Gradient Descent

We start by looking at convex functions that are Lipschitz cont.

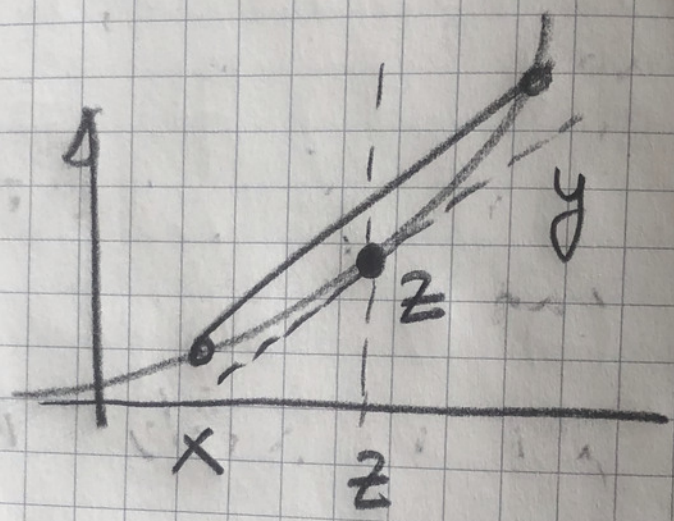
__Def:__ Let S be an open and convex set. Let $f : S \to \mathbb{R}$.
We say that f is convex if for all $x, y \in S$ and $\lambda \in [0,1]$

$$f(z) \leq \lambda f(x) + \bar{\lambda} f(y) \text{ where}$$
$$z = \lambda x + \bar{\lambda} y.$$

An alternative characterisation is the following.

__Lemma:__ Let S be an open and convex set. Let $f : S \to \mathbb{R}$.
Then f is convex if $\forall z \in S$ there exists $\underline{v}$ so that $\forall x \in S$

$$f(x) \geq f(z) + \langle v, x - z \rangle . \quad (*)$$


y
z
x
z

Proof:

Assume the statement of the Lemma. Let $z = \lambda x + \bar{\lambda} y$ for some $x, y \in S$ and $\lambda \in [0,1]$. By the lemma

$$f(x) \geq f(z) + \langle v_z, x - z \rangle \quad | \ \lambda$$
$$f(y) \geq f(z) + \langle v_z, y - z \rangle \quad | \ \bar{\lambda}$$

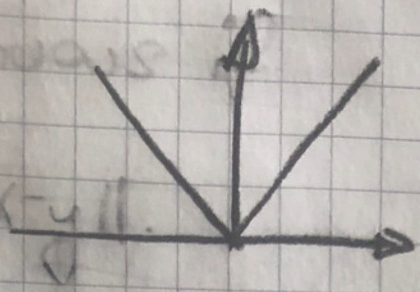
$\lambda f(x) + \bar{\lambda} f(y) \geq f(z)$, hence convex according to def.

Conversely, assume def.

If f has gradient at z then clear that the promised v_z is the gradient of f at z and is unique.

Example:

$$f(x) = |x|$$



$$\partial f(x) = \begin{cases} \{-1\}, & x < 0 \\ \{1\}, & x > 0 \\ [-1,1], & x = 0 \end{cases}$$

Example: Let $g_1, \ldots, g_r$ be convex differtible functions. Let

$$g(z) = \max_{i \in [r]} g_i(z)$$

Let $j = \operatorname{argmax}_i g_i(z)$. Then

$$\nabla g_j(z) \subseteq \partial g(z).$$

Subgradient:

Consider the previous lemma. Any v_z that fulfills (*) is called a subgradient of f at z. If f is differentiable at z then v_z is unique. We say $v_z \in \partial f(z)$ and call $\partial f(z)$ the differential set.

We have for all $x \in S$, (54)

$$
\begin{aligned}
g(x) &\geq g_j(x) \\
&\geq \underset{\parallel}{g_j(z)} + \langle \nabla g_j(z), x - z \rangle \\
&= g(z) + \langle \nabla g_j(z), x - z \rangle
\end{aligned}
$$

$\blacksquare$

We say that f is ρ-Lipschitz if for all $x, y \in S$
$|f(x) - f(y)| \leq \rho \|x - y\|$.

__Lemma:__ Let S be an open convex set and let $f: S \to \mathbb{R}$ be a convex function. Then f is ρ-Lipschitz __iff__ for all $z \in S$ and $v_z \in \partial f(z)$, $\|v_z\| \leq \rho$.

