LECTURE 24: GRADIENT DESCENT

Today: Another cool way of finding the local max/min of a function

What is wrong with our previous $D^2 f$ method? Nothing, it's just that it requires you to

(1) Calculate second derivatives

(2) Find the critical points = lots of algebra

(3) Calculate determinants = computationally expensive

The methods below are fairly quick and only require first derivatives, although they just give you an approximate answer.

1. Introduction

Let's motivate this with an application

Date: Tuesday, December 6, 2022.

Example 1: To the Moon!

Suppose you're building a portfolio using two kinds of stocks: GME (Gamestop) and AMC

$$\begin{aligned} x_1 &= \text{number of GME shares} \\ x_2 &= \text{number of AMC shares} \end{aligned}$$

Each stock has a return and a risk, given by expected values and variances respectively

Assume the expected return of GME is 20% and for AMC is 16%

Then a (deterministic) model for your wealth is

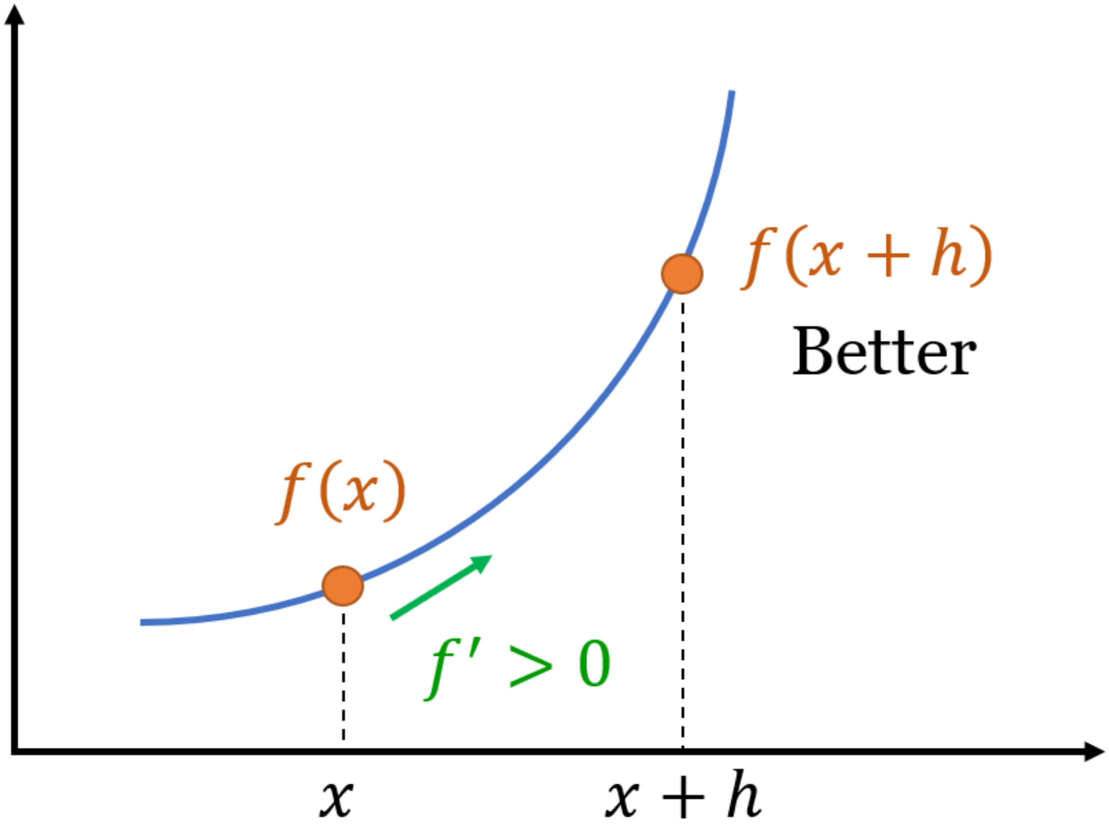
$$\begin{aligned} &\max f(x_1, x_2) \\ &\quad f(x_1, x_2) = \ 20x_1 + 16x_2 - \left(2\left(x_1\right)^2 + \left(x_2\right)^2 + \left(x_1 + x_2\right)^2\right) \end{aligned}$$

Note: It's possible to turn this into a probabilistic model, but this is outside the scope of the course.

