \text{(by the gradient step in line 4)}\\
&= \frac{1}{2\eta}\Big(\big\|x_i - y_{i+1}\big\|_2^2 + \|x_i - x^*\|_2^2 - \big\|y_{i+1} - x^*\big\|_2^2 \Big) \qquad \text{(by the cosine law (3.1))}\\
&\leq \frac{1}{2\eta}\Big(\big\|x_i - y_{i+1}\big\|_2^2 + \|x_i - x^*\|_2^2 - \|x_{i+1} - x^*\|_2^2 \Big).
\end{aligned}$$

The last line uses Claim 3.4: since x_{i+1} is the projected point $\Pi_{\mathcal{X}}(y_{i+1})$ and $x^* \in \mathcal{X}$, the corollary yields that $\|x_{i+1} - x^*\|_2^2 \leq \|y_{i+1} - x^*\|_2^2$. To analyze the average error, sum the previous displayed equation over i. The last two terms telescope, yielding

$$\begin{aligned}
\sum_{i=1}^{T} \big(f(x_i) - f(x^*)\big) &\leq \frac{1}{2\eta}\Bigg(\Big(\sum_{i=1}^{T} \big\|x_i - y_{i+1}\big\|_2^2 \Big) + \|x_1 - x^*\|_2^2 - \|x_{T+1} - x^*\|_2^2 \Bigg)\\
&\leq \frac{1}{2\eta}\Big(\sum_{i=1}^{T} \|\eta g_i\|_2^2 + \|x_1 - x^*\|_2^2 \Big) \qquad \text{(by the gradient step in line 4)}\\
&\leq \frac{\eta T}{2} + \frac{1}{2\eta} \qquad \text{(by (3.4) and the assumption on } x_1\text{)}
\end{aligned}$$

Dividing by T and using Jensen's inequality (Lemma 3.8) and the definition of η gives

$$f\left(\sum_{i=1}^{T}\frac{x_i}{T}\right) - f(x^*) \;\leq\; \sum_{i=1}^{T}\frac{1}{T}\big(f(x_i) - f(x^*)\big) \;\leq\; \frac{\eta}{2} + \frac{1}{2T\eta} \;=\; \frac{1}{\sqrt{T}},$$

as required. □

1.3 Online setting

Suppose that at time step i, the algorithm proposes a point x_i, the adversary chooses a function f_i, and the algorithm receives a subgradient $g_i \in \partial f_i(x_i)$. The algorithm's cost of this iteration is $f_i(x_i)$. The goal is to minimize the ***regret*** (or ***total regret***), which is defined to be

$$\operatorname{Regret}(T) \;:=\; \sum_{i=1}^{T} f_i(x_i) - \min_{x^* \in \mathcal{X}} \sum_{i=1}^{T} f_i(x^*).$$

This is the algorithm's cost minus the cost of the best *fixed* point.

The projected gradient descent algorithm (Algorithm 2) works in this setting with only trivial changes: replacing f with f_i throughout. The modified algorithm is shown in Algorithm 3. The theorem is a slight modification of Theorem 1.3.

Algorithm 3 Online projected gradient descent for Lipschitz functions.

```
1: procedure OnlineProjectedGradientDescent($\mathcal{X} \subseteq \mathbb{R}^n,\ x_1 \in \mathcal{X},\ T \in \mathbb{N}$)
2:     Let $\eta = 1/\sqrt{T}$
3:     for $i \leftarrow 1, \ldots, T-1$ do
           ▷ Incur cost $f_i(x_i)$, receive a subgradient $g_i \in \partial f_i(x_i)$
4:         $y_{i+1} \leftarrow x_i - \eta g_i$
5:         $x_{i+1} \leftarrow \Pi_{\mathcal{X}}(y_{i+1})$
```

Theorem 1.4. Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set. Suppose that $f_1, f_2, \ldots : \mathcal{X} \to \mathbb{R}$ are convex and 1-Lipschitz (with respect to $\|\cdot\|_2$). Fix a starting point $x_1 \in \mathcal{X}$. Define $\eta = \frac{1}{\sqrt{T}}$. Suppose that $\|x_1 - x^*\|_2 \leq 1$. Then the regret satisfies

$$\operatorname{Regret}(T) \;=\; \sum_{i=1}^{T}\big(f_i(x_i) - f_i(x^*)\big) \;\leq\; \sqrt{T}.$$

Proof. We bound the error on the i^{th} iteration as follows:

$$\begin{aligned}
f_i(x_i) - f_i(x^*) &\leq \langle\, g_i,\, x_i - x^* \,\rangle && \text{(by the subgradient inequality (3.2))}\\
&= \frac{1}{\eta}\langle\, x_i - y_{i+1},\, x_i - x^* \,\rangle && \text{(by the gradient step in line 4)}\\
&= \frac{1}{2\eta}\Big(\|x_i - y_{i+1}\|_2^2 + \|x_i - x^*\|_2^2 - \|y_{i+1} - x^*\|_2^2\Big) && \text{(by the cosine law (3.1))}\\
&\leq \frac{1}{2\eta}\Big(\|x_i - y_{i+1}\|_2^2 + \|x_i - x^*\|_2^2 - \|x_{i+1} - x^*\|_2^2\Big).
\end{aligned}$$

The last line uses Claim 3.4: since x_{i+1} is the projected point $\Pi_{\mathcal{X}}(y_{i+1})$ and $x^* \in \mathcal{X}$, the corollary yields that $\|x_{i+1} - x^*\|_2^2 \leq \|y_{i+1} - x^*\|_2^2$. To analyze the average error, sum the previous displayed equation over i. The last two terms telescope, yielding

$$\begin{aligned}
\sum_{i=1}^{T} \big(f_i(x_i) - f_i(x^*) \big) &\leq \frac{1}{2\eta} \Bigg(\Big(\sum_{i=1}^{T} \|x_i - y_{i+1}\|_2^2 \Big) + \|x_1 - x^*\|_2^2 - \|x_{T+1} - x^*\|_2^2 \Bigg) \\
&\leq \frac{1}{2\eta} \Big(\sum_{i=1}^{T} \|\eta g_i\|_2^2 + \|x_1 - x^*\|_2^2 \Big) \qquad \text{(by the gradient step in line 4)} \\
&\leq \frac{\eta T}{2} + \frac{1}{2\eta} \qquad \text{(by (3.4) and definition of } R\text{)}
\end{aligned}$$

Jensen's inequality is not needed here, as we wish to bound the regret: the total error of the iterates. Substituting η completes the proof of the regret bound. □

1.4 Unknown time horizon (with diameter bound)

A disadvantage of the preceding algorithms is that they require the step size η to be chosen with knowledge of T, the iteration at which a good approximation is desired. What if T is not known at the time that gradient descent starts executing? It is possible to make GD oblivious to the value of T by allowing the step size to depend on the iteration.

