Gradient Descent (Cont'd)

Recall: $x^{(t+1)} = x^{(t)} - \mu^{(t)} \nabla f(x^{(t)})$

we had two questions:

(2) When do we stop?
(we now have nec. & suff. cond. for optimality).

(1) How do we choose $\mu^{(t)}$?

We now focus on this question

Along the way, we will develop an understanding of the rate of convergence of gradient descent.

(i.e. how does the error: $f(x^{(t)}) - f(x^*)$ behave as a function of the iteration t).

We will need a def'n:

Def'n: A function $f: \Omega \subset \mathbb{R}^n \to \mathbb{R}$ is *L-Lipschitz* if $\forall x, y \in \Omega$

$$|f(x) - f(y)| \leq L \|x - y\|$$

So how much f changes between x & y is controlled by how far x & y are from each other!

Example: $f: \mathbb{R} \to \mathbb{R}$

$f(x) = |x|$ is Lipschitz because

$$|f(x) - f(y)| = \big||x| - |y|\big| \leq |x - y|$$

(why?)

and its Lipschitz constant is one.

Fact: If f is L-lipschitz, convex, and differentiable, then

$$\|\nabla f(x)\| \leq L \quad \forall x \in \Omega$$

proof: For any x, y:

convex $\Rightarrow f(x) - f(y) \geq \nabla f(y)^T (x-y)$

Lipschitz $\Rightarrow L\|x-y\| \geq |f(x)-f(y)|$

$\forall x, y$: $\geq |\nabla f(y)^T (x-y)|$

Pick $x = y + \nabla f(y)$ & sub. in above eq.

$\Rightarrow L\|\nabla f(y)\| \geq |\nabla f(y)^T \nabla f(y)|$

so $L\|\nabla f(y)\| \geq \|\nabla f(y)\|^2$

so $\|\nabla f(y)\| \leq L$.



Theorem: $\|\nabla f(x)\| \leq L \quad \forall x \in \mathbb{R}^n$

$\Rightarrow |f(x) - f(y)| \leq L\|x-y\|$

↳ Bounded gradient $\Rightarrow$ Lipschitz

Proof: By Taylor:

$f(x) - f(y) = \nabla f(tx + (1-t)y)^T (x-y)$
for some $t \in (0,1)$

$\Rightarrow |f(x) - f(y)| = |\nabla f(tx + (1-t)y)^T (x-y)|$

$\leq \|\nabla f(tx + (1-t)y)\| \|(x-y)\|$

bec $|a^T b| \leq \|a\| \|b\|$

$\leq L\|x-y\|$

↳ bec all gradients are bounded (by hypothesis).

∎

Choice of μ & convergence of gradient descent:

Theorem: Let f be convex, differentiable, L-Lipschitz

Let $\begin{cases} \|x^{(0)} - x^*\| \leq R \\ \|\nabla f(x)\| \leq L \end{cases}$

& let $\mu = \frac{R}{L\sqrt{t}}$

Then $f\left(\frac{1}{t}\sum_{s=0}^{t-1} x^{(s)}\right) - f(x^*) \leq \frac{RL}{\sqrt{t}}$

($f(x^*)$: optimal value; left-hand side: gap to optimal value)

Convergence rate theorem

So, if we plan to run GD for t steps, we can pick $\mu = \frac{R}{L\sqrt{t}}$ (so it is constant) and after t-steps the

average of our iterates

$\frac{1}{t}\sum_{s=0}^{t-1} x^{(s)}$ gives an error

$$f(\text{ave}) - f(x^*) \leq RL/\sqrt{t}$$

Example: Consider the function

$f(x_1, x_2) = \sin(x_1) + x_2$

and notice that $\nabla f(x_1, x_2) = \begin{pmatrix} \cos(x_1) \\ 1 \end{pmatrix}$

so that $\|\nabla f\| \leq \sqrt{\cos^2(x_1) + 1} \leq \sqrt{2}$

so f is Lipschitz with $L = \sqrt{2}$

$\Rightarrow$ we can run GD for t-steps

with $\mu = \frac{R}{\sqrt{2}\sqrt{t}}$ and we'll get

within $\frac{R\sqrt{2}}{\sqrt{t}}$ of a local minimum

(In practice, we won't know R, more later)