2. Motivation

1D Motivation: Given a 1D (differentiable) function f we have

$$f'(x) = \lim_{h \to 0^+} \frac{f(x+h) - f(x)}{h}$$



If $f'(x) > 0$ then this means that for small $h > 0$ we have

$$\frac{f(x+h) - f(x)}{h} \approx f'(x) > 0$$

So for small $h > 0$ we get $f(x+h) - f(x) > 0 \Rightarrow f(x+h) > f(x)$.

This means that $f(x+h)$ is a *bigger* value than $f(x)$, and so $f(x+h)$ is a better candidate for a max.

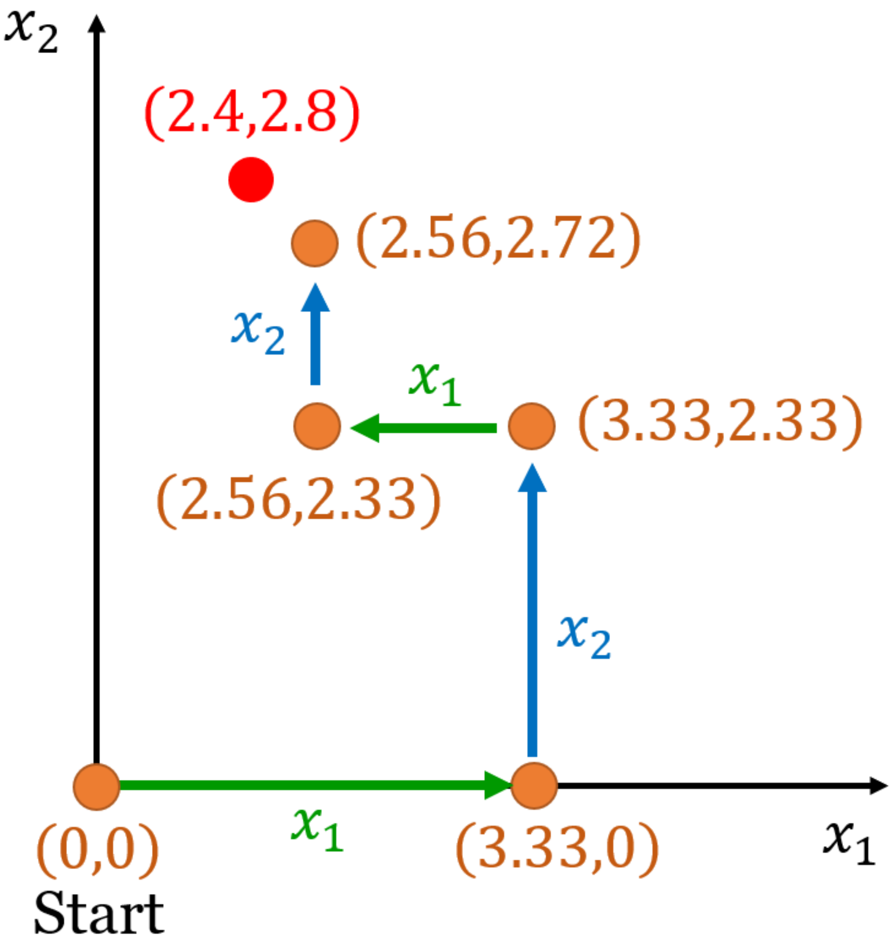
But then we can repeat the same thing with $x+h$ instead of x and get an even bigger value. Until when can we do that? Until $f' = 0$

3. Method 1: Cyclic Ascent

Example 2:

Apply cyclic ascent to find an approximate local max of

$$f(x_1,x_2) = 20x_1 + 16x_2 - \left(2\,(x_1)^2 + (x_2)^2 + (x_1+x_2)^2\right)$$



Main Idea: Same idea, but we start with f_{x_1} then do f_{x_2} then do f_{x_3} until we run out of variables. Then start again with f_{x_1}, f_{x_2}, and so on and so forth. It's called a cyclic method because we're cycling through the variables.

