

Otimização Não Linear

CM106/CMM204/CMI043

Tópico 03 - Minimização Irrestrita

Abel Soares Siqueira - UFPR

2020/s1

Minimização Irrestrita

- Buscamos o mínimo de $f : \mathbb{R}^n \rightarrow \mathbb{R}$ para todo $x \in \mathbb{R}^n$.
- Idealmente buscamos um minimizador global, i.e., $x^* \in \mathbb{R}^n$ tal que

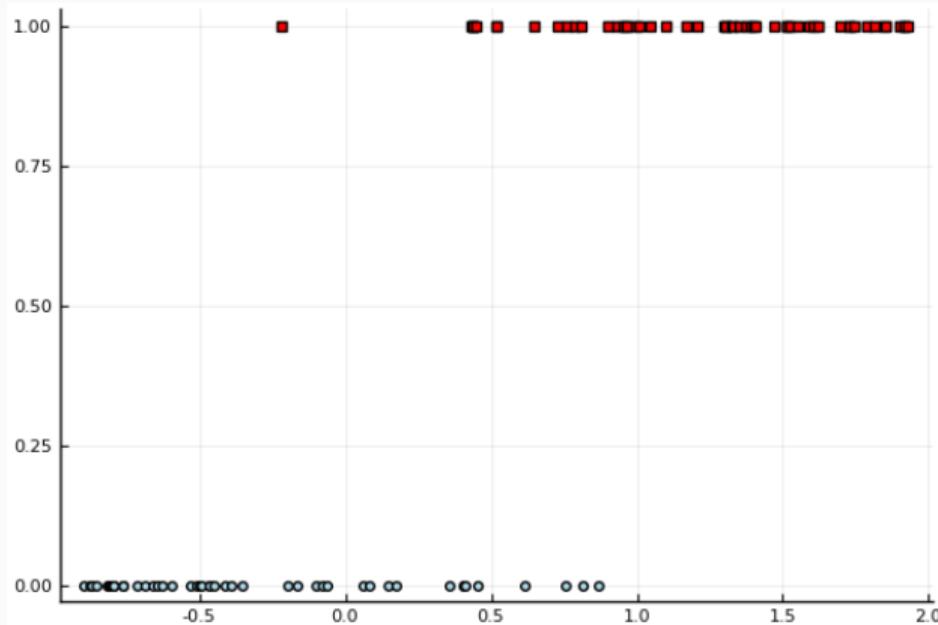
$$f(x^*) \leq f(x), \quad \forall x \in \mathbb{R}^n.$$

- No entanto em geral só conseguimos verificar se estamos com um minimizador local, i.e., $x^* \in \mathbb{R}^n$ tal que

$$\exists \epsilon > 0 : \quad f(x^*) \leq f(x), \quad \forall x \in \mathbb{R}^n \cap B(x^*, \epsilon).$$

Regressão Logística

- $\{(x_i, y_i), i = 1, \dots, m\} \subset \mathbb{R} \times \{0, 1\}$.



