* Note that projected GD can be modified in the same way for L-smooth functions to obtain ~~the~~ same convergence rate when solving

$$\min_{x \in \Omega} f(x) \quad \text{where } \Omega \subset \mathbb{R}^n \text{ is convex.}$$

Picking μ in Practice:

- The convergence theorems that we've seen so far require knowing L, the Lipschitz constant (or the smoothness constant) in order to set μ, but we may not know L in practice.

- Idea to get around this issue:

Use the best $\mu^{(t)}$ at every iteration.

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

$\Rightarrow$ Pick $\mu^{(t)}$ to minimize :

$$f(x^{(t+1)}) = f\Big(\underbrace{x^{(t)}}_{\text{known}} - \underbrace{\mu^{(t)}}_{\text{the variable we're optimizing}} \underbrace{\nabla f(x^{(t)})}_{\text{known}}\Big)$$

Issue : Solving for $\mu^{(t)}$ exactly is often hard, so we often settle for an approximation

Possible sol'n : Use, at every iteration t, a "backtracking Line Search"

How does this work?

Note that for any descent direction $\vec{p}$:

$$f(x) - \nabla f(x)^T p < f(x-p) \le \underbrace{f(x) - \gamma \nabla f(x)^T p}_{(*)}$$

for some small $\gamma > 0$.

Pick $p = \mu \nabla f(x)$, as we do in GD

and plug it into $(*)$:

$$\underbrace{f(x - \mu \nabla f(x)) \leqslant f(x) - \gamma \mu \|\nabla f(x)\|^2}_{(**)}$$

So we expect that for μ that is small enough, $(**)$ should hold

Idea: • Fix γ, e.g. $\gamma = 0.5$,

• Start with $\mu = 1$ (for example)

• decrease μ iteratively till

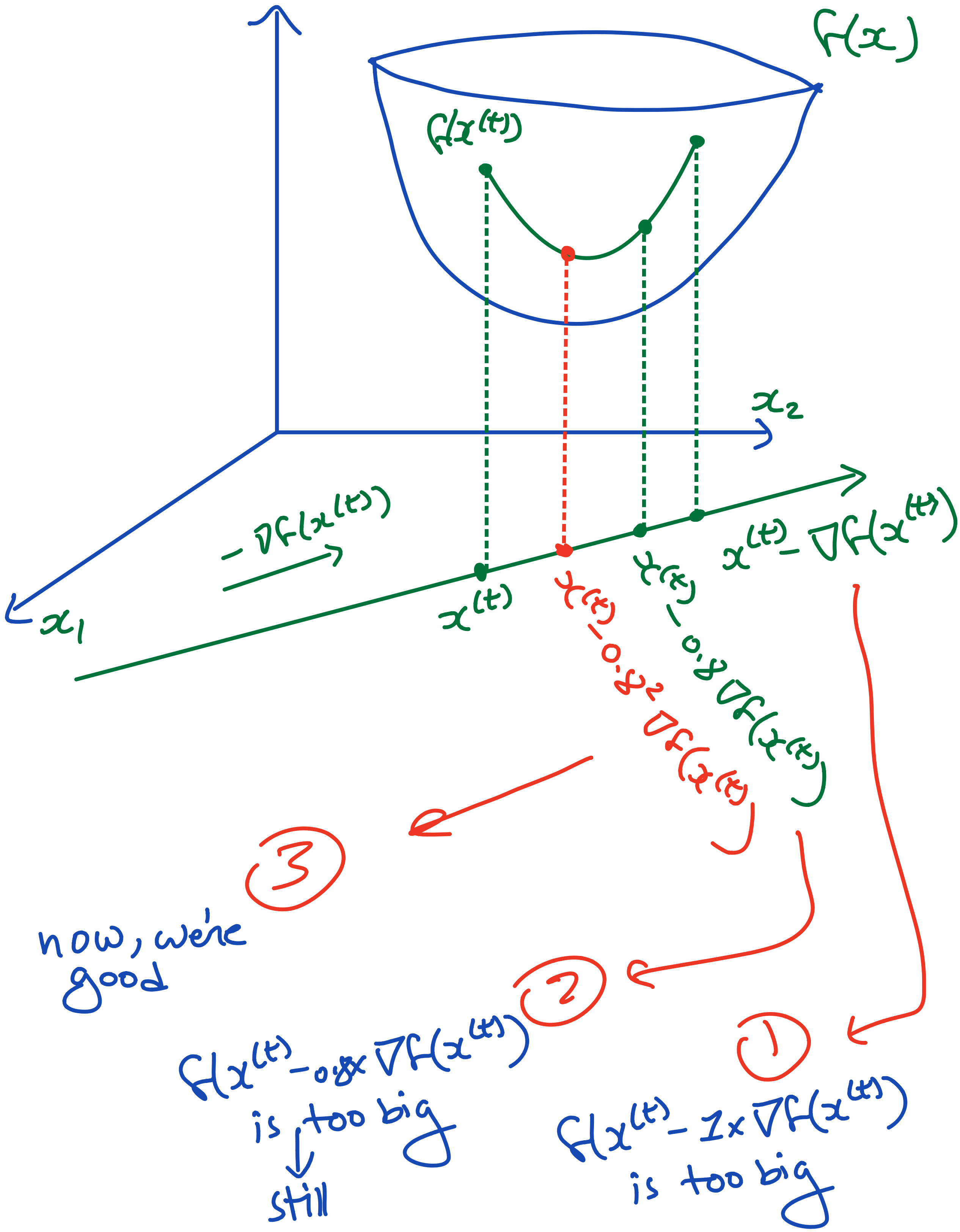
$$f(x^{(t+1)}) = f(x^{(t)} - \mu^{(t)} \nabla f(x^{(t)}))$$