STEP 0: Start with a point, say $(0, 0)$ (any point works)

Candidate for max: $(0, 0)$

(It makes sense in terms of our portfolio context, because you start out with no shares before buying them)

Idea: Like the simplex method, start at $(0, 0)$ and move in a direction

STEP 1: x_1-direction

$$f_{x_1} = 20 - 4x_1 - 2(x_1 + x_2) = 20 - 6x_1 - 2x_2$$

Along our path, we have $x_2 = 0$ and so

$$f_{x_1}(x_1, 0) = 20 - 6x_1 - 2(0) = 20 - 6x_1$$

And once again it's the same idea: Increase x_1 until $f_{x_1} = 0$

$$f_{x_1} = 0 \Rightarrow 20 - 6x_1 = 0 \Rightarrow x_1 = \frac{20}{6} = \frac{10}{3}$$

Candidate for max: $\left(\frac{10}{3}, 0\right) \approx (3.33, 0)$

We have increased x_1 as much as we could, and so we continue with x_2

STEP 2: x_2-direction

$$f_{x_2} = 16 - 2x_2 - 2\left(x_1 + x_2\right) = 16 - 4x_2 - 2x_1$$

In this case we have $x_1 = \frac{10}{3}$ and so we get

$$f_{x_2}\left(\frac{10}{3}, x_2\right) = 16 - 4x_2 - 2\left(\frac{10}{3}\right) = \frac{48}{3} - \frac{20}{3} - 4x_2 = \frac{28}{3} - 4x_2 = 0$$

$$x_2 = \frac{28}{(3)(4)} = \frac{7}{3}$$

Candidate for max: $\left(\frac{10}{3}, \frac{7}{3}\right) \approx (3.33, 2.33)$

Note: If we had x_3, we would continue with x_3, but since we've exhausted all our variables, we start again at x_1

STEP 3: x_1−direction

Start at $x_2 = \frac{7}{3}$ and change x_1

$$f_{x_1} = 20 - 6x_1 - 2x_2$$

$$f_{x_1}\left(x_1, \frac{7}{3}\right) = 20 - 6x_1 - 2\left(\frac{7}{3}\right) = \frac{46}{3} - 6x_1 = 0$$

$$x_1 = \frac{46}{3(6)} = \frac{23}{9} \approx 2.56$$

Candidate for max: $\left(\frac{23}{9}, \frac{7}{3}\right) \approx (2.56, 2.33)$

Notice this time x_1 decreased instead of increased

STEP 4+ And then you continue:

Fix $x_1 = \frac{23}{9}$ and change x_2, to get $(2.56, 2.72)$

Then fix $x_2 = 2.72$ and change x_1, and so on

Question: When do you stop?

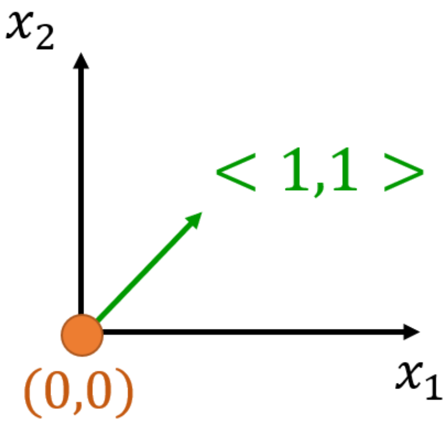
Answer: Whenever you want!

Notice in fact that the points are getting closer and closer to each other, so you stop whenever you find an answer that is within your desired threshold. For instance, if you want an answer correct to 1 decimal place, then you would continue until you get $(2.4, 2.8)$

This makes sense in context of the problem: What difference does it really make if you buy 23.12 shares or 23.13 shares?

Note: While this method is faster than the second derivative one, it's not always super efficient.

Example: What if, instead of increasing x_1 or x_2 separately, we increase (x_1, x_2) in the $(1, 1)$ direction?



This leads us to…

4. Review of Directional Derivatives

Example 3:

$$f(x_1, x_2) = 3x_1x_2 + 4(x_2)^2$$

Calculate the following:

(a) ∇f

(b) The directional derivative of f in the direction $\langle 3, 4\rangle$

(c) The direction of the greatest rate of increase/decrease of f at $(1, 2)$

$$\text{(a)}\ \ \nabla f = \langle f_{x_1}, f_{x_2}\rangle = \langle 3x_2, 3x_1 + 8x_2\rangle$$