(We can do better if we assume more about the function, more later)

Proof of convergence rate theorem

Convexity implies that

$$f(x^*) \geq f(x^{(s)}) + \langle \nabla f(x^{(s)}), x^* - x^{(s)} \rangle$$

$$\Leftrightarrow \boxed{f(x^{(s)}) - f(x^*) \leq \langle \nabla f(x^{(s)}), x^{(s)} - x^* \rangle} \quad (1)$$

But by GD: $x^{(s+1)} = x^{(s)} - \mu \nabla f(x^{(s)})$

$$\Rightarrow \boxed{\nabla f(x^{(s)}) = \frac{x^{(s)} - x^{(s+1)}}{\mu}} \quad (2)$$

So that

(1) & (2) $\Rightarrow$

$$\boxed{f(x^{(s)}) - f(x^*) \leq \frac{1}{\mu} \langle x^{(s)} - x^{(s+1)}, x^{(s)} - x^* \rangle} \quad (3)$$

we need the simple fact that for $a, b \in \mathbb{R}^n$

$$\langle a, b \rangle = \frac{\|a\|^2 + \|b\|^2 - \|a-b\|^2}{2}$$

Use this in (3) to get:

$$f(x^{(s)}) - f(x^*) \leq \frac{1}{2\mu} \left(\underbrace{\|x^{(s)} - x^*\|^2}_{b} + \underbrace{\|x^{(s)} - x^{(s+1)}\|^2}_{a} - \underbrace{\|x^{(s+1)} - x^*\|^2}_{a-b} \right)$$

and Recall that $x^{(s)} - x^{(s+1)} = \mu \nabla f(x^{(s)})$ to get

$$\boxed{f(x^{(s)}) - f(x^*) \leq \frac{1}{2\mu} \left(\|x^{(s)} - x^*\|_2^2 - \|x^{(s+1)} - x^*\|_2^2 \right) + \frac{\mu}{2} \|\nabla f(x_s)\|^2}$$

(4)

Summing from $s = 0$ to $t-1$:

telescoping sum

$$\sum_{s=0}^{t-1}\left(f(x^{(s)}) - f(x^*)\right) \leq \frac{1}{2\mu}\left(\underbrace{\|x^{(0)}-x^*\|_2^2}_{\leq R^2} - \underbrace{\|x^{(t)}-x^*\|^2}_{\geq 0}\right) + \mu/2 \sum_{s=0}^{t-1} \underbrace{\|\nabla f(x^{(s)})\|^2}_{\leq L^2}$$

$$\leq \frac{1}{2\mu}\left(R^2\right) + \frac{\mu}{2} L^2 t$$

divide both sides by t

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

now, by convexity

$$f\left(\sum_{s=0}^{t-1} \frac{1}{t} x^{(s)}\right) \leq \sum_{s=0}^{t-1} \frac{1}{t} f(x^{(s)})$$

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

$$\leq \frac{RL}{\sqrt{t}} \qquad \text{dec } \mu = \frac{R}{L\sqrt{t}}$$

■

Remark: In reality, we won't know the value of $\|x^{(0)} - x^*\| = R$ so when we pick μ, we can choose $\mu = \frac{C}{L\sqrt{t}}$ & we get (C: some small const.) an error bound

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

$$\leq \frac{R^2 L}{2C\sqrt{t}} + \frac{CL}{2\sqrt{t}}$$

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

Remark: Since the error decays like $\frac{1}{\sqrt{t}}$, if we want an error of ε, then we need something like $t \approx \varepsilon^{-2}$ iterations.