Régressão Logística

- $Y \sim \text{Bernoulli}(p)$: $P(Y = 1) = p$, $P(Y = 0) = 1 - p$, ou seja

$$P(Y = y) = p^y(1 - p)^{1-y}.$$

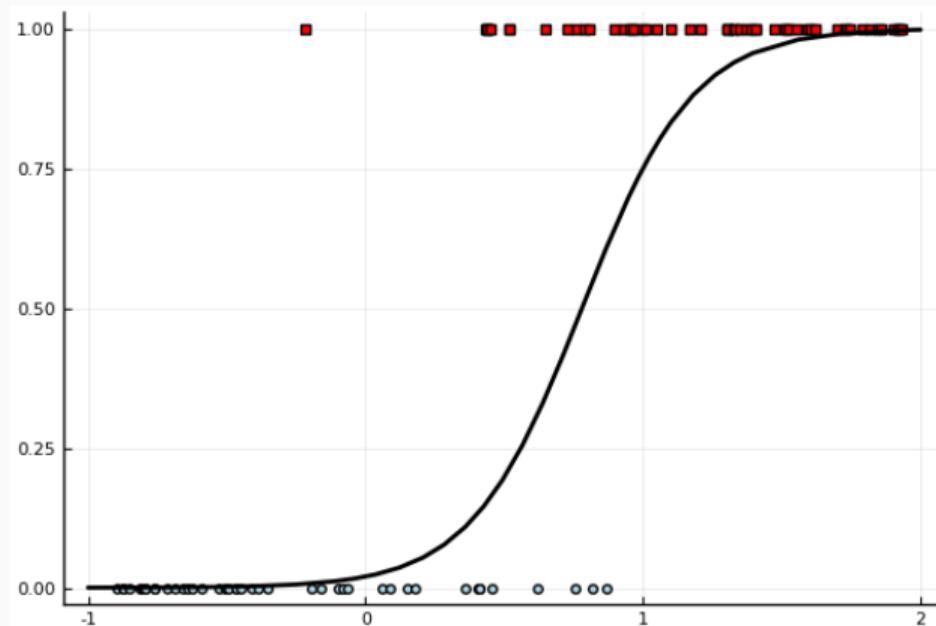
- Associamos p à x_i . Como p deve estar entre 0 e 1, usamos

$$p(\beta; x) = \frac{1}{1 + e^{-\beta_0 - \beta_1 x}}.$$

- A probabilidade de $Y = y_i$ dado que $X = x_i$ é dado por

$$P(Y = y_i) = p(\beta; x_i)^{y_i} (1 - p(\beta; x_i))^{1-y_i}.$$

Regressão Logística



Regressão Logística

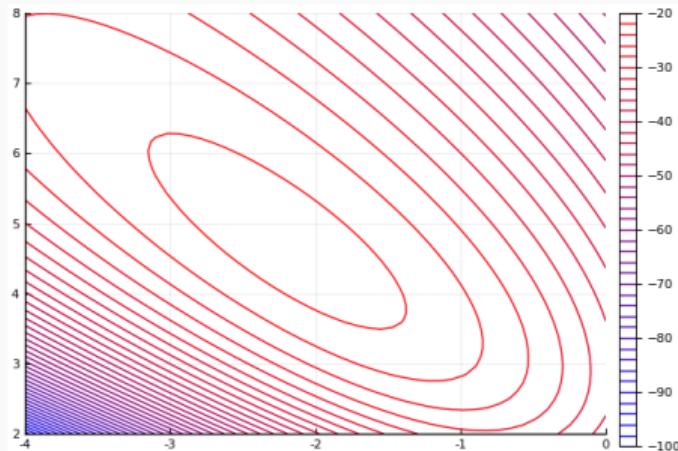
- A probabilidade de todos esses eventos acontecerem - supondo independência - é dada pela função de **Verossimilhança**

$$L(\beta) = \prod_{i=1}^m p(\beta; x_i)^{y_i} (1 - p(\beta; x_i))^{1-y_i}.$$

- Como essa função é difícil de tratar, consideramos $\ell(\beta) = \ln L(\beta)$, a função de log-verossimilhança.

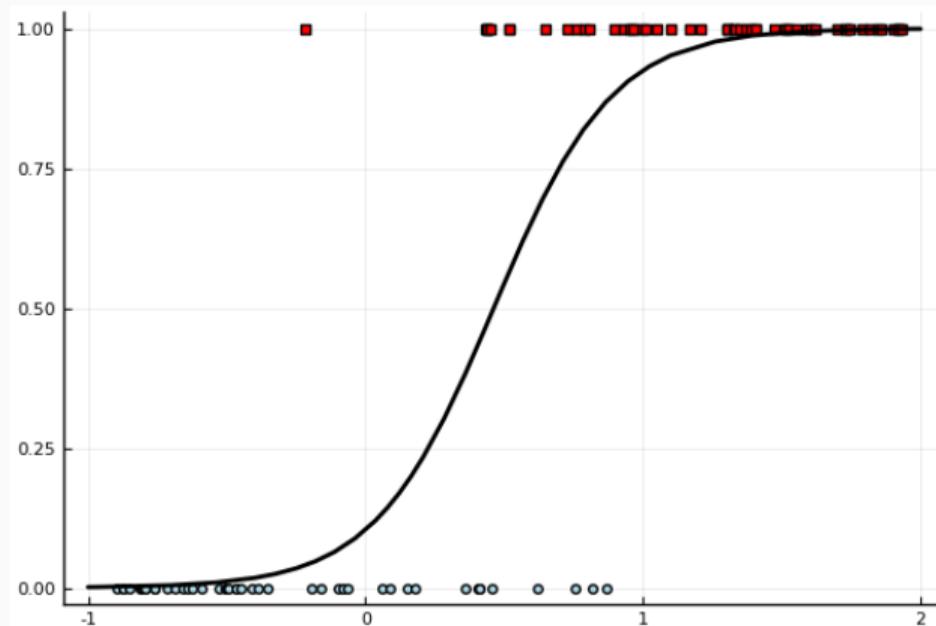
$$\ell(\beta) = \sum_{i=1}^m y_i \ln(p(\beta; x_i)) + (1 - y_i) \ln(1 - p(\beta; x_i)).$$