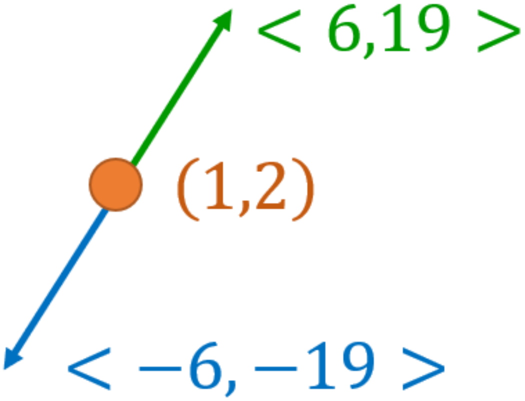
$$\begin{aligned}\text{(b) } D_{\langle 3,4\rangle} f &= (\nabla f) \cdot \langle 3, 4\rangle = \langle 3x_2, 3x_1 + 8x_2\rangle \cdot \langle 3, 4\rangle \\ &= 3(3x_2) + 4(3x_1 + 8x_2) = 12x_1 + 41x_2\end{aligned}$$

Note: Some books require $\langle 3, 4\rangle$ to be a unit vector, but we won't take that approach here.

(c) Largest increase is in the direction of the gradient, so

$$\textbf{Answer: } \nabla f(1, 2) = \langle 3(2), 3(1) + 8(2)\rangle = \langle 6, 19\rangle$$

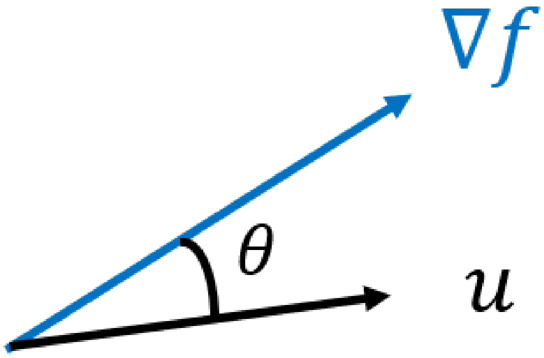
And the largest decrease is in $-\langle 6, 19\rangle = \langle -6, -19\rangle$



Note: (c) follows because

$$D_u f = \nabla f \cdot u = |\nabla f|\,|u|\cos(\theta)$$

Which is largest if $\theta = 0$, that is if u points the same direction as ∇f, and smallest if $\theta = \pi$, u points the opposite direction of ∇f



This tells us that, in order to maximize f, it's best to move in the direction of ∇f

5. Method 2: Steepest Ascent

Quite fun, because you turn a max problem into another max problem!

Example 4:

Apply steepest ascent to

$$f(x_1, x_2) = 20x_1 + 16x_2 - \left(2\left(x_1\right)^2 + \left(x_2\right)^2 + \left(x_1 + x_2\right)^2\right)$$