Let us illustrate this technique by analyzing the online, projected setting. This result requires a slightly stronger hypothesis: a bound on the *diameter* of $\mathcal{X}$. The proof is very similar to the proof of Theorem 1.4.

Algorithm 4 Online projected gradient descent for Lipschitz functions, with unknown time horizon.

```
1: procedure OnlineProjectedGradientDescent($\mathcal{X} \subseteq \mathbb{R}^n,\ x_1 \in \mathcal{X}$)
2:     Let $\eta_i = 1/\sqrt{i}$ for all $i \in \mathbb{N}$
3:     $i \leftarrow 1$
4:     repeat
           ▷ Incur cost $f_i(x_i)$, receive a subgradient $g_i \in \partial f_i(x_i)$
5:         $y_{i+1} \leftarrow x_i - \eta_i g_i$
6:         $x_{i+1} \leftarrow \Pi_{\mathcal{X}}(y_{i+1})$
7:         $i \leftarrow i + 1$
8:     until solution desired in iteration $T$
```

Theorem 1.5. Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set. Suppose that $f_1, f_2, \ldots : \mathcal{X} \to \mathbb{R}$ are convex and 1-Lipschitz (with respect to $\|\cdot\|_2$). Let $\eta_i = \frac{1}{\sqrt{2i}}$. Suppose that $\operatorname{diam}(\mathcal{X}) \leq 1$. Then the regret satisfies

$$\operatorname{Regret}(T) \; = \; \sum_{i=1}^{T} \big(f_i(x_i) - f_i(x^*) \big) \; \leq \; \sqrt{2T} \qquad \forall T \geq 1.$$

Consequently, in the offline setting where each $f_i = f$,

$$f\left(\frac{1}{T}\sum_{i=1}^{T} x_i\right) - f(x^*) \;\leq\; \sqrt{\frac{2}{T}} \qquad \forall T \geq 1.$$

References. This result is originally due to [8, Theorem 1], with slightly worse parameters. See also [3, Theorem 3.1].

Proof. We bound the error on the i^{th} iteration as follows:

$$\begin{aligned}
f_i(x_i) - f_i(x^*) \;&\leq\; \langle\, g_i,\, x_i - x^* \,\rangle \qquad \text{(by the subgradient inequality (3.2))}\\
&= \frac{1}{\eta_i}\langle\, x_i - y_{i+1},\, x_i - x^* \,\rangle \qquad \text{(by the gradient step in line 5)}\\
&= \frac{1}{2\,\eta_i}\Big(\|x_i - y_{i+1}\|_2^2 + \|x_i - x^*\|_2^2 - \|y_{i+1} - x^*\|_2^2 \Big) \qquad \text{(by the cosine law (3.1))}\\
&\leq \frac{1}{2\,\eta_i}\Big(\|x_i - y_{i+1}\|_2^2 + \|x_i - x^*\|_2^2 - \|x_{i+1} - x^*\|_2^2 \Big).
\end{aligned}$$

The last line uses Claim 3.4: since x_{i+1} is the projected point $\Pi_{\mathcal{X}}(y_{i+1})$ and $x^* \in \mathcal{X}$, the corollary yields that $\|x_{i+1} - x^*\|_2^2 \leq \|y_{i+1} - x^*\|_2^2$. To analyze the average error, sum the previous displayed equation over i. The last two terms no longer telescope, but nearly do, and this is enough to get a good upper bound:

$$\begin{aligned}
&\sum_{i=1}^{T}\big(f_i(x_i) - f_i(x^*)\big)\\
&\leq \frac{1}{2}\left(\sum_{i=1}^{T} \frac{\|x_i - y_{i+1}\|_2^2}{\eta_i} + \frac{\|x_1 - x^*\|_2^2}{\eta_1} + \sum_{i=2}^{T}\left(\frac{1}{\eta_i} - \frac{1}{\eta_{i-1}}\right)\|x_i - x^*\|_2^2 - \frac{\|x_{T+1} - x^*\|_2^2}{\eta_{T+1}} \right)\\
&\leq \frac{1}{2}\left(\sum_{i=1}^{T} \frac{\big\|\,\eta_i\, g_i\big\|_2^2}{\eta_i} + \sqrt{2}\sum_{i=1}^{T}\left(\sqrt{i} - \sqrt{i-1}\right) \right) \qquad \text{(by definition of } \eta_i \text{ and the diameter bound)}\\
&= \frac{1}{2}\left(\sum_{i=1}^{T} \frac{1}{\sqrt{2i}} + \sqrt{2T} \right) \qquad \text{(by (3.4) and telescoping)}\\
&\leq \frac{1}{2}\left(\frac{2\sqrt{T}}{\sqrt{2}} + \sqrt{2T} \right) \qquad \text{(by Claim 3.2)}\\
&\leq \sqrt{2T}.
\end{aligned}$$

This completes the proof of the regret bound.

In the offline setting, we divide by T and use Jensen's inequality (Lemma 3.8), obtaining

$$f\left(\sum_{i=1}^{T} \frac{x_i}{T}\right) - f(x^*) \;\leq\; \sum_{i=1}^{T} \frac{1}{T}\big(f(x_i) - f(x^*)\big) \;\leq\; \sqrt{2T}.$$

□

1.4.1 Removing the diameter bound

Theorem 1.5 is nice in that it allows decreasing step sizes. However, unlike our previous theorems, it requires a *diameter bound* on $\mathcal{X}$ instead of simply bounding the distance $\|x_1 - x^*\|$. This additional assumption is distasteful and, as it turns out, unnecessary. In this section we remove this assumption by introducing a ***stabilization trick***, in which each iterate x_i is always mixed with a certain fraction of the starting point x_1.

