1 Deterministic Optimization

For brevity, everywhere differentiable functions will be called smooth. Similarly, not everywhere differentiable functions will be called nonsmooth.

First, we extend the Lipschitz continuity definition (Definition 6.4) to higher dimensions.

**Definition 7.1** (Lipschitz continuity). A function $\phi: \mathbb{R}^p \to \mathbb{R}$ is called $K$-Lipschitz continuous with respect to the norm $\|\cdot\|$ if and only if there is a constant $K < +\infty$ such that

$$\forall (w, u) \in \mathbb{R}^p : |\phi(w) - \phi(u)| \leq K \|w - u\|$$

A smooth function $\phi: \mathbb{R}^p \to \mathbb{R}$ is called $K$-Lipschitz continuous with respect to the norm $\|\cdot\|$ if and only if there is a constant $K < +\infty$ such that

$$\forall w \in \mathbb{R}^p : \|\nabla \phi(w)\| \leq K$$

Recall that the gradient of a smooth convex function $\phi: \mathbb{R}^p \to \mathbb{R}$ at $w$ fulfills:

$$\forall u \in \mathbb{R}^p : \phi(u) - \phi(w) \geq \langle \nabla \phi(w), u - w \rangle$$

**Definition 7.2** (Subgradient). For a (possibly nonsmooth) convex function $\phi: \mathbb{R}^p \to \mathbb{R}$, we can define a subdifferential set as follows:

$$\partial \phi(w) = \{ g \mid \forall u \in \mathbb{R}^p : \phi(u) - \phi(w) \geq \langle g, u - w \rangle \}$$

Each element $g \in \partial \phi(w)$ is called a subdifferential or subgradient of $\phi$ at $w$.

Clearly, in the above definition, if $\phi: \mathbb{R}^p \to \mathbb{R}$ is smooth, then $\partial \phi(w)$ has a single element for every $w \in \mathbb{R}^p$. If $\partial \phi(w)$ is nonsmooth there exist some $w \in \mathbb{R}^p$ for which $\partial \phi(w)$ has more than one element.

Consider for instance the nonsmooth function $\phi(w) = |w|$ where $w \in \mathbb{R}$. By Definition 7.2, we have:

$$\partial \phi(0) = \{ g \mid \forall u \in \mathbb{R} : |u| \geq g u \}$$

Thus, clearly $\partial \phi(0) = [-1, +1]$. 

1
Now, consider the following optimization problem where \( f : \mathbb{R}^p \rightarrow \mathbb{R} \) is convex and \( K \)-Lipschitz with respect to the \( \ell_2 \)-norm:

\[
\hat{w} = \arg \min_{w \in \mathbb{R}^p} f(w) \tag{1}
\]

Let \( \eta_t \) be the step size at iteration \( t \geq 1 \). Specifically, let \( \beta \) be a constant factor and define:

\[
\eta_t = \frac{\beta}{K \sqrt{t}}
\]

Consider the next subgradient descent algorithm for solving the above problem:

**Algorithm 7.1 Subgradient descent algorithm**

**Input:** Number of iterations \( T \geq 1 \), factor \( \beta > 0 \), initial point \( w^{(1)} \in \mathbb{R}^p \) (The setting of \( w^{(1)} \) can be uninformed, e.g., \( w^{(1)} = 0 \))

for \( t = 1 \ldots T - 1 \) do
  \( w^{(t+1)} \leftarrow w^{(t)} - \eta_t g^{(t)} \) where \( g^{(t)} \in \partial f(w^{(t)}) \)
end for

**Output:** \( \tilde{w}^{(T)} \leftarrow \frac{\sum_{t=1}^T \eta_t w^{(t)}}{\sum_{t=1}^T \eta_t} \)

In what follows, we state our main result regarding convergence rates for Algorithm 7.1.