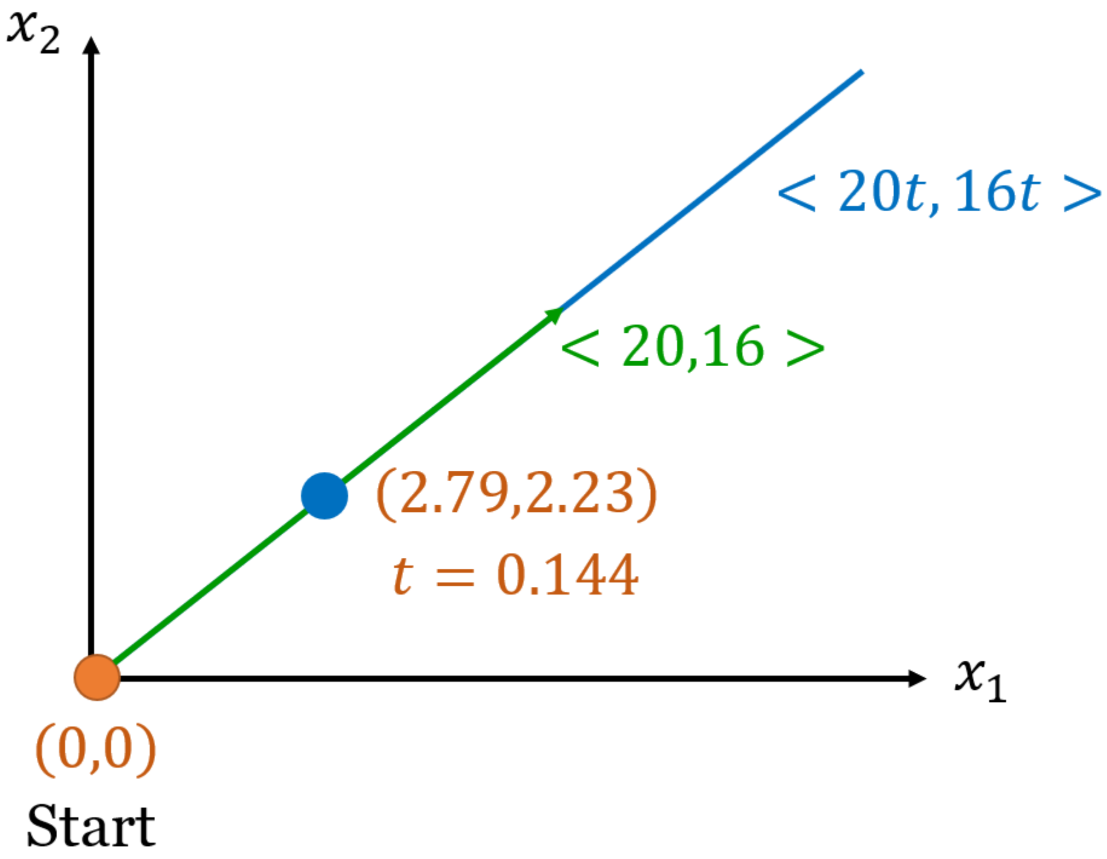
STEP 0: Start at $(0,0)$

STEP 1: Calculate $\nabla f(0,0)$

$$\begin{aligned} f_{x_1} &= 20 - 4x_1 - 2\left(x_1 + x_2\right) \\ f_{x_2} &= 16 - 2x_2 - 2\left(x_1 + x_2\right) \end{aligned}$$

$$\nabla f(0,0) = \langle 20, 16 \rangle$$

So from the above, to maximize f, it makes sense to start at $(0,0)$ and move in the direction of $\langle 20, 16 \rangle$



By how much should we move? Cannot apply the previous method and increase until ∇f is zero because that's unlikely to happen.

STEP 2: Solve a 1D max problem:

The line starting at $(0,0)$ and direction $\langle 20,16\rangle$ is parametrized as

$$\langle 0,0\rangle + t\,\langle 20,16\rangle = \langle 20t, 16t\rangle$$

$$\begin{aligned} f(20t,16t) =& 20\,(20t) + 16\,(16t) - \left(2\,(20t)^2 + (16t)^2 + (20t+16t)^2\right) \\ =& 656t - 2352t^2 = g(t) \end{aligned}$$

Maximize this with respect to t

$$g'(t) = 656 - 2352(2t) = 0 \Rightarrow t = \frac{656}{2(2352)} = \frac{41}{294} \approx 0.1395$$

And $g''(t) = -2(2352) < 0$ so indeed a local max.

Candidate for max

This means you start at $(0,0)$ and move by $(0.1395)\left\langle 20, 16\right\rangle$ and so