Algorithm 5 Stabilized online projected gradient descent for Lipschitz functions.

```
procedure StabilizedGradientDescent($\mathcal{X} \subseteq \mathbb{R}^n,\ x_1 \in \mathcal{X},\ \eta : \mathbb{N} \to \mathbb{R},\ \gamma : \mathbb{N} \to \mathbb{R}$)
    repeat
        ▷ Incur cost $f_i(x_i)$, receive a subgradient $g_i \in \partial f_i(x_i)$
        $y_{i+1} \leftarrow x_i - \eta_i g_i$
        $x_{i+1} \leftarrow \gamma_i \Pi_{\mathcal{X}}(y_{i+1}) + (1-\gamma_i) x_1$    (project $y_{i+1}$ onto $\mathcal{X}$ then mix with $x_1$)
        $i \leftarrow i + 1$
    until solution desired in iteration $T$
```

Theorem 1.6. Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set. Suppose that $f_1, f_2, \ldots : \mathcal{X} \to \mathbb{R}$ are convex and 1-Lipschitz (with respect to $\|\cdot\|_2$). Assume that $\|x_1 - x^*\|_2 \leq 1$. Let $\eta_i = \frac{1}{\sqrt{2i}}$ and $\gamma_i = \eta_{i+1}/\eta_i$. Then the regret satisfies

$$\operatorname{Regret}(T) \;=\; \sum_{i=1}^{T} \big(f_i(x_i) - f_i(x^*)\big) \;\leq\; \sqrt{2T} \qquad \forall T \geq 1.$$

Proof. Defining x_{i+1} as the mixture $x_{i+1} = \gamma_i \Pi_{\mathcal{X}}(y_{i+1}) + (1-\gamma_i)x_1$ has a useful consequence due to convexity.

$$\begin{aligned}
\|x_{i+1} - x^*\|_2^2 \;&\leq\; \gamma_i \,\|\Pi_{\mathcal{X}}(y_{i+1}) - x^*\|_2^2 + (1-\gamma_i)\,\|x_1 - x^*\|_2^2 \qquad \text{(by convexity of } \|\cdot\|_2^2\text{)}\\
\Longrightarrow \qquad \|\Pi_{\mathcal{X}}(y_{i+1}) - x^*\|_2^2 \;&\geq\; \frac{1}{\gamma_i}\Big(\|x_{i+1} - x^*\|_2^2 - (1-\gamma_i)\,\|x_1 - x^*\|_2^2 \Big)\\
&=\; \frac{\eta_i}{\eta_{i+1}} \|x_{i+1} - x^*\|_2^2 - \Big(\frac{\eta_i}{\eta_{i+1}} - 1\Big) \|x_1 - x^*\|_2^2\,, \qquad (1.1)
\end{aligned}$$

by the definition of γ_i.

Next, we follow the proof of Theorem 1.5. The error on the i^{th} iteration is bounded as follows:

$$
\begin{aligned}
&f_i(x_i) - f_i(x^*) \\
&\quad\le \frac{1}{2\eta_i}\Big(\|x_i - y_{i+1}\|_2^2 + \|x_i - x^*\|_2^2 - \|y_{i+1} - x^*\|_2^2 \Big) \\
&\quad\le \frac{1}{2\eta_i}\Big(\|x_i - y_{i+1}\|_2^2 + \|x_i - x^*\|_2^2 - \|\Pi_{\mathcal{X}}(y_{i+1}) - x^*\|_2^2 \Big) \qquad \text{(by Claim 3.4)} \\
&\quad\le \frac{1}{2\eta_i}\left(\|x_i - y_{i+1}\|_2^2 + \|x_i - x^*\|_2^2 - \frac{\eta_i}{\eta_{i+1}}\|x_{i+1} - x^*\|_2^2 + \Big(\frac{\eta_i}{\eta_{i+1}} - 1\Big)\|x_1 - x^*\|_2^2 \right) \qquad \text{(by Eq. (1.1))} \\
&\quad= \frac{1}{2}\left(\frac{\|x_i - y_{i+1}\|_2^2}{\eta_i} + \underbrace{\frac{\|x_i - x^*\|_2^2}{\eta_i} - \frac{\|x_{i+1} - x^*\|_2^2}{\eta_{i+1}}}_{\text{telescopes}} + \underbrace{\Big(\frac{1}{\eta_{i+1}} - \frac{1}{\eta_i}\Big)}_{\text{telescopes}}\|x_1 - x^*\|_2^2 \right).
\end{aligned}
$$

Summing this over i yields

$$
\begin{aligned}
&\sum_{i=1}^{T}\big(f_i(x_i) - f_i(x^*)\big) \\
&\quad\le \frac{1}{2}\left(\sum_{i=1}^{T}\frac{\|x_i - y_{i+1}\|_2^2}{\eta_i} + \sum_{i=1}^{T}\Big(\frac{\|x_i - x^*\|_2^2}{\eta_i} - \frac{\|x_{i+1} - x^*\|_2^2}{\eta_{i+1}}\Big) + \sum_{i=1}^{T}\Big(\frac{1}{\eta_{i+1}} - \frac{1}{\eta_i}\Big)\|x_1 - x^*\|_2^2 \right) \\
&\quad\le \frac{1}{2}\left(\sum_{i=1}^{T}\frac{\|x_i - y_{i+1}\|_2^2}{\eta_i} + \frac{\|x_1 - x^*\|_2^2}{\eta_T} \right) \qquad \text{(telescoping)} \\
&\quad= \frac{1}{2}\Big(\sum_{i=1}^{T}\frac{\|\eta_i g_i\|_2^2}{\eta_i} + \frac{\|x_1 - x^*\|_2^2}{\eta_T} \Big) \qquad \text{(by the gradient step in line 3)} \\
&\quad\le \frac{1}{2}\Big(\sum_{i=1}^{T}\eta_i + \frac{1}{\eta_T} \Big) \qquad \text{(by the assumptions } \|g_i\|_2 \le 1 \text{ and } \|x_1 - x^*\|_2 \le 1\text{)} \\
&\quad< \frac{1}{2}\Big(\frac{2}{\sqrt{2}}\sqrt{T} + \sqrt{2T} \Big) \qquad \text{(by Claim 3.2 and the definition of } \eta_i\text{)}
\end{aligned}
$$

□

1.5 Stochastic gradient setting

Now we consider the setting in which we have a *stochastic* gradient oracle. When executed at a point x, it returns a vector $\hat{g}$ such that the *expectation* of $\hat{g}$ (conditioned on the past) is in $\partial f(x)$. The stochastic gradient descent algorithm, shown in Algorithm 6, is a trivial modification of Algorithm 2 to use this stochastic oracle.