**Theorem 7.1 (Adapted from [1, 2]).** Assume that \( f : \mathbb{R}^p \rightarrow \mathbb{R} \) is convex and \( K \)-Lipschitz with respect to the \( \ell_2 \)-norm in the problem of eq.(1). Recall that \( \hat{w} \) is the optimal solution of the problem of eq.(1). Assume that Algorithm 7.1 runs for a number of iterations \( T \), factor \( \beta \) and initial point \( w^{(1)} \), and that the algorithm outputs \( \tilde{w}^{(T)} \). We have:

\[
f(\tilde{w}^{(T)}) - f(\hat{w}) \leq \frac{K \|w^{(1)} - \hat{w}\|^2}{4 \beta (\sqrt{T} - 1)} + \frac{\beta K (1 + \log T)}{4 (\sqrt{T} - 1)}
\]

**Proof.** Let \( a^{(t)} = \|w^{(t)} - \hat{w}\|^2 \). Note that since \( g^{(t)} \in \partial f(w^{(t)}) \), by Definition 7.2 we have:

\[
f(\tilde{w}) - f(w^{(t)}) \geq \langle g^{(t)}, \tilde{w} - w^{(t)} \rangle
\]

In fact, the above holds \( \forall \tilde{w} \in \mathbb{R}^p \) but in our problem we care about a unique \( \tilde{w} \). By the Lipschitz continuity of \( f \), we know that \( (\forall t) \|g^{(t)}\|_2 \leq K \). Therefore:

\[
a^{(t+1)} = \|w^{(t+1)} - \hat{w}\|^2 \\
= \|(w^{(t)} - \hat{w}) - \eta_t g^{(t)}\|^2 \\
= \|w^{(t)} - \hat{w}\|^2 - 2 \eta_t \langle g^{(t)}, w^{(t)} - \hat{w} \rangle + \eta_t^2 \|g^{(t)}\|^2 \\
\leq a^{(t)} + 2 \eta_t \left( f(\tilde{w}) - f(w^{(t)}) \right) + \eta_t^2 K^2
\]
Reorganizing the above, we obtain:

\[ \eta_t \left( f(w^{(t)}) - f(\hat{w}) \right) \leq \frac{1}{2} \left( a^{(t)} - a^{(t+1)} + \eta_t^2 K^2 \right) \]

Summing over \( t \), we get:

\[ \sum_{t=1}^{T} \eta_t \left( f(w^{(t)}) - f(\hat{w}) \right) \leq \frac{1}{2} \sum_{t=1}^{T} \left( a^{(t)} - a^{(t+1)} + \eta_t^2 K^2 \right) \]

\[ = \frac{1}{2} \left( a^{(1)} - a^{(T+1)} + K^2 \sum_{t=1}^{T} \eta_t^2 \right) \]

\[ \leq \frac{1}{2} \left( \|w^{(1)} - \hat{w}\|_2^2 + \beta^2 \sum_{t=1}^{T} \frac{1}{t} \right) \]

\[ \leq \frac{1}{2} \left( \|w^{(1)} - \hat{w}\|_2^2 + \beta^2 (1 + \log T) \right) \]

where we used the fact that \( \sum_{t=1}^{T} 1/t \leq 1 + \log T \). By Jensen’s inequality and convexity of \( f \), we have:

\[ f(\hat{w}^{(T)}) - f(\hat{w}) = f \left( \frac{\sum_{t=1}^{T} \eta_t w^{(t)}}{\sum_{t=1}^{T} \eta_t} \right) - f(\hat{w}) \]

\[ \leq \frac{\sum_{t=1}^{T} \eta_t f(w^{(t)})}{\sum_{t=1}^{T} \eta_t} - f(\hat{w}) \]

\[ = \frac{\sum_{t=1}^{T} \eta_t (f(w^{(t)}) - f(\hat{w}))}{\sum_{t=1}^{T} \eta_t} \]

\[ = \frac{\sum_{t=1}^{T} \eta_t (f(w^{(t)}) - f(\hat{w}))}{\beta K \sum_{t=1}^{T} \frac{1}{\sqrt{t}}} \]

\[ \leq \frac{1}{2} \left( \|w^{(1)} - \hat{w}\|_2^2 + \beta^2 (1 + \log T) \right) \]

\[ \leq \frac{2\beta}{K} \left( \sqrt{T} - 1 \right) \]

where we used the fact that \( 2(\sqrt{T} - 1) \leq \sum_{t=1}^{T} 1/t \). This proves our claim. \( \Box \)