Regressão Logística



- ℓ parece um pouco com uma função quadrática, mas não é, no entanto podemos maximizá-la. Veremos como.

Regressão Logística



Reduzindo a quadráticas

Teorema de Taylor

Teo.: Seja $f : \mathbb{R}^n \rightarrow \mathbb{R}$ continuamente diferenciável e $d \in \mathbb{R}^n$. Então,

$$f(x + d) - f(x) = \nabla f(x + td)^T d = \nabla f(x)^T d + o(\|d\|),$$

para algum $t \in (0, 1)$.

Teo.: Seja $f : \mathbb{R}^n \rightarrow \mathbb{R}$ continuamente diferenciável até segunda ordem e $d \in \mathbb{R}^n$. Então,

$$f(x + d) - f(x) - \nabla f(x)^T d = \frac{1}{2} d^T \nabla^2 f(x + td) d = \frac{1}{2} d^T \nabla^2 f(x) d + o(\|d\|^2),$$

para algum $t \in (0, 1)$.

Modelo Quadrático

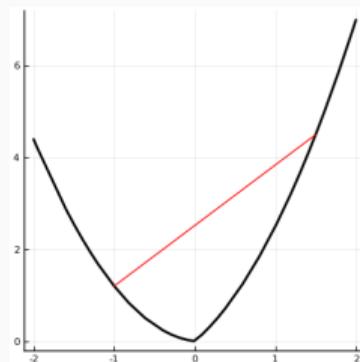
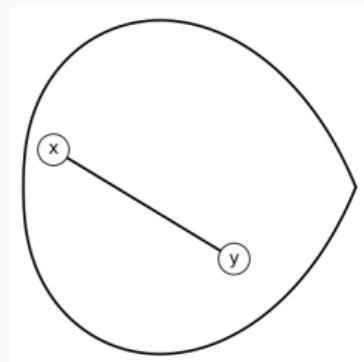
- Sabemos encontrar o mínimo de uma quadrática.
- Usando o Teorema de Taylor, conseguimos aproximar f por uma quadrática.
- Minimizando essa quadrática, se ela tiver matriz definida positiva, encontramos uma aproximação para o mínimo de f .

Convexidade

Def.: Um conjunto $S \subset \mathbb{R}^n$ é dito convexo se para todo $x, y \in S$, o segmento $[x, y] := \{\lambda x + (1 - \lambda)y : \lambda \in [0, 1]\} \subset S$.

Def.: Uma função $f : S \rightarrow \mathbb{R}$ em S convexo é dita convexa se para todo $x, y \in S$, vale

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$



Convexidade

Teo.: Seja $f : S \rightarrow \mathbb{R}$ diferenciável em S convexo. Então f é convexa se, e somente se,

$$f(y) \geq f(x) + \nabla f(x)^T(y - x), \quad \forall x, y \in S.$$

Também vale que f é convexa se, e somente se,

$$(\nabla f(y) - \nabla f(x))^T(y - x) \geq 0, \quad \forall x, y \in S.$$

Se f é duas vezes diferenciável, então f é convexa se, e somente se,

$$\nabla^2 f(x) \quad \text{é semi-definida positiva.}$$

Análogo: estritamente convexo mudando para desigualdades estritas e definida positiva.

