$$f\left(\frac{1}{t}\sum_{s=0}^{t-1} x^{(s)}\right) - f(x^*) \leq \frac{1}{2\mu t}R^2 + \frac{\mu}{2}L^2$$

$$\leq \frac{R^2 L}{2C\sqrt{t}} + \frac{CL}{2\sqrt{t}} = \frac{1}{\sqrt{t}}(\sim)$$

(Still $\sim \frac{1}{\sqrt{t}}$)

An interpretation of GD

Recall that by Taylor:

$$\underbrace{f(z)}_{\text{goal is to min. } f(z)} = f(x^{(t)}) + \nabla f(x^{(t)})^T (z - x^{(t)})$$

$$+ \frac{1}{2}(z - x^{(t)})^T \nabla^2 f(\xi)(z - x^{(t)})$$

ξ is between $x^{(t)}$ & z

GD approximates $\nabla^2 f(\xi)$ by

$\frac{1}{\mu} I$

$\uparrow$ Identity matrix

$$f(z) \approx \underbrace{f(x^{(t)}) + \nabla f(x^{(t)})^T (z - x^{(t)}) + \frac{1}{2\mu} \|z - x^{(t)}\|^2}_{g(z)}$$

Notice that $g(z)$ is convex so to minimize it we want to solve

$$\nabla g(z^*) = 0$$

$$\iff \nabla f(x^{(t)}) + \frac{1}{\mu}(z^* - x^{(t)}) = 0$$

$$\Rightarrow z^* = x^{(t)} - \mu \nabla f(x^{(t)})$$

set $x^{(t+1)}$ to this, and repeat

$\Rightarrow$ GD!

Another interpretation of GD & the method of Steepest descent:

One way we can think of GD, is that we pick at each step, $t+1$, $x^{(t+1)}$

so that

$$f(x^{(t+1)}) \approx f(x^{(t)}) - \mu \|\nabla f(x^{(t)})\|_2^2$$

(bec, in general, $f(z) \approx f(x^{(t)}) - \mu \nabla f(x^{(t)})^T (z - x^{(t)})$

& in GD, $x^{(t+1)} = x^{(t)} - \mu \nabla f(x^{(t)})$)

Another idea: think of GD as picking a direction that minimizes the function $h(\vec{p}) = \nabla f(x^{(t)})^T \frac{p}{\|p\|}$

In other words, if I think of algorithms that update x via

$$x^{(t+1)} = x^{(t)} - \mu \vec{p}$$

GD is the one that picks $\vec{p}$ so that $h(p)$ is minimized

i.e. we solve

$$\min_{p} h(p) \iff \min_{p} \frac{\nabla f(x^{(t)})^T \vec{p}}{\|p\|_2}$$

$$\|\vec{p}\|_2 = \sqrt{\sum_{i=1}^{N} p_i^2}$$

This suggests that we can come up with other algorithms by replacing $\|p\|_2$ in the denominator, by other norms.

Def'n: Let $x \in \mathbb{R}^N$, then we define

$$\|x\|_1 := \sum_{i=1}^{N} |x_i| \rightarrow \text{1-norm or } \ell_1\text{-norm}$$

$$\|x\|_2 := \sqrt{\sum_{i=1}^{N} |x_i|^2} \rightarrow \text{2-norm or } \ell_2\text{-norm}$$

$$\|x\|_\infty = \max_i |x_i|$$

more generally for $p \in (1, \infty)$

$$\|x\|_p = \left(\sum_{i=1}^{p} |x_i|^p\right)^{1/p}$$

Now, consider algorithms of the type

$$x^{(t+1)} = x^{(t)} - \mu p$$

where p minimizes $\dfrac{\nabla f(x^{(t)})^T p}{\|p\|_1}$

or $\dfrac{\nabla f(x^{(t)})^T p}{\|p\|_\infty}$ (or some other norm)

For 1-norm:

$$\min_p \frac{\nabla f(x^{(t)})^T p}{\|p\|_1} \implies$$

$$p^* = -\operatorname{sign}\left(\frac{\partial f}{\partial x_{j^*}}\right) \begin{array}{r} \\ \\ \\ \\ j^{*\text{th}} \text{ entry} \rightarrow \\ \\ \end{array} \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ \delta \\ \vdots \\ \vdots \end{pmatrix}$$

↳ a vector

where $\delta = \|\nabla f(x)\|_\infty$
$= \left|\dfrac{\partial f}{\partial x_{j^*}}(x^{(t)})\right|$

where j^* indexes the entry of $\nabla f(x^{(t)})$ with the largest magnitude.

<u>Example</u>: suppose

$$\nabla f(x^{(t)}) = \begin{pmatrix} 1 \\ -2 \\ -0.5 \end{pmatrix} \leftarrow$$

then $$P^* = \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix}$$

This algorithm is called "coordinate descent" (move along one coordinate at each iteration)

For ∞-norm:

$$\min_{p} \frac{\nabla f(x^{(t)})^T p}{\|p\|_\infty}$$

$$\Rightarrow p^* = -\delta \begin{pmatrix} \text{sign}(\nabla_1 f(x^{(t)})) \\ \text{sign}(\nabla_2 f(x^{(t)})) \\ \vdots \\ \text{sign}(\nabla_n f(x^{(t)})) \end{pmatrix}$$

where $\delta = \|\nabla f(x^{(t)})\|_1$

So $p^* = -\|\nabla f(x^{(t)})\|_1 \begin{pmatrix} \pm 1 \\ \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{pmatrix}$

Example: suppose

$$\nabla f(x^{(t)}) = \begin{pmatrix} 1 \\ -2 \\ -0.5 \end{pmatrix}$$

then $$p^* = -3.5 \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}.$$

(GD would have given

$$p^* = \begin{pmatrix} -1 \\ 2 \\ 0.5 \end{pmatrix}$$
)

Aside: the proof of the optimality of the p^* chosen in the examples above relies on Hölder's inequality:

$$|x^T y| \leq \|x\|_\infty \|y\|_1 \quad \left\{ |x^T y| \leq \|x\|_2 \|y\|_2 \right.$$

& $$|x^T y| \leq \|x\|_1 \|y\|_\infty$$