The expected error of the stochastic gradient descent algorithm is easy to analyze. The proof is a modification of Theorem 1.3 that just requires a bit of care with conditional expectations.

First let us introduce some notation. Let $\mathcal{F}_i$ denote the sigma-field generated by $\hat{g}_1, ..., \hat{g}_i$. If

Algorithm 6 Stochastic gradient descent for minimizing convex, Lipschitz functions over a convex set.

```
procedure StochasticGradientDescent($\mathcal{X} \subseteq \mathbb{R}^n,\ x_1 \in \mathcal{X},\ t \in \mathbb{N}$)
    Let $\eta = 1/\sqrt{t}$
    for $i \leftarrow 1, ..., t$ do
        Let $\hat{g}_i$ be a random vector obtained from the subgradient oracle at $x_i$
        ▷ So $\mathrm{E}[\,\hat{g}_i \mid \mathcal{F}_{i-1}\,] \in \partial f(x_i)$
        $y_{i+1} \leftarrow x_i - \eta\, \hat{g}_i$ ,
        $x_{i+1} \leftarrow \Pi_{\mathcal{X}}(y_{i+1})$
    return $\sum_{i=1}^{t} x_i/t$
```

that is an uncomfortable notion for you, just think of $\mathcal{F}_i$ as being the vector $(\hat{g}_1, ..., \hat{g}_i)$. Define

$$\begin{aligned} \text{Expected subgradient:} \quad g_i &= \mathrm{E}[\,\hat{g}_i \mid \mathcal{F}_{i-1}\,] \\ \text{Noise in subgradient:} \quad \hat{z}_i &= g_i - \hat{g}_i \end{aligned}$$

Theorem 1.7. Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set. Suppose that $f : \mathcal{X} \to \mathbb{R}$ is convex. Assume that:

(a) $g_i \in \partial f(x_i)$ for all i (with probability 1).

(b) $\mathrm{E}\left[\, \|\hat{g}_i\|_2^2 \,\right] \leq 1$ for all i.

Fix an optimal solution $x^* \in \operatorname{argmin}_x f(x)$ and a starting point $x_1 \in \mathcal{X}$. Define $\eta = \frac{1}{\sqrt{T}}$. Suppose that $\|x_1 - x^*\|_2 \leq 1$. Then

$$\mathrm{E}\left[f\left(\frac{1}{T} \sum_{i=1}^{T} x_i \right) \right] - f(x^*) \;\leq\; \frac{1}{\sqrt{T}}.$$

References. [6, Theorem 14.8], [2, Section 6.1].

Proof. Hypothesis (a) and the subgradient inequality (3.2) imply that

$$f(x_i) - f(x^*) \;\leq\; \langle\, g_i,\, x_i - x^* \,\rangle \;=\; \langle\, \mathrm{E}[\,\hat{g}_i \mid \mathcal{F}_{i-1}\,],\, x_i - x^* \,\rangle \qquad \text{(with probability 1).}$$

Observe that both the left- and right-hand side are $\mathcal{F}_{i-1}$-measurable random variables. To see this, note that $x_i = x_1 - \eta \sum_{j=1}^{i-1} \hat{g}_j$, so the randomness of x_i is completely determined by $\hat{g}_1, ..., \hat{g}_{i-1}$. Taking the (unconditional) expectation

$$\begin{aligned} \mathrm{E}[\,f(x_i)\,] - f(x^*) &\leq \mathrm{E}[\ \langle\, \mathrm{E}[\,\hat{g}_i \mid \mathcal{F}_{i-1}\,],\, x_i - x^* \,\rangle\ ] \\ &= \mathrm{E}[\ \ \mathrm{E}[\,\langle\, \hat{g}_i,\, x_i - x^* \,\rangle \mid \mathcal{F}_{i-1}\,]\ \ ] \\ &= \mathrm{E}[\,\langle\, \hat{g}_i,\, x_i - x^* \,\rangle\,], \end{aligned}$$

since $\mathrm{E}[\,\mathrm{E}[\,A \mid F\,]\,] = \mathrm{E}[\,A\,]$ for any random variable (or sigma-field) F.

We bound the error on the i^{th} iteration as follows:

$$\begin{aligned}
\mathrm{E}\,[\,f(x_i)\,] - f(x^*) &= \mathrm{E}\,[\,\langle\, \hat{g}_i,\, x_i - x^*\,\rangle\,] \\
&= \mathrm{E}\left[\,\frac{1}{\eta}\langle\, x_i - y_{i+1},\, x_i - x^*\,\rangle\right] \qquad \text{(by the gradient step)} \\
&= \mathrm{E}\left[\,\frac{1}{2\eta}\Big(\,\|x_i - y_{i+1}\|_2^2 + \|x_i - x^*\|_2^2 - \|y_{i+1} - x^*\|_2^2\Big)\right] \\
&\leq \mathrm{E}\left[\,\frac{1}{2\eta}\Big(\,\|x_i - y_{i+1}\|_2^2 + \|x_i - x^*\|_2^2 - \|x_{i+1} - x^*\|_2^2\Big)\right].
\end{aligned}$$