$$(0,0) + (0.1395)(20,16) = (2.79, 2.23)$$

Which is much closer to $(2.4, 2.8)$ than our previous method

STEP 3: Start at $(2.79, 2.23)$ and calculate $\nabla f(2.79, 2.23)$

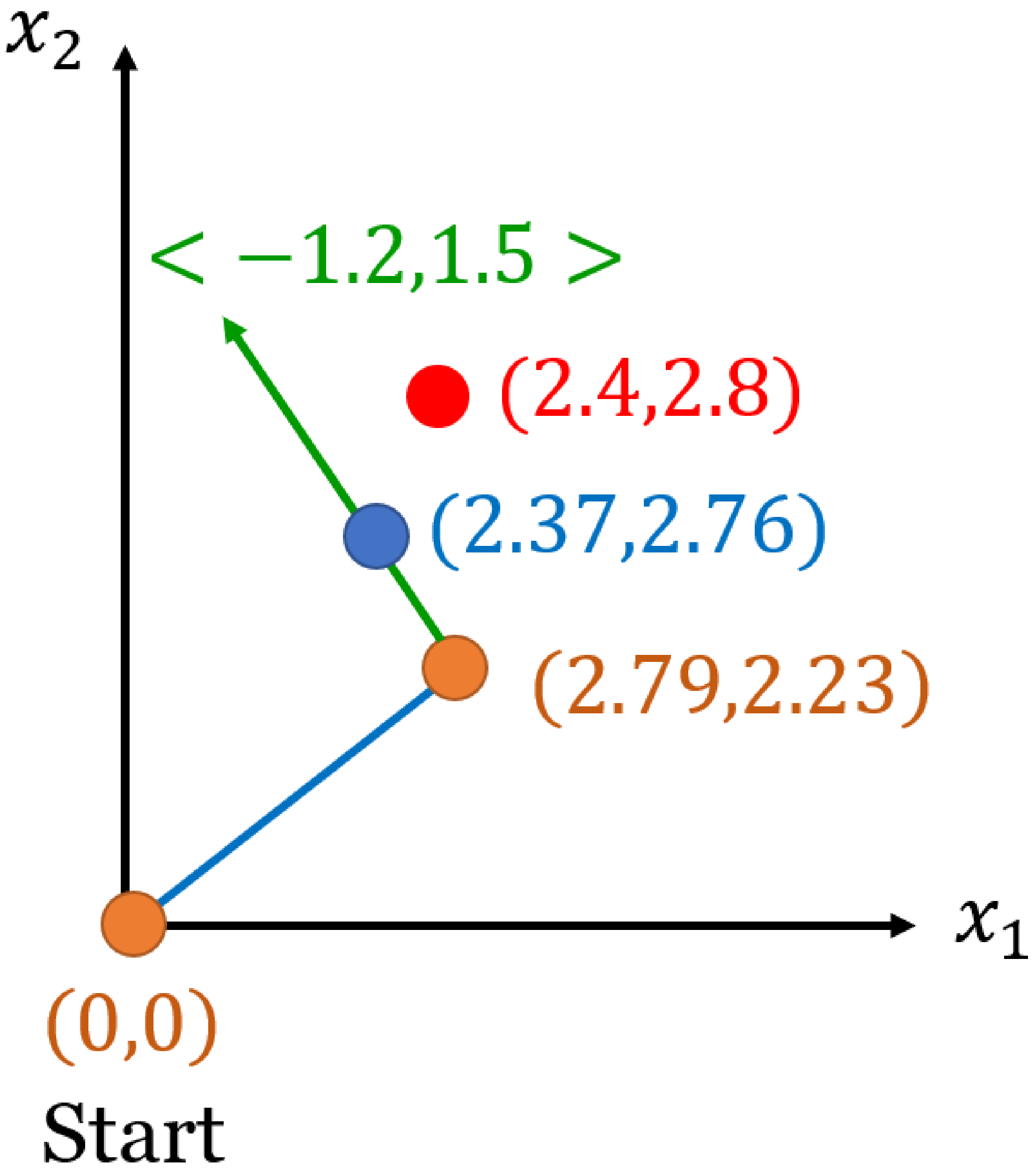
$$\begin{aligned} f_{x_1} =& 20 - 4x_1 - 2\left(x_1 + x_2\right) \\ f_{x_2} =& 16 - 2x_2 - 2\left(x_1 + x_2\right) \end{aligned}$$

$$\nabla f(2.79, 2.23) = \left\langle -1.2, 1.5\right\rangle$$

Here the line is parametrized by

$$\left\langle 2.79, 2.23\right\rangle + t\left\langle -1.2, 1.5\right\rangle = \left\langle 2.79 - 1.2t, 2.23 + 1.5t\right\rangle$$

$$f(2.79 - 1.2t, 2.23 + 1.5t) = -5.22t^2 + -3.69t + 45.7385 = g(t)$$



$$g'(t) = -10.44t - 3.69 = 0 \Rightarrow t = \frac{3.69}{10.44} \approx 0.353$$

$g''(t) = -10.44 < 0$ so we indeed have a max.

Candidate for max

$$(2.79, 2.23) + (0.353)(-1.2, 1.5) = (2.37, 2.76)$$

Which is even closer to $(2.4, 2.8)$

Note: Technically, the two methods give you a local max, and it's possible in practice that your iterations get closer and closer to a min that you don't care about, like a marble being stuck in a pinball machine. In that case you choose a very large value of t, just so you can leave the region that you're stuck in, like shaking a pinball machine.

This officially ends our exploration of nonlinear programs! This is really just the tip of the iceberg, there are several other methods that prove to be useful as well, such as Lagrange Multipliers (maximizing a function with a constraint) and Newton's method (useful to find zeros of functions, which is useful because critical points are literally zeros of derivatives)

The End