Remark: The convergence rate is independent of n!

A useful interpretation of GD:

For smooth enough functions we have, by Taylor:

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

(ξ between z & $x^{(t)}$)

↓

Want to minimize this

approx $\nabla^2 f(\xi)$ by $\frac{1}{\mu}$

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

& notice that $g(z)$ is convex, so to minimize it, we

Want

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

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

$0 =$

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

set $x^{(t+1)}$ to this!

Pros & Cons of GD:

* Simple: it only requires computing a gradient
* Each iteration is cheap (see above)
* Can be slow to converge
* Can't handle non-differentiable functions.

Gradient Descent when we have constraints:

we have been solving the opt. problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

by running the GD iterations

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

But what changes should we make to GD if we want to solve

$$\min f(x) \quad \text{subject to} \quad x \in \Omega$$

where Ω is a strict subset of $\mathbb{R}^n$?

In particular, the issue is that even if an iterate $x^{(t)} \in \Omega$

there is no guarantee that

$x^{(t+1)} = x^{(t)} - \mu \nabla F(x^{(t)}) \in \Omega$.

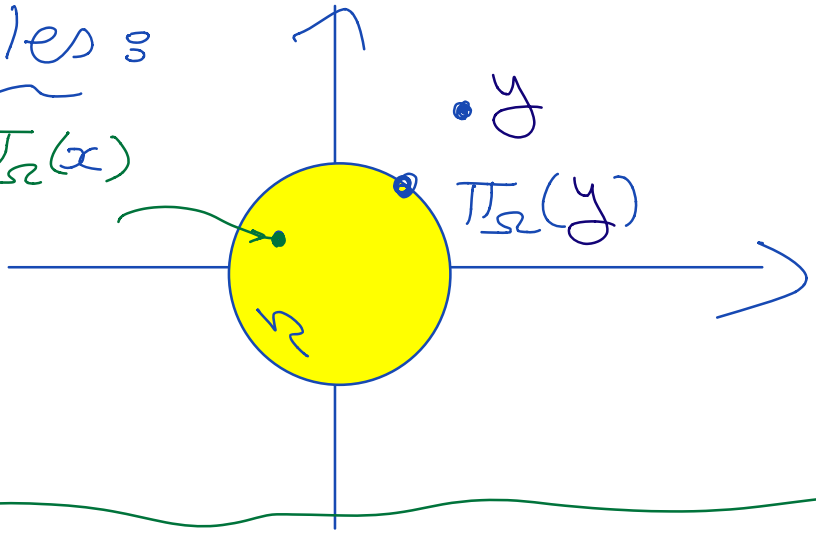
Def'n: The projection of a point x onto a set Ω is defined as

$$\Pi_{\Omega}(x) = \underset{y \in \Omega}{\operatorname{argmin}} \|x - y\|$$

↳ $\Pi_{\Omega}(x)$ is the closest pt to x that lives in Ω.

Examples:

$x = \Pi_{\Omega}(x)$

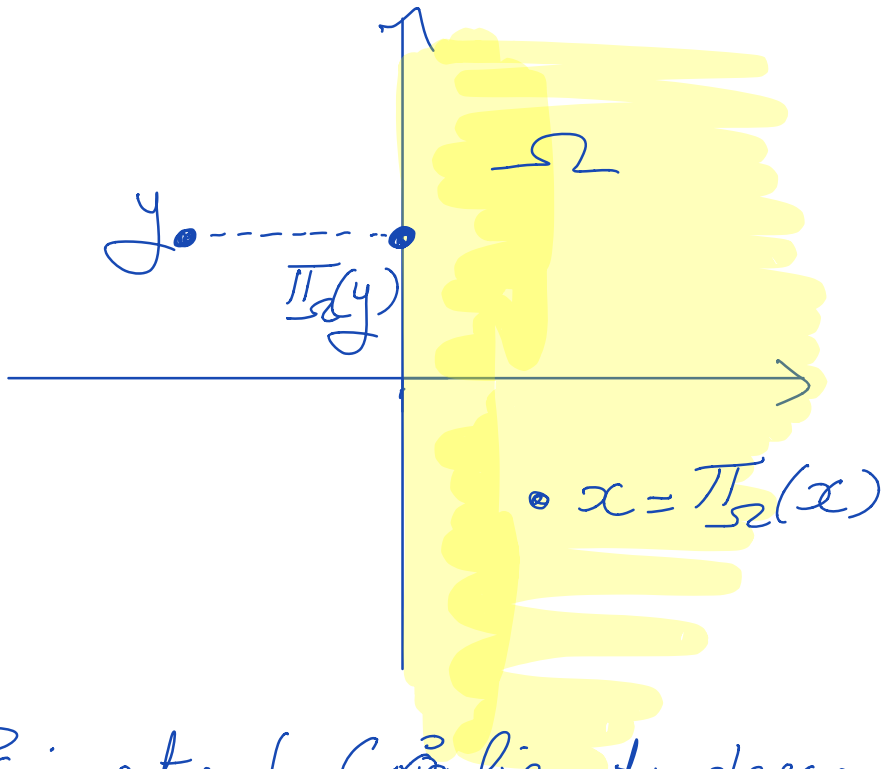


If

$$\Omega = \{x : \|x\| \leq 1\} \Rightarrow \Pi_{\Omega}(x) = \begin{cases} x, & \|x\| \leq 1 \\ \frac{x}{\|x\|}, & \|x\| > 1 \end{cases}$$

If $\Omega = \{x : x_1 \geq 0\}$

then $\Pi_\Omega(x) = \begin{cases} x & \text{if } x_1 \geq 0 \\ (0, x_2, x_3, \ldots x_n) & \text{if } x_1 < 0 \end{cases}$



Projected Gradient descent:

To solve: $\min_{x \in \Omega} f(x)$ where $\Omega \subset \mathbb{R}^n$ is convex, we run the iteration

$$x^{(t+1)} = \Pi_{\Omega}\left(x^{(t)} - \mu^{(t)} \nabla f(x^{(t)})\right)$$

Regular GD step

projected onto Ω

This is equivalent to doing:

$$y^{(t+1)} = x^{(t)} - \mu^{(t)} \nabla f(x^{(t)})$$

$$x^{(t+1)} = \underset{x \in \Omega}{\operatorname{argmin}} \| y^{(t+1)} - x \|$$

In general $x^{(t+1)} = \underset{x \in \Omega}{\operatorname{argmin}} \| y^{(t+1)} - x \|$

$$= \Pi_{\Omega}(y^{(t+1)})$$

is itself a convex optimization problem. (!)

But, in many cases, this has a closed form solution.