The last line uses Claim 3.4: since x_{i+1} is the projected point $\Pi_{\mathcal{X}}(y_{i+1})$ and $x^* \in \mathcal{X}$, the corollary yields that $\|x_{i+1} - x^*\|_2^2 \leq \|y_{i+1} - x^*\|_2^2$. To analyze the average error, sum the previous displayed equation over i. The last two terms telescope, yielding

$$\begin{aligned}
\mathrm{E}\left[\sum_{i=1}^{T}\big(f(x_i) - f(x^*)\big)\right] &\leq \mathrm{E}\left[\,\frac{1}{2\eta}\bigg(\Big(\sum_{i=1}^{T}\|x_i - y_{i+1}\|_2^2\Big) + \|x_1 - x^*\|_2^2 - \|x_{T+1} - x^*\|_2^2\bigg)\right] \\
&\leq \frac{1}{2\eta}\Big(\sum_{i=1}^{T}\mathrm{E}\left[\big\|\eta\,\hat{g}_i\big\|_2^2\right] + \|x_1 - x^*\|_2^2\Big) \quad \text{(by the gradient step)} \\
&\leq \frac{\eta T}{2} + \frac{1}{2\eta} \qquad \text{(by hypothesis (b) and assumption on } x_1\text{)}
\end{aligned}$$

Dividing by T and using Jensen's inequality (Lemma 3.8) and the definition of η gives

$$\mathrm{E}\left[f\bigg(\sum_{i=1}^{T}\frac{x_i}{T}\bigg)\right] - f(x^*) \;\leq\; \mathrm{E}\left[\sum_{i=1}^{T}\frac{1}{T}\big(f(x_i) - f(x^*)\big)\right] \;\leq\; \frac{\eta}{2} + \frac{1}{2T\eta} \;=\; \frac{1}{\sqrt{T}},$$

as required. □

Remark. Suppose that f is 1/2-Lipschitz and that $\mathrm{E}\left[\|\hat{z}_i\|^2\right] \leq 1/4$ for each i. By Theorem 3.6 we have $\|g_i\| \leq 1/2$ for all i (with probability 1). Furthermore,

$$\mathrm{E}\left[\|\hat{g}_i\|^2\right] \;=\; \mathrm{E}\left[\|g_i - \hat{z}_i\|^2\right] \;\leq\; 2\,\mathrm{E}\left[\|g_i\|^2 + \|\hat{z}_i\|^2\right] \;\leq\; 2\big(\tfrac{1}{4} + \mathrm{E}\left[\|\hat{z}_i\|^2\right]\big) \;\leq\; 1.$$

So Theorem 1.7 applies.

1.6 Analysis of the last iterate

Theorem 1.8. Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set. Suppose that $f : \mathcal{X} \to \mathbb{R}$ is convex and 1-Lipschitz (with respect to $\|\cdot\|_2$). Define $\eta_i = \frac{1}{\sqrt{i}}$. Suppose that $\operatorname{diam}(\mathcal{X}) \leq 1$. Then Algorithm 7 satisfies

$$f(x_T) - f(x^*) \;\leq\; \frac{3(2 + \log T)}{2\sqrt{T}}.$$

References. Shamir-Zhang [7, Theorem 2].

Algorithm 7 Projected gradient descent for minimizing a convex, 1-Lipschitz function f with an unknown time horizon.

```
procedure ProjectedGradientDescent($\mathcal{X} \subseteq \mathbb{R}^n,\ x_1 \in \mathcal{X}$)
    for $i \leftarrow 1, 2, \ldots$ do
        Let $\eta_i = 1/\sqrt{i}$
        $y_{i+1} \leftarrow x_i - \eta_i g_i^{\mathsf{T}}$, where $g_i \in \partial f(x_i)$.
        $x_{i+1} \leftarrow \Pi_{\mathcal{X}}(y_{i+1})$
```

Proof. The first step is identical to the proof of Theorem 1.5:

$$\begin{aligned}
f(x_i) - f(w) &\leq \langle\, g_i,\, x_i - w \,\rangle \\
&= \frac{1}{\eta_i} \langle\, x_i - y_{i+1},\, x_i - w \,\rangle \\
&= \frac{1}{2\eta_i} \Big(\|x_i - y_{i+1}\|_2^2 + \|x_i - w\|_2^2 - \|y_{i+1} - w\|_2^2 \Big) \\
&\leq \frac{1}{2\eta_i} \Big(\|x_i - y_{i+1}\|_2^2 + \|x_i - w\|_2^2 - \|x_{i+1} - w\|_2^2 \Big).
\end{aligned}$$

The next step is similar to the proof of Theorem 1.5, except that the sum starts at $i = T - k$. Crucially, instead of substituting $w = x^*$, we substitute $w = x_{T-k}$, which causes the quantity $\|x_{T-k} - w\|_2^2$ to vanish.

$$\begin{aligned}
&\sum_{i=T-k}^{T} \big(f(x_i) - f(x_{T-k}) \big) \\
&\leq \sum_{i=T-k}^{T} \frac{\|x_i - y_{i+1}\|_2^2}{2\eta_i} + \frac{1}{2} \sum_{i=T-k+1}^{T} \Big(\frac{1}{\eta_i} - \frac{1}{\eta_{i-1}} \Big) \|x_i - x_{T-k}\|_2^2 - \frac{1}{\eta_T} \|x_{T+1} - x_{T-k}\|_2^2 \\
&\leq \sum_{i=T-k}^{T} \frac{\|\eta_i g_i\|_2^2}{2\eta_i} + \frac{1}{2} \sum_{i=T-k+1}^{T} \Big(\sqrt{i} - \sqrt{i-1} \Big) \qquad \text{(diameter bound)} \\
&= \frac{1}{2} \sum_{i=T-k}^{T} \frac{1}{\sqrt{i}} + \frac{1}{2} (\sqrt{T} - \sqrt{T-k}) \qquad \text{(Lipschitz assumption)} \\
&\leq \frac{3}{2} (\sqrt{T} - \sqrt{T-k-1}),
\end{aligned}$$

due to the bound

$$\sum_{i=a}^{b} \frac{1}{\sqrt{i}} \;\leq\; \int_{a-1}^{b-1} \frac{1}{\sqrt{x}}\, dx \;\leq\; 2(\sqrt{b} - \sqrt{a-1}).$$

Thus, using Claim 3.3, we have

$$\sum_{i=T-k}^{T} \big(f(x_i) - f(x_{T-k}) \big) \;\leq\; \frac{3}{2} \cdot \frac{k+1}{\sqrt{T} + \sqrt{T-k-1}}.$$