2 A more general setting?

Now, consider the following optimization problem where \( f, r : \mathbb{R}^p \to \mathbb{R} \) are convex and \( K \)-Lipschitz with respect to the \( \ell_2 \)-norm:

\[ \hat{w} = \arg \min_{w \in \mathbb{R}^p} f(w) + r(w) \quad (2) \]

Consider the next subgradient descent algorithm for solving the above problem:
Algorithm 7.2 Subgradient descent algorithm

**Input:** Number of iterations $T \geq 1$, factor $\beta > 0$, initial point $w^{(1)} \in \mathbb{R}^p$ (The setting of $w^{(1)}$ can be uninformed, e.g., $w^{(1)} = 0$)

for $t = 1 \ldots T - 1$ do

\[
w^{(t+1/2)} \leftarrow w^{(t)} - \eta_t g^{(t)} \text{ where } g^{(t)} \in \partial f(w^{(t)})
\]

\[
w^{(t+1)} \leftarrow \arg \min_{w \in \mathbb{R}^p} \left( \frac{1}{2} \|w - w^{(t+1/2)}\|^2_2 + \eta_{t+1} r(w) \right)
\]

end for

**Output:** $\bar{w}^{(T)} \leftarrow \frac{\sum_{t=1}^{T} \eta_t w^{(t)}}{\sum_{t=1}^{T} \eta_t}$

Before going into more general observations, let’s first consider an example in order to show the usefulness of the above.

**Sparse optimization example.** Let $r(w) = \lambda \|w\|_1$. In this case, the second assignment reduces to $p$ independent assignments of the form:

\[
(\forall j = 1, \ldots, p) \quad w_j^{(t+1)} \leftarrow \arg \min_{w_j \in \mathbb{R}} \left( \frac{1}{2} (w_j - w_j^{(t+1/2)})^2 + \eta_{t+1} \lambda |w_j| \right)
\]

\[
\leftarrow \text{sgn}(w_j^{(t+1/2)}) \max(0, |w_j^{(t+1/2)}| - \eta_{t+1} \lambda)
\]

While the Algorithm 7.2 might look like a drastic generalization of Algorithm 7.1, in fact it is not. Note that:

\[
w^{(t+1)} = \arg \min_{w \in \mathbb{R}^p} \left( \frac{1}{2} \|w - w^{(t+1/2)}\|^2_2 + \eta_{t+1} r(w) \right)
\]

(3)

Recall that $g^{(t)} \in \partial f(w^{(t)})$ and let $s^{(t)} \in \partial r(w^{(t)})$. Note that $w^{(t+1)}$ is optimal if and only if $0$ belongs to the subdifferential set of eq.(3) evaluated at $w^{(t+1)}$.

\[
0 \in \partial \left( \frac{1}{2} \|w - w^{(t+1/2)}\|^2_2 + \eta_{t+1} r(w) \right) \bigg|_{w = w^{(t+1)}}
\]

\[
0 \in \left( w - w^{(t+1/2)} + \eta_{t+1} \partial r(w) \right) \bigg|_{w = w^{(t+1)}}
\]

\[
0 = w^{(t+1)} - w^{(t+1/2)} + \eta_{t+1} s^{(t+1)}
\]

\[
= w^{(t+1)} - w^{(t)} + \eta_t g^{(t)} + \eta_{t+1} s^{(t+1)}
\]

Reorganizing the above, we obtain:

\[
w^{(t+1)} = w^{(t)} - \eta_t g^{(t)} - \eta_{t+1} s^{(t+1)}
\]

which looks a lot like a regular subgradient descent step.
3 Exercise

Prove convergence of Algorithm 7.2.

4 Stochastic Optimization

In several optimization problems, it is possible to compute a stochastic version of the subgradient a lot faster than the true subgradient.