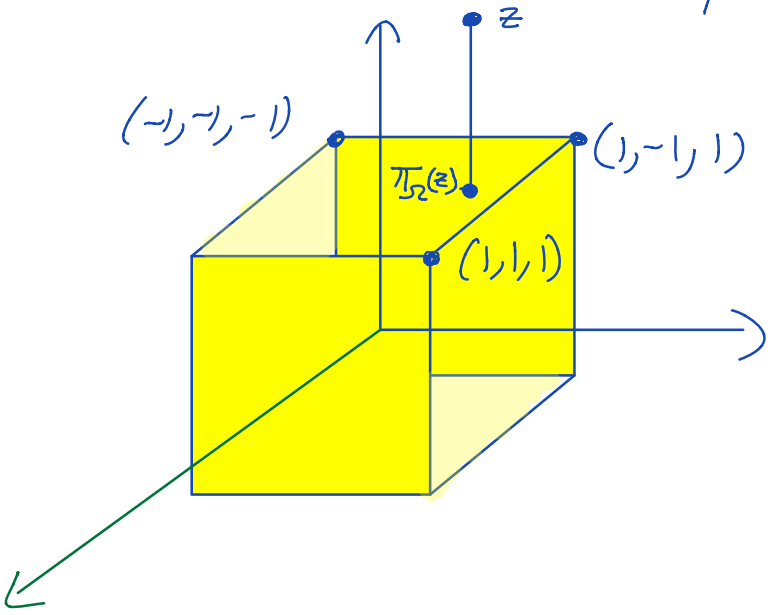
Example: We've seen a couple already

(1) $\Omega = \{ x \in \mathbb{R}^n : \|x\| \leq 1 \} \Rightarrow$

$$\Pi_\Omega(x) = \begin{cases} x, & x \in \Omega \\ \dfrac{x}{\|x\|_2}, & x \notin \Omega \end{cases}$$

(2) $\Omega = \{ x \in \mathbb{R}^n : |x_i| \leq 1 \}$

$$\Pi_\Omega(x) = \begin{cases} x, & x \in \Omega \\ y : y_i = x_i \text{ if } |x_i| \leq 1 \\ \quad y_i = \dfrac{x_i}{|x_i|} \text{ if } |x_i| > 1 \end{cases}$$



Some Facts

Lemma:

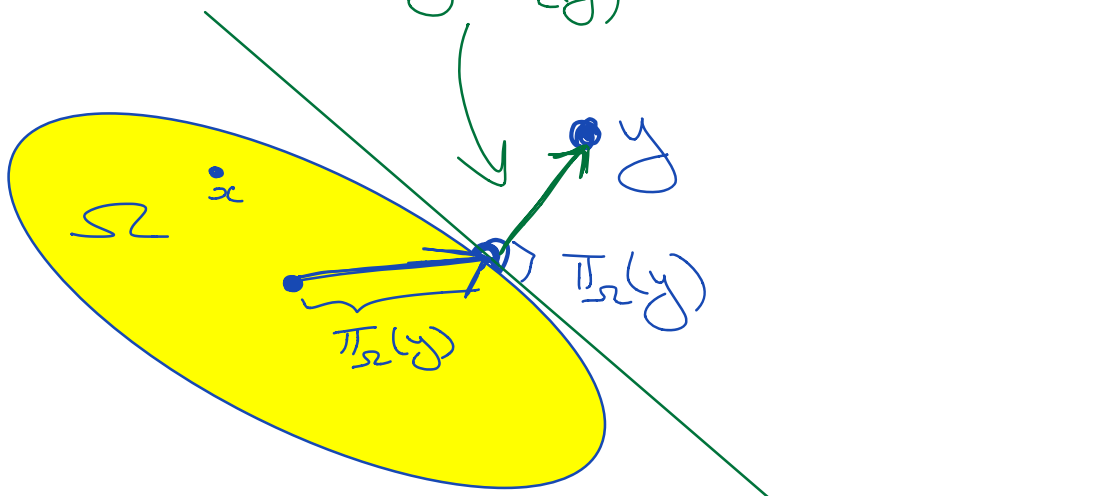
If $\Omega \subset \mathbb{R}^n$ is convex & closed (& not empty) and if $x \in \Omega$, $y \in \mathbb{R}^n$

then (I) $\pi_\Omega(y)$ is unique

(II) $\pi_\Omega(\pi_\Omega(y)) = \pi_\Omega(y)$

(III) $\langle \pi_\Omega(y) - x,\ \pi_\Omega(y) - y \rangle \leq 0$

"proof"



so

$$\langle y - \pi_\Omega(y),\ \pi_\Omega(y) \rangle \geq \langle y - \pi_\Omega(y),\ x \rangle$$

$\forall x \in \Omega$

(IV) $\|\pi_\Omega(y) - x\|^2 + \|y - \pi_\Omega(y)\|^2 \leq \|y - x\|^2$

(V) $\|\pi_\Omega(y) - x\| \leq \|y - x\|$