Now divide this by $k+1$ and define $S_k = \frac{1}{k+1} \sum_{i=T-k}^{T} f(x_i)$ to obtain

$$S_k - f(x_{T-k}) \;\leq\; \frac{3}{2\sqrt{T}}.$$

Observe that $kS_{k-1} = (k+1)S_k - f(x_{T-k})$. Combining this with the previous inequality yields

$$kS_{k-1} \;=\; kS_k + \big(S_k - f(x_{T-k})\big) \;\leq\; kS_k + \frac{3}{2\sqrt{T}}.$$

Dividing by k, we obtain

$$S_{k-1} \;\leq\; S_k + \frac{3}{2k\sqrt{T}}.$$

Thus, by induction,

$$f(x_T) \;=\; S_0 \;\leq\; S_{T-1} + \frac{3}{2\sqrt{T}} \sum_{k=1}^{T-1} \frac{1}{k} \;\leq\; S_{T-1} + \frac{3(1+\log T)}{2\sqrt{T}}.$$

Finally, in the proof of Theorem 1.5, we have already shown that

$$S_{T-1} - f(x^*) \;=\; \frac{1}{T}\sum_{i=1}^{T}\big(f(x_i) - f(x^*)\big) \;\leq\; \frac{3}{2\sqrt{T}}.$$

Combining the last two inequalities yields

$$f(x_T) - f(x^*) \;\leq\; \frac{3}{2\sqrt{T}} + \frac{3(1+\log T)}{2\sqrt{T}},$$

completing the proof. □

1.7 General reduction from arbitrary scale & Lipschitz value

Our analyses above make two assumptions

- *Scale of codomain:* the given function f is 1-Lipschitz, and
- *Scale of domain:* $\|x_1 - x^*\|_2 \leq 1$.

How can we handle a general scale, say an L-Lipschitz function with $\|x_1 - x^*\|_2 \leq R$? There is a general reduction that can handle such scenarios.

Meta-theorem. Suppose that we have a theorem giving a convergence rate guarantee $c(T)$ for gradient descent assuming that $f : \hat{\mathcal{X}} \to \mathbb{R}$ is 1-Lipschitz and assuming $\|x_1 - x^*\|_2 \leq 1$. Suppose that $h : \mathcal{X} \to \mathbb{R}$ is a convex function that is L-Lipschitz, and such that $\|x_1 - x^*\| \leq R$. Then there is a black-box reduction from h to f, showing that gradient descent on h achieves convergence rate $RL \cdot c(T)$.

Proof of meta-theorem. Let $OPT = \min_{x\in\mathcal{X}} h(x)$. Define $\hat{\mathcal{X}} = \mathcal{X}/R$ and $f : \hat{\mathcal{X}} \to \mathbb{R}$ by

$$f(x) \;=\; \frac{1}{RL}(h(Rx) - OPT).$$

Thus,

$$\begin{aligned} h(x) &\;=\; RL \cdot f(x/R) + OPT \qquad (1.2) \\ \min_{x\in\hat{\mathcal{X}}} f(x) &\;=\; 0. \end{aligned}$$

Claim 1.9. $v \in \partial h(x)$ iff $v/L \in \partial f(x/R)$.

Consider running gradient descent on h with step sizes $\eta_t = \frac{R}{L\sqrt{t}}$ from the starting point x_1, producing iterates $x_2, x_3, \ldots.$ Let g_i be the subgradient used in the i^{th} iteration. Define $\hat{g}_i = g_i/L$.

Simulataneously, imagine running gradient descent on f with step sizes $\hat{\eta_t} = \frac{1}{\sqrt{t}} = \frac{L}{R}\eta_t$ and vectors $\hat{g}_i = g_i/L$, from the starting point $\hat{x_1} = x_1/R$. Let $\hat{x_2}, \hat{x_3}, \ldots$ be the vectors produced.

Claim 1.10. $\hat{x_i} = x_i/R$ for all $i \geq 1$.

Proof. By induction, the case $i = 1$ true by definition. So suppose true up to i. By definition $g_i \in \partial h(x_i)$, so Claim 1.9 implies that $\hat{g}_i \in \partial f(x_i/R) = \partial f(\hat{x}_i)$. Then

$$\hat{x_{i+1}} \;=\; \hat{x_i} - \hat{\eta_i} \cdot \hat{g_i} \;=\; \frac{1}{R}x_i - \frac{L}{R}\eta_i \cdot \frac{1}{L}g_i \;=\; \frac{1}{R}(x_i - \eta_i g_i) \;=\; \frac{1}{R}x_{i+1}. \quad \square$$

To illustrate the meta-theorem, we apply it to Theorem 1.1, obtaining:

Theorem 1.11. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and L-Lipschitz (with respect to $\|\cdot\|_2$). Fix an optimal solution $x^* \in \operatorname{argmin}_x f(x)$ and a starting point $x_1 \in \mathbb{R}^n$. Define $\eta = \frac{L}{R\sqrt{T}}$. Suppose that $\|x_1 - x^*\|_2 \leq R$. Then

$$\begin{aligned} h\left(\sum_{i=1}^{T} \frac{x_i}{T}\right) - h(x^*) &= RL \cdot f\left(\sum_{i=1}^{T} \frac{x_i}{RT}\right) \qquad \text{(by (1.2))} \\ &= RL \cdot f\left(\sum_{i=1}^{T} \frac{\hat{x_i}}{T}\right) \qquad \text{(by Claim 1.10)} \\ &\leq \frac{RL}{\sqrt{T}} \qquad \text{(by Theorem 1.1).} \end{aligned}$$

2 Strongly convex and Lipschitz functions

In this section we consider a stronger assumption on the function f: we assume it is α-strongly convex and L-Lipschitz.

2.1 Online setting

First of all we do the online, projected setting, with unknown time horizon. The theorem is a modification of Theorem 1.5.