__Proof:__ Assume that for all $z \in S$ and $v_z \in \partial f(z)$, $\|v_z\| \leq \rho$.
Then, since f is convex and v_z is a subgradient of z, $\forall x \in S$

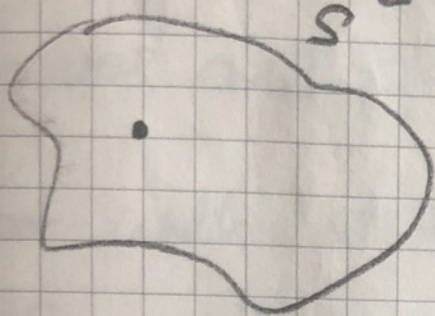
$$f(x) \geq f(z) + \langle v_z, x - z \rangle$$

$$f(z) - f(x) \leq \langle v_z, z - x \rangle$$

$$\leq \rho \|z - x\|$$

We get the second inequality by switching roles of x and z.

For the converse assume that $z \in S$ and $v_z \in \partial f(z)$.



Let $x = z + \varepsilon \frac{v_z}{\|v_z\|}$.

Then $\|x - z\| = \varepsilon$. By Lipschitz

$$|f(x) - f(z)| \leq \rho \|x - z\| = \varepsilon \rho.$$

By convexity,

$$f(x) \geq f(z) + \langle v_z, x - z \rangle$$

$$\varepsilon \rho \geq f(x) - f(z) \geq \|v_z\| \varepsilon$$

Hence $\rho \geq \|v_z\|$.

Let us now look at the GD algorithm applied to a convex and ρ-Lipschitz function f.

Start with $w^{(1)} = 0$. Then at each step $w^{(t)} = w^{(t-1)} - \eta \nabla f(w^{(t-1)})$, for some step size η. If the function f does not have a gradient at $w^{(t)}$ then pick any subgradient from $\partial f(w^{(t)})$. After the step T, output the estimate

$$\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w^{(t)}.$$

How close will we be to the optimal value?

<u>Lemma:</u> Apply the above algorithm. Let w^* be

$$w^* = \underset{w : \|w\| \leq B}{\operatorname{argmin}} f(w). \text{ Then}$$

$$f(\bar{w}) - f(w^*) \leq \frac{B\rho}{\sqrt{T}}.$$

In other words, we need at most $\frac{B^2\rho^2}{\varepsilon^2}$ steps to get ε close.

Here the stepsize is

$$\eta = \sqrt{\frac{B^2}{\rho^2 T}}.$$

Proof:

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

$$\overset{\text{convexity}}{\leq} \frac{1}{T}\sum_{t=1}^{T}\left(f(w^{(t)}) - f(w^*)\right)$$

$$\overset{\text{convex}}{\leq} \frac{1}{T}\sum_{t=1}^{T} \langle w^{(t)} - w^*, \nabla f(w^{(t)}) \rangle$$

$$\overset{\text{see the next page}}{\leq} \frac{1}{T}\left[\frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\nabla f(w^{(t)})\|^2\right]$$

$$\leq \frac{1}{T}\left[\frac{B^2}{2\eta} + \frac{\eta}{2} T \rho^2\right]$$

$$= \frac{B^{\cancel{2}} \rho \sqrt{T}}{2T \cancel{B}} + \frac{B \cancel{T}}{2\rho\sqrt{T}\cancel{T}} \rho^{\cancel{2}}$$

$$= \frac{\rho B}{2\sqrt{T}} + \frac{\rho B}{2\sqrt{T}} = \frac{\rho B}{\sqrt{T}}.$$

Note: We want both error terms to be of equal size

$$\Longrightarrow \eta = \sqrt{\frac{B^2}{\rho^2 T}}$$

$$\frac{1}{T\eta}\sum_{t=1}^{T} \langle w^{(t)} - w^*, \eta \nabla f(w^{(t)}) \rangle$$

$$= \frac{1}{2T\eta}\sum_{t=1}^{T} -\|w^{(t)} - w^* - \eta \nabla f(w^{(t)})\|^2 + \|w^{(t)} - w^*\|^2 + \eta^2 \|\nabla f(w^{(t)})\|^2$$

$$\overset{\text{GD step}}{=} \frac{1}{2T\eta}\sum_{t=1}^{T} -\|w^{(t+1)} - w^*\|^2 + \|w^{(t)} - w^*\|^2 + \eta^2 \|\nabla f(w^{(t)})\|^2$$

$$\overset{w^{(1)}=0}{=} \frac{1}{2T\eta}\left(\|w^{(1)} - w^*\|^2 - \|w^{(T+1)} - w^*\|^2\right) + \frac{\eta}{2T}\sum_{t=1}^{T} \|\nabla f(w^{(t)})\|^2$$

$$\leq \frac{1}{2T\eta}\|w^*\|^2 + \frac{\eta}{2T}\sum_{t=1}^{T} \|\nabla f(w^{(t)})\|^2$$

<u>Note:</u> $\langle a, b \rangle = -\frac{\|a-b\|^2}{2} + \frac{\|a\|^2}{2} + \frac{\|b\|^2}{2}$

parallelogram law