1 Deterministic Optimization

For brevity, everywhere differentiable functions will be called \textit{smooth}. Similarly, not everywhere differentiable functions will be called \textit{nonsmooth}.

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

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

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

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

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

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

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

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

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

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

Clearly, in the above definition, if $\phi : \mathbb{R}^p \rightarrow \mathbb{R}$ 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:

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

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 \textit{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:

\textbf{Algorithm 7.1} Subgradient descent algorithm

\begin{itemize}
  \item \textbf{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., \( w^{(1)} = 0 \))
  \item \textbf{for} \( t = 1 \ldots T - 1 \) \textbf{do}
    \begin{itemize}
      \item \( w^{(t+1)} \leftarrow w^{(t)} - \eta_t g^{(t)} \) where \( g^{(t)} \in \partial f(w^{(t)}) \)
    \end{itemize}
  \textbf{end for}
  \item \textbf{Output:} \( \tilde{w}^{(T)} \leftarrow \frac{\sum_{t=1}^{T} \eta_t w^{(t)}}{\sum_{t=1}^{T} \eta_t} \)
\end{itemize}

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

\textbf{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:

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

\textbf{Proof.} Let \( a^{(t)} \equiv \|w^{(t)} - \hat{w}\|_2^2 \). Note that since \( g^{(t)} \in \partial f(w^{(t)}) \), by Definition 7.2 we have:

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

In fact, the above holds \( \forall \hat{w} \in \mathbb{R}^p \) but in our problem we care about a unique \( \hat{w} \).

By the Lipschitz continuity of \( f \), we know that \( \forall t \|g^{(t)}\|_2 \leq K \). Therefore:

\[
a^{(t+1)} = \|w^{(t+1)} - \hat{w}\|_2^2
\]

\[
= \|w^{(t)} - \hat{w}\|_2^2 - 2\eta_t \langle g^{(t)}, w^{(t)} - \hat{w} \rangle + \eta_t^2 \|g^{(t)}\|_2^2
\]

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

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

Summing over $t$, we get:

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

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

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

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

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:

$$f(\bar{w}^{(T)}) - f(\bar{w}) = f \left( \frac{\sum_{t=1}^{T} \eta_t w^{(t)}}{\sum_{t=1}^{T} \eta_t} \right) - f(\bar{w})$$

$$\leq \frac{\sum_{t=1}^{T} \eta_t f(w^{(t)})}{\sum_{t=1}^{T} \eta_t} - f(\bar{w})$$

$$= \frac{\sum_{t=1}^{T} \eta_t \left( f(w^{(t)}) - f(\bar{w}) \right)}{\sum_{t=1}^{T} \eta_t}$$

$$= \frac{\sum_{t=1}^{T} \eta_t \left( f(w^{(t)}) - f(\bar{w}) \right)}{\beta \sum_{t=1}^{T} \frac{1}{\sqrt{t}}}$$

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

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

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

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

2 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 collec-
tion of functions (∀i) \( f_i : \mathbb{R}^p \times \mathcal{Z} \to \mathbb{R} \). Let:

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

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:

\[
g_j(w) = \frac{\partial f_j}{\partial w}(w, z^{(j)}) \quad (2)
\]

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

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

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

\[
= \partial f(w)
\]

**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:

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

where \( g_{j_b}(w) \) was defined in eq.(2). Let \( j_1 \ldots j_B \) be i.i.d. uniformly distributed random variables with support on \( \{1 \ldots n\} \), we can see that:

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

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

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

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

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

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

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

$$
g_j(w) = p e^{(j)} \frac{\partial f}{\partial w_j}(w)
$$

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

$$
E_j[g_j(w)] = \sum_{k=1}^{p} P_j[j = k] g_k(w) \\
= \frac{1}{p} \sum_{k=1}^{p} p e^{(k)} \frac{\partial f}{\partial w_k}(w) \\
= \frac{\partial f}{\partial w}(w)
$$

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 $t$:

$$
E[g^{(t)}] \in \partial f(w^{(t)})
$$

Consider the next subgradient descent algorithm for solving eq.(1):