Application 1: Stochastic Sample. For instance, in the problem of eq.(1), consider functions that depend on \( n \) data samples \( z^{(1)} \ldots z^{(n)} \in \mathcal{Z} \) and a collection of functions \((\forall i) f_i : \mathbb{R}^p \times \mathcal{Z} \to \mathbb{R}\). Let:

\[
f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w, z^{(i)})
\]

Given two sets \( A, B \) and a scalar \( c \), define \( A + B = \{a + b \mid a \in A \text{ and } b \in B\} \) and \( cA = \{ca \mid a \in A\} \). Clearly:

\[
\partial f(w) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial f_i}{\partial w}(w, z^{(i)})
\]

Assume that we pick a data sample \( j \) uniformly at random from \( \{1 \ldots n\} \), and use the following stochastic subgradient:

\[
g_j(w) \in \frac{\partial f_j}{\partial w}(w, z^{(j)})
\]

Let \( j \) be a uniformly distributed random variable with support on \( \{1 \ldots n\} \), we can see that:

\[
\mathbb{E}_j[g_j(w)] = \sum_{i=1}^{n} \mathbb{P}_j[j = i] \ g_i(w)
\]

\[
\in \frac{1}{n} \sum_{i=1}^{n} \frac{\partial f_i}{\partial w}(w, z^{(i)})
\]

\[
= \partial f(w)
\]

Application 2: Stochastic Minibatch of Samples. Similarly, we can consider a minibatch instead of a single sample. Assume that we pick \( B \) data samples \( j_1 \ldots j_B \) independently (with replacement), uniformly at random from \( \{1 \ldots n\} \), and use the following stochastic subgradient:

\[
g_{j_1 \ldots j_B}(w) \in \frac{1}{B} \sum_{b=1}^{B} g_{j_b}(w)
\]
where \(g_{j_b}(w)\) was defined in eq.(4). Let \(j_1 \ldots j_B\) be i.i.d. uniformly distributed random variables with support on \(\{1 \ldots n\}\), we can see that:

\[
\mathbb{E}_{j_1 \ldots j_B} [g_{j_1 \ldots j_B}(w)] = \frac{1}{B} \sum_{b=1}^{B} \mathbb{E}_{j_b} [g_{j_b}(w)]
\]

\[
= \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} P_{j_b}[j_b = i] g_i(w)
\]

\[
\in \frac{1}{Bn} \sum_{b=1}^{B} \sum_{i=1}^{n} \frac{\partial f_i}{\partial w}(w, z^{(i)})
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \frac{\partial f_i}{\partial w}(w, z^{(i)})
\]

\[
= \partial f(w)
\]

**Application 3: Stochastic Coordinate.** Consider a general function \(f : \mathbb{R}^p \to \mathbb{R}\). Let \(e^{(k)}\) be the vector defined by \(e^{(k)}_i = 1[i = k]\). The gradient of \(f\) is:

\[
\partial f(w) = \begin{bmatrix}
\frac{\partial f}{\partial w_1}(w) \\
\vdots \\
\frac{\partial f}{\partial w_p}(w)
\end{bmatrix} = \sum_{k=1}^{p} e^{(k)} \frac{\partial f}{\partial w_k}(w)
\]

Assume that we pick a coordinate \(j\) uniformly at random from \(\{1 \ldots p\}\), and use the following stochastic subgradient:

\[
g_j(w) \in p e^{(j)} \frac{\partial f}{\partial w_j}(w)
\]

Let \(j\) be a uniformly distributed random variable with support on \(\{1 \ldots p\}\), we can see that:

\[
\mathbb{E}_j [g_j(w)] = \sum_{k=1}^{p} P_j[j = k] g_k(w)
\]

\[
\in \frac{1}{p} \sum_{k=1}^{p} p e^{(k)} \frac{\partial f}{\partial w_k}(w)
\]

\[
= \partial f(w)
\]

Clearly, we can also use the minibatch trick.