Algorithm 8 Online projected gradient descent for strongly convex, Lipschitz functions, with unknown time horizon.

```
1: procedure OnlineProjectedGradientDescent($\mathcal{X} \subseteq \mathbb{R}^n$, $x_1 \in \mathcal{X}$)
2:     Let $\eta_i = 1/i$ for all $i \in \mathbb{N}$
3:     $i \leftarrow 1$
4:     repeat
           ▷ Incur cost $f_i(x_i)$, receive a subgradient $g_i \in \partial f_i(x_i)$
5:         $y_{i+1} \leftarrow x_i - \eta_i g_i$
6:         $x_{i+1} \leftarrow \Pi_{\mathcal{X}}(y_{i+1})$
7:         $i \leftarrow i + 1$
8:     until solution desired in iteration $T$
9:     return $\sum_{i=1}^{T} \frac{x_i}{T}$
```

Theorem 2.1. Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set. Suppose that $f_1, f_2, \ldots : \mathcal{X} \to \mathbb{R}$ are 1-strongly convex and 1-Lipschitz (with respect to $\|\cdot\|_2$). Set $\eta_i = \frac{1}{i}$. Then

$$\operatorname{Regret}(t) \;=\; \sum_{i=1}^{T} \big(f_i(x_i) - f_i(x^*) \big) \;\leq\; \frac{1 + \ln T}{2} \qquad \forall T \geq 1.$$

Consequently, in the offline setting where each $f_i = f$,

$$f\left(\frac{1}{T} \sum_{i=1}^{T} x_i \right) - f(x^*) \;\leq\; \frac{1 + \ln T}{2T} \qquad \forall T \geq 1.$$

References. In the online setting, this result originally appeared as [4, Theorem 1]. See also [3, Theorem 3.3], [1, Theorem 2.3].

Proof. We bound the error on the i^{th} iteration as follows:

$$\begin{aligned}
f_i(x_i) - f_i(x^*) \;&\leq\; \langle\, g_i,\, x_i - x^* \,\rangle - \frac{1}{2} \|x_i - x^*\|_2^2 && \text{(by (3.5))} \\
&=\; \frac{1}{\eta_i} \langle\, x_i - y_{i+1},\, x_i - x^* \,\rangle - \frac{1}{2} \|x_i - x^*\|_2^2 && \text{(by the gradient step in line 5)} \\
&=\; \frac{1}{2\eta_i} \Big(\|x_i - y_{i+1}\|_2^2 + \|x_i - x^*\|_2^2 - \|y_{i+1} - x^*\|_2^2 \Big) - \frac{1}{2} \|x_i - x^*\|_2^2 && \text{(by (3.1))}
\end{aligned}$$

$$\leq \frac{1}{2\eta_i}\Big(\|x_i - y_{i+1}\|_2^2 + \|x_i - x^*\|_2^2 - \|x_{i+1} - x^*\|_2^2\Big) - \frac{1}{2}\|x_i - x^*\|_2^2.$$

The last line uses Claim 3.4: since x_{i+1} is the projected point $\Pi_{\mathcal{X}}(y_{i+1})$ and $x^* \in \mathcal{X}$, the corollary yields that $\|x_{i+1} - x^*\|_2^2 \leq \|y_{i+1} - x^*\|_2^2$. To analyze the average error, sum the previous displayed equation over i. The parameter η_i is chosen so that the sum involving the last three terms will telescope:

$$\begin{aligned}
&\sum_{i=1}^{T}\big(f_i(x_i) - f_i(x^*)\big)\\
&\leq \sum_{i=1}^{T}\frac{\|x_i - y_{i+1}\|_2^2}{2\eta_i} + \frac{1}{2}\Big(\frac{1}{\eta_1} - 1\Big)\|x_1 - x^*\|_2^2 + \frac{1}{2}\sum_{i=2}^{T}\Big(\frac{1}{\eta_i} - \frac{1}{\eta_{i-1}} - 1\Big)\|x_i - x^*\|_2^2\\
&= \sum_{i=1}^{T}\frac{\|\eta_i g_i\|_2^2}{2\eta_i} + \frac{1}{2}\underbrace{\Big(1-1\Big)}_{=0}\|x_1 - x^*\|_2^2 + \frac{1}{2}\sum_{i=2}^{T}\underbrace{\Big(i-(i-1)-1\Big)}_{=0}\|x_i - x^*\|_2^2\\
&= \frac{1}{2}\sum_{i=1}^{T}\frac{\|g_i\|_2^2}{i}\\
&\leq \frac{1+\ln T}{2}.
\end{aligned}$$

This completes the proof of the regret bound. Dividing by T and using Jensen's inequality (Lemma 3.8), we obtain

$$f\left(\sum_{i=1}^{T}\frac{x_i}{T}\right) - f(x^*) \leq \sum_{i=1}^{T}\frac{1}{T}\big(f(x_i) - f(x^*)\big) \leq \frac{1+\ln T}{2T},$$

as required. □

2.2 Improved bound in the offline setting

In the offline setting, we can improve the analysis by a factor of $\log(T)$ through the use of a non-uniform convex combination. The algorithm, shown in Algorithm 9, is a small variant of Algorithm 8, with non-uniform averaging. The theorem is a modification of Theorem 2.1.

Theorem 2.2. Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set. Suppose that $f : \mathcal{X} \to \mathbb{R}$ is 1-strongly convex and 1-Lipschitz (with respect to $\|\cdot\|_2$). Set $\eta_i = \frac{2}{i+1}$. Then, letting $\lambda_i = \frac{i}{T(T+1)/2}$,

$$f\Big(\sum_{i=1}^{T}\lambda_i x_i\Big) - f(x^*) \leq \frac{2}{T+1}.$$

References. [2, Theorem 3.9], [1, Theorem 2.4], [5].

Remark. Theorem 2.2 is optimal [2, Theorem 3.13].