**Algorithm 7.2** Stochastic subgradient descent algorithm

<table>
<thead>
<tr>
<th>Input:</th>
<th>Number of iterations $T \geq 1$, factor $\beta &gt; 0$, initial point $w^{(1)} \in \mathbb{R}^p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(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.)</td>
<td></td>
</tr>
<tr>
<td>for</td>
<td>$t = 1 \ldots T - 1$ do</td>
</tr>
<tr>
<td></td>
<td>$w^{(t+1)} \leftarrow w^{(t)} - \eta \theta g^{(t)}$ where $E[g^{(t)}] \in \partial f(w^{(t)})$</td>
</tr>
<tr>
<td>end for</td>
<td></td>
</tr>
<tr>
<td>Output:</td>
<td>$\bar{w}^{(T)} \leftarrow \frac{\sum_{t=1}^{T} \eta_t w^{(t)}}{\sum_{t=1}^{T} \eta_t}$</td>
</tr>
</tbody>
</table>

In what follows, we state our main result regarding convergence rates for Algorithm 7.2.
Theorem 7.2. Assume that $f : \mathbb{R}^p \to \mathbb{R}$ 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 7.2 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 $(\forall t) \mathbb{E} \left[ \|g^{(t)}\|^2 \right] \leq K^2$. We have:

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

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:

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

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

$$
a^{(t)} \equiv \mathbb{E}_{g^{(1:t-1)}} \left[ \|w^{(t)} - \tilde{w}\|^2 \right]
$$

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

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

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

$$
a^{(t+1)} = \mathbb{E}_{g^{(1:t)}} \left[ \|w^{(t+1)} - \tilde{w}\|^2 \right]
$$

$$
= \mathbb{E}_{g^{(1:t)}} \left[ \|w^{(t)} - \tilde{w}\|^2 - \eta_t g^{(t)} \right] + \eta_t^2 \mathbb{E}_{g^{(1:t)}} \left[ \|g^{(t)}\|^2 \right]
$$

$$
= \mathbb{E}_{g^{(1:t-1)}} \left[ \|w^{(t)} - \tilde{w}\|^2 - 2\eta_t \langle g^{(t)}, w^{(t)} - \tilde{w} \rangle + \eta_t^2 \mathbb{E}_{g^{(1:t)}} \left[ \|g^{(t)}\|^2 \right] \right]
$$

$$
\leq a^{(t)} - 2\eta_t \mathbb{E}_{g^{(1:t-1)}} \left[ \mathbb{E}_{g^{(t)}} \left[ \langle g^{(t)}, w^{(t)} - \tilde{w} \rangle \right] \right] + \eta_t^2 K^2
$$

$$
= a^{(t)} - 2\eta_t \mathbb{E}_{g^{(1:t-1)}} \left[ \mathbb{E}_{g^{(t)}} \left[ g^{(t)} \right], w^{(t)} - \tilde{w} \right] + \eta_t^2 K^2
$$

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

Reorganizing the above, we obtain:

$$
\eta_t \left( \mathbb{E}_{g^{(1:t-1)}} \left[ f(w^{(t)}) \right] - f(\tilde{w}) \right) \leq \frac{1}{2} \left( a^{(t)} - a^{(t+1)} + \eta_t^2 K^2 \right)
$$
Summing over $t$, we get:

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

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

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

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

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:

$$
\mathbb{E}_{g^{(1:T)}}[f(\hat{w}^{(T)})] - f(\hat{w}) = \mathbb{E}_{g^{(1:T)}} \left[ f \left( \frac{\sum_{t=1}^{T} \eta_t w^{(t)}}{\sum_{t=1}^{T} \eta_t} \right) \right] - f(\hat{w})
$$

$$
\leq \mathbb{E}_{g^{(1:T)}} \left[ \frac{\sum_{t=1}^{T} \eta_t f(w^{(t)})}{\sum_{t=1}^{T} \eta_t} \right] - f(\hat{w})
$$

$$
= \sum_{t=1}^{T} \eta_t \left( \mathbb{E}_{g^{(t:t-1)}}[f(w^{(t)})] - f(\hat{w}) \right)
$$

$$
= \sum_{t=1}^{T} \eta_t \left( \frac{\beta}{K} \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \right)
$$

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

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

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

$$
P[z < Z/\delta] \geq P[z < \mathbb{E}[z]/\delta]
$$

$$
= 1 - P[z \geq \mathbb{E}[z]/\delta]
$$

$$
\geq 1 - \delta
$$

which proves our second claim.  \hfill \Box
References