**Back to the General Problem.** Note that above, randomness does not come from uncertainty in the data but from generating a fast approximate version of the gradient. We will in general only assume that at every iteration \(t\):

\[
\mathbb{E}[g^{(t)}] \in \partial f(w^{(t)})
\]
Consider the next subgradient descent algorithm for solving eq.(1):

Algorithm 7.3 Stochastic subgradient descent algorithm

Input: Number of iterations \( T \geq 1 \), factor \( \beta > 0 \), initial point \( w^{(1)} \in \mathbb{R}^p \)
(The setting of \( w^{(1)} \) can be uninformed, e.g., \( w^{(1)} = 0 \).
The setting of \( w^{(1)} \) should be deterministic, i.e., non-stochastic.)

for \( t = 1 \ldots T - 1 \) do

\( w^{(t+1)} \leftarrow w^{(t)} - \eta_t g^{(t)} \) where \( E[g^{(t)}] \in \partial f(w^{(t)}) \)

end for

Output: \( \tilde{w}^{(T)} \leftarrow \frac{\sum_{t=1}^{T} \eta_t w^{(t)}}{\sum_{t=1}^{T} \eta_t} \)

In what follows, we state our main result regarding convergence rates for Algorithm 7.3.

Theorem 7.2. Assume that \( f : \mathbb{R}^p \rightarrow \mathbb{R} \) is convex and \( K \)-Lipschitz with respect to the \( \ell_2 \)-norm in the problem of eq.(1). Recall that \( \tilde{w} \) is the optimal solution of the problem of eq.(1). Assume that Algorithm 7.3 runs for a number of iterations \( T \), factor \( \beta \) and initial point \( w^{(1)} \), and that the algorithm outputs \( \tilde{w}^{(T)} \). Assume that the stochastic subgradients \( g^{(1)} \ldots g^{(T)} \) are independent and that they fulfill the condition \((\forall t) E[\|g^{(t)}\|_2^2] \leq K^2\). We have:

\[
E_{g^{(1)}\ldots g^{(T)}}[f(\tilde{w}^{(T)})] - f(\tilde{w}) \leq \frac{K\|w^{(1)} - \tilde{w}\|_2^2}{4\beta(\sqrt{T} - 1)} + \frac{\beta K (1 + \log T)}{4(\sqrt{T} - 1)}
\]

Fix \( \delta \in (0, 1) \). With probability at least \( 1 - \delta \) with respect to the choice of the stochastic subgradients \( g^{(1)} \ldots g^{(T)} \), we have:

\[
f(\tilde{w}^{(T)}) - f(\tilde{w}) \leq \frac{1}{\delta} \left( \frac{K\|w^{(1)} - \tilde{w}\|_2^2}{4\beta(\sqrt{T} - 1)} + \frac{\beta K (1 + \log T)}{4(\sqrt{T} - 1)} \right)
\]

Proof. Let \( g^{(1:1)} \equiv g^{(1)} \ldots g^{(j)} \). Note that \( w^{(t)} \) depend only on \( g^{(1:t-1)} \). Let:

\[
a^{(t)} = E_{g^{(1:t-1)}}[\|w^{(t)} - \tilde{w}\|_2^2]
\]

Note that since \( E[g^{(t)}] \in \partial f(w^{(t)}) \), by Definition 7.2 we have:

\[
f(\tilde{w}) - f(w^{(t)}) \geq E[g^{(t)}], \tilde{w} - w^{(t)}
\]

In fact, the above holds \( \forall \tilde{w} \in \mathbb{R}^p \) but in our problem we care about a unique \( \tilde{w} \). By assumption of the theorem, we know that \( g^{(1)} \ldots g^{(T)} \) are independent
and that they fulfill the condition \((\forall t) \mathbb{E} \left[ \|g(t)\|^2_2 \right] \leq K^2\). Therefore:

\[
d_{t+1} = \mathbb{E}_{g^{(1:t)}} \left[ \|w^{(t+1)} - \hat{w}\|^2_2 \right] \\
= \mathbb{E}_{g^{(1:t)}} \left[ \|w^{(t)} - \hat{w}\|^2_2 - 2\eta_t \langle g^{(t)}, w^{(t)} - \hat{w} \rangle + \eta_t^2 \|g^{(t)}\|^2_2 \right] \\
= \mathbb{E}_{g^{(1:t-1)}} \left[ \|w^{(t)} - \hat{w}\|^2_2 - 2\eta_t \mathbb{E}_{g^{(1:t)}} \left[ \langle g^{(t)}, w^{(t)} - \hat{w} \rangle \right] + \eta_t^2 \mathbb{E} \left[ \|g^{(t)}\|^2_2 \right] \right] \\
\leq a(t) - 2\eta_t \mathbb{E}_{g^{(1:t-1)}} \left[ \mathbb{E}_{g^{(t)}} \left[ \langle g^{(t)}, w^{(t)} - \hat{w} \rangle \right] \right] + \eta_t^2 K^2 \\
= a(t) - 2\eta_t \mathbb{E}_{g^{(1:t-1)}} \left[ \langle E_{g^{(t)}}, w^{(t)} - \hat{w} \rangle \right] + \eta_t^2 K^2 \\
\leq a(t) + 2\eta_t \mathbb{E}_{g^{(1:t-1)}} \left[ f(\hat{w}) - f(w^{(t)}) \right] + \eta_t^2 K^2 \\
= a(t) + 2\eta_t \left( f(\hat{w}) - \mathbb{E}_{g^{(1:t-1)}} [f(w^{(t)})] \right) + \eta_t^2 K^2
\]

Reorganizing the above, we obtain:

\[
\eta_t \left( \mathbb{E}_{g^{(1:t-1)}} [f(w^{(t)})] - f(\hat{w}) \right) \leq \frac{1}{2} \left( a(t) - a(t+1) + \eta_t^2 K^2 \right)
\]

Summing over \(t\), we get:

\[
\sum_{t=1}^{T} \eta_t \left( \mathbb{E}_{g^{(1:t-1)}} [f(w^{(t)})] - f(\hat{w}) \right) \leq \frac{1}{2} \sum_{t=1}^{T} \left( a(t) - a(t+1) + \eta_t^2 K^2 \right) \\
= \frac{1}{2} \left( a(1) - a(T+1) + K^2 \sum_{t=1}^{T} \eta_t^2 \right) \\
\leq \frac{1}{2} \left( \|w^{(1)} - \hat{w}\|^2_2 + \beta^2 \sum_{t=1}^{T} \frac{1}{t} \right) \\
\leq \frac{1}{2} \left( \|w^{(1)} - \hat{w}\|^2_2 + \beta^2 (1 + \log T) \right)
\]

where we used the fact that \(\sum_{t=1}^{T} 1/t \leq 1 + \log T\). By Jensen’s inequality and
convexity of $f$, we have:

\[
E_g^{(1:T)}[f(\tilde{w}(T)) - f(\hat{w})] = E_g^{(1:T)} \left[ f \left( \sum_{t=1}^{T} \eta_t w^{(t)} \right) \right] - f(\hat{w}) \\
\leq E_g^{(1:T)} \left[ \sum_{t=1}^{T} \eta_t f(w^{(t)}) \right] - f(\hat{w}) \\
= \sum_{t=1}^{T} \eta_t \left( E_g^{(t-1)}[f(w^{(t)})] - f(\hat{w}) \right) \\
\leq \frac{\sqrt{T}}{\beta K} \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \\
\leq \frac{1}{2} \left( \|w^{(1)} - \hat{w}\|_2^2 + \beta^2(1 + \log T) \right) \\
2 \frac{\beta}{\sqrt{T} - 1}
\]

where we used the fact that $2(\sqrt{T} - 1) \leq \sum_{t=1}^{T} 1/\sqrt{t}$. This proves our first claim.

Define the random variable $z = f(\tilde{w}(T)) - f(\hat{w})$. Note that $z \geq 0$ and that $E[z] = E[f(\tilde{w}(T))] - f(\hat{w})$. Recall that in the above we provided an upper bound $Z$ for $E[z]$. That is, we showed that $E[z] \leq Z$. By Markov’s inequality (Theorem 1.1), we have:

\[
P[z < Z/\delta] \geq P[z < E[z]/\delta] \\
= 1 - P[z \geq E[z]/\delta] \\
\geq 1 - \delta
\]

which proves our second claim. \qed

## References