Algorithm 9 Projected gradient descent for strongly convex, Lipschitz functions, with an unknown time horizon.

```
1: procedure StrongGDNonUniform($\mathcal{X} \subseteq \mathbb{R}^n,\ x_1 \in \mathcal{X}$)
2:     Let $\eta_i = \frac{2}{i+1}$ for all $i \in \mathbb{N}$
3:     $i \leftarrow 1$
4:     repeat
5:         $y_{i+1} \leftarrow x_i - \eta_i g_i$, where $g_i \in \partial f(x_i)$
6:         $x_{i+1} \leftarrow \Pi_{\mathcal{X}}(y_{i+1})$
7:         $i \leftarrow i + 1$
8:     until solution desired in iteration $T$
9:     return $\sum_{i=1}^{T} \frac{i}{T(T+1)/2} x_i$
```

Proof. We bound the error on the i^{th} iteration as follows:

$$\begin{aligned}
f(x_i) - f(x^*) &\leq \langle\, g_i,\, x_i - x^* \,\rangle - \frac{1}{2}\|x_i - x^*\|_2^2 && \text{(by (3.5))}\\
&= \frac{1}{\eta_i}\langle\, x_i - y_{i+1},\, x_i - x^* \,\rangle - \frac{1}{2}\|x_i - x^*\|_2^2 && \text{(by the gradient step in line 5)}\\
&= \frac{1}{2\eta_i}\Big(\|x_i - y_{i+1}\|_2^2 + \|x_i - x^*\|_2^2 - \|y_{i+1} - x^*\|_2^2\Big) - \frac{1}{2}\|x_i - x^*\|_2^2 && \text{(by (3.1))}\\
&\leq \frac{1}{2\eta_i}\Big(\|x_i - y_{i+1}\|_2^2 + \|x_i - x^*\|_2^2 - \|x_{i+1} - x^*\|_2^2\Big) - \frac{1}{2}\|x_i - x^*\|_2^2 .
\end{aligned}$$

The last line uses Claim 3.4: since x_{i+1} is the projected point $\Pi_{\mathcal{X}}(y_{i+1})$ and $x^* \in \mathcal{X}$, the corollary yields that $\|x_{i+1} - x^*\|_2^2 \leq \|y_{i+1} - x^*\|_2^2$.

To avoid a harmonic sum arising from the first term, we first multiply this inequality by i before summing. First, we simplify as follows:

$$\begin{aligned}
i \cdot \big(f(x_i) - f(x^*)\big) &\leq \frac{i\,\|\eta_i g_i\|_2^2}{2\eta_i} + i\Big(\frac{1}{2\eta_i} - \frac{1}{2}\Big)\|x_i - x^*\|_2^2 - \frac{i}{2\eta_i}\|x_{i+1} - x^*\|_2^2\\
&= \frac{i\,\|g_i\|_2^2}{i+1} + \Big(\frac{i(i+1)}{4} - \frac{2i}{4}\Big)\|x_i - x^*\|_2^2 - \frac{i(i+1)}{4}\|x_{i+1} - x^*\|_2^2\\
&\leq 1 + \frac{1}{4}\cdot\Big(i(i-1)\,\|x_i - x^*\|_2^2 - i(i+1)\,\|x_{i+1} - x^*\|_2^2\Big).
\end{aligned}$$

Now, summing over i, the right-hand side telescopes and we obtain

$$\sum_{i=1}^{T} i \cdot \big(f(x_i) - f(x^*)\big) \;\leq\; T - \frac{1}{4}T(T+1)\,\|x_{T+1} - x^*\|_2^2 \;\leq\; T.$$

Dividing by $T(T+1)/2$ and applying Jensen's inequality completes the proof.

□

3 Basic Facts

Claim 3.1 (Cosine Law).

$$\|a-b\|_2^2 = \|a\|_2^2 - 2a^{\mathsf{T}}b + \|b\|_2^2 \qquad \forall a, b \in \mathbb{R}^n. \tag{3.1}$$

Claim 3.2. For any $n \in \mathbb{N}$, $2\sqrt{n} - 2 \leq \sum_{i=1}^{n} \frac{1}{\sqrt{i}} \leq 2\sqrt{n} - 1$.

Claim 3.3 (Difference of square roots). $\sqrt{a} - \sqrt{a-b} \ = \ \frac{b}{\sqrt{a}+\sqrt{a-b}}$.

Proof. Note that $(\sqrt{a} - \sqrt{a-b})(\sqrt{a} + \sqrt{a-b}) = \sqrt{a}^2 - \sqrt{a-b}^2 = b$. □

Claim 3.4 (Projection decreases Euclidean distance). $\|\Pi_{\mathcal{X}}(y) - x\|_2 \leq \|y - x\|_2$ for all $x \in \mathcal{X}$.

Definition 3.5 (Subgradient). Let $f : \mathcal{X} \to \mathbb{R}^n$ be a function. Recall that a ***subgradient*** of f at x is any vector g satisfying:

$$f(y) \ \geq \ f(x) + \langle\, g,\, y - x \,\rangle \qquad \forall y \in \mathcal{X}. \tag{3.2}$$

Theorem 3.6 (Lipschitz equivalence). Let $\mathcal{X}$ be convex and open. Let $f : \mathcal{X} \to \mathbb{R}$ be convex. For any norm $\|\cdot\|$, the following conditions are equivalent.

- $f : \mathcal{X} \to \mathbb{R}$ is L-Lipschitz with respect to $\|\cdot\|$:

$$|f(x) - f(y)| \ \leq \ L\,\|x - y\| \qquad \forall x, y \in \mathcal{X}. \tag{3.3}$$

- f has bounded subgradients:

$$\|g\|_* \leq L \qquad \forall w \in X,\ g \in \partial f(w). \tag{3.4}$$

Claim 3.7. Suppose that f is α-strongly convex and $g \in \partial f(x)$. Then

$$f(y) \ \geq \ f(x) + \langle\, g,\, y - x \,\rangle + \frac{\alpha}{2}\,\|x - y\|_2^2 \quad \forall y \in \mathcal{X}. \tag{3.5}$$

Lemma 3.8 (Jensen's inequality). Let $f : \mathcal{X} \to \mathbb{R}$ be a convex function. Let $x_1, ..., x_n \in \mathcal{X}$. Let $\lambda_1, ..., \lambda_n \in [0, 1]$ satisfy $\sum_{i=1}^{n} \lambda_i = 1$. Then $f(\sum_{i=1}^{n} \lambda_i x_i) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$.

References

[1] Nikhil Bansal and Anupam Gupta. Potential-function proofs for first-order methods, 2017.

[2] Sebastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3–4), 2015.

[3] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3–4), 2015.

[4] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

[5] Simon Lacoste-Julien, Mark W. Schmidt, and Francis R. Bach. A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method, 2012. arXiv:1212.2002.

[6] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[7] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. *Proceedings of the 30th International Conference on Machine Learning, PMLR*, 28(1):71–79, 2013.

[8] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of ICML*, pages 928–936, 2003.