$$\leqslant \underbrace{f(x^{(t)}) - \mu^{(t)} \gamma \|\nabla f(x^{(t)})\|^2}_{\text{Armijo Condition}}$$

For example

$$\mu^{(t)} = 1 \longrightarrow \mu^{(t)} = 0.8 \longrightarrow \mu^{(t)} = 0.8^2$$

$$\times \qquad\qquad \times \qquad\qquad \checkmark$$



$f(x)$
$f(x^{(t)})$
x_2
x_1
$-\nabla f(x^{(t)})$
$x^{(t)}$
$x^{(t)} - 0.8^2 \nabla f(x^{(t)})$
$x^{(t)} - 0.8 \nabla f(x^{(t)})$
$x^{(t)} - \nabla f(x^{(t)})$
3
now, we're good
2
$f(x^{(t)} - 0.8 \times \nabla f(x^{(t)}))$ is still too big
1
$f(x^{(t)} - 1 \times \nabla f(x^{(t)}))$ is too big

Then we discussed Backtracking Linesearch for selecting μ.

Pick $\beta < 1$, $\gamma < 1$

At each GD step t:

① Set $v = -\nabla f(x^{(t)})$

② Set $\mu^{(t)} = 1$

③ If $f(x^{(t)} - \mu^{(t)} \nabla f(x^{(t)})) \leq f(x^{(t)}) - \mu \gamma \|\nabla f(x^{(t)})\|^2$

then: keep $\mu^{(t)}$

Else: Set $\mu^{(t)} \leftarrow \beta \mu^{(t)}$
and repeat ③

Example: Consider

$$f(x_1, x_2) = (x_1 - 1)^4 + (x_1 + x_2 - 1)^2$$

(and note that $x^* = (1,0)$)

$$\nabla f = \left(4(x_1-1)^3 + 2(x_1+x_2-1),\ 2(x_1+x_2-1)\right)$$

Suppose that $x^{(0)} = (0,0)$

$$x^{(1)} = x^{(0)} - \mu \nabla f(\overset{(0,0)}{x^{(0)}})$$
$$= (0,0) - \mu(-6,-2)$$

Ideally, want to pick μ to minimize

$$f((0,0) - \mu(-6,-2)) = f(6\mu, 2\mu)$$
$$= (6\mu - 1)^4 + (8\mu - 1)^2$$

Picking μ this way entails solving a nonlinear opt. problem, which may be hard.

So let's use backtracking line search:

We start with $\mu = 1$:

$$\underbrace{f(x^{(1)})}_{x^{(0)}-\mu\nabla f(x^{(0)})} = 674 > \underbrace{\underbrace{f(x^{(0)})}_{2} - \underbrace{\mu}_{1}\,\underbrace{\gamma}_{0.5}\,\underbrace{\|\nabla f(x^{(0)})\|^2}_{40}}_{-18}$$

So we try $\mu = 1 \times 0.8 = 0.8$

$$f(x^{(1)}) = 237.6736 > \underbrace{f(x^{(0)}) - 0.8 \times 0.5\, \|\nabla f(x^{(0)})\|^2}_{-14}$$

So we try $\mu = 0.8^2$
(also fails)

So we keep going untill $\mu = 0.8^{11}$

gives

$$f(x^{(1)}) = 0.1530 \leq f(x^{(0)}) - 0.5 \times 0.8^{11}\, \|\nabla f(x^{(0)})\|^2$$

So we choose this μ $(\mu = 0.8^{11})$

and set $x^{(1)} = \underbrace{x^{(0)}}_{(0,0)} - \underbrace{\mu}_{0.8^{11}}\, \underbrace{\nabla f(x^{(0)})}_{\nabla f(0,0)}$

and we repeat this process for
$t = 2, 3, \dots$
(for example $x^{(1000)} = (0.979, 0.021)$)

$\boxed{\begin{array}{l}\text{Theorem: For an } L\text{-smooth convex function with } \mu^{(t)} \text{ set by backtracking line search, GD gives} \\ f(x^{(t)}) - f(x^*) \leq \dfrac{1}{2t \min\limits_{s=1,\dots,t}(\mu^{(s)})} \\ \text{and } \min\limits_{s=1,\dots,t} \mu^{(s)} \geq \min\left(1, \dfrac{\beta}{L}\right)\end{array}}$

$\hookrightarrow$ this guarantees

$$f(x^{(t)}) - f(x^*) \leq \frac{L}{2t\beta}$$