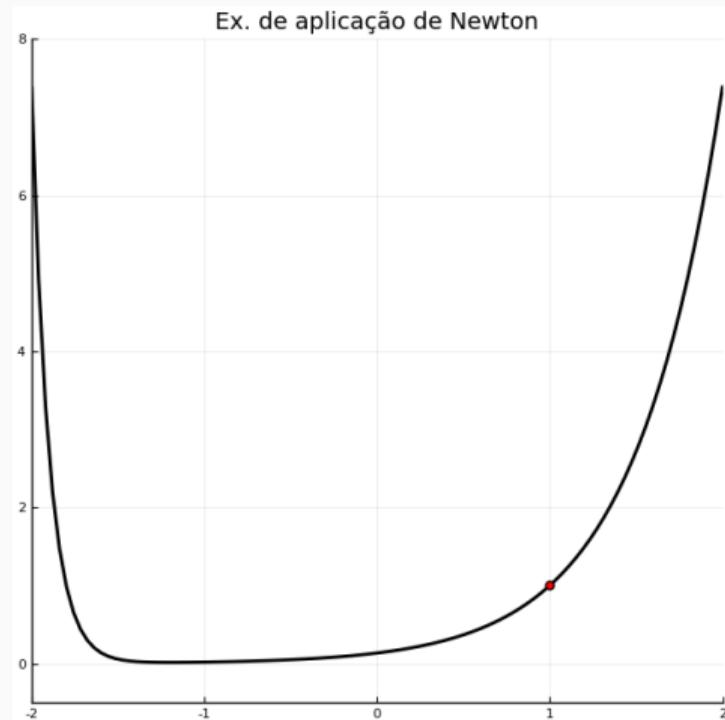
Método de Newton

- Se f é duas vezes diferenciável e estritamente convexa em \mathbb{R}^n , então para qualquer $x_k \in \mathbb{R}^n$, temos a seguinte aproximação quadrática:

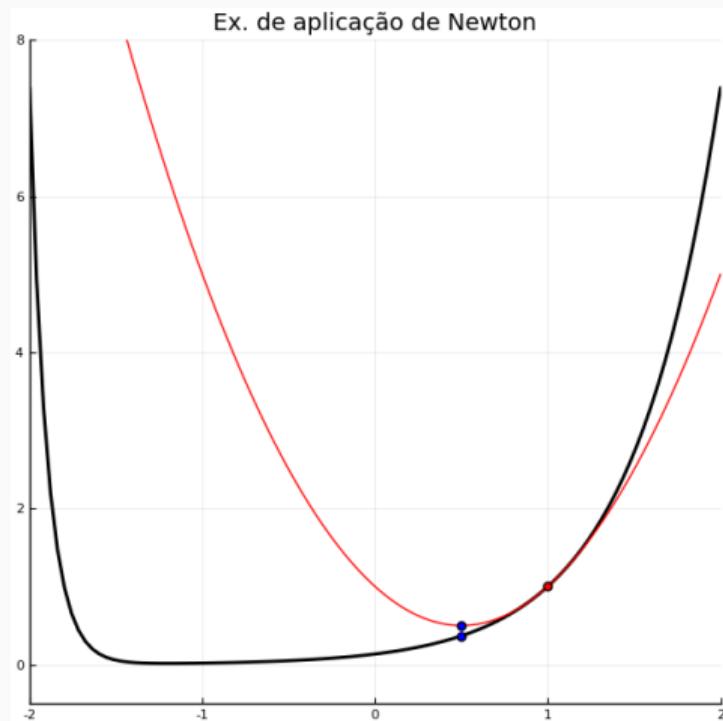
$$m_k(d) = f(x_k) + \underbrace{d^T \nabla f(x_k)}_{g_k} + \frac{1}{2} d^T \underbrace{\nabla^2 f(x_k)}_{B_k} d.$$

- Como f é estritamente convexa, temos B_k definida positiva. Logo m_k tem um minimizador global $d_k = -B_k^{-1} g_k$.
- Definimos $x_{k+1} = x_k + d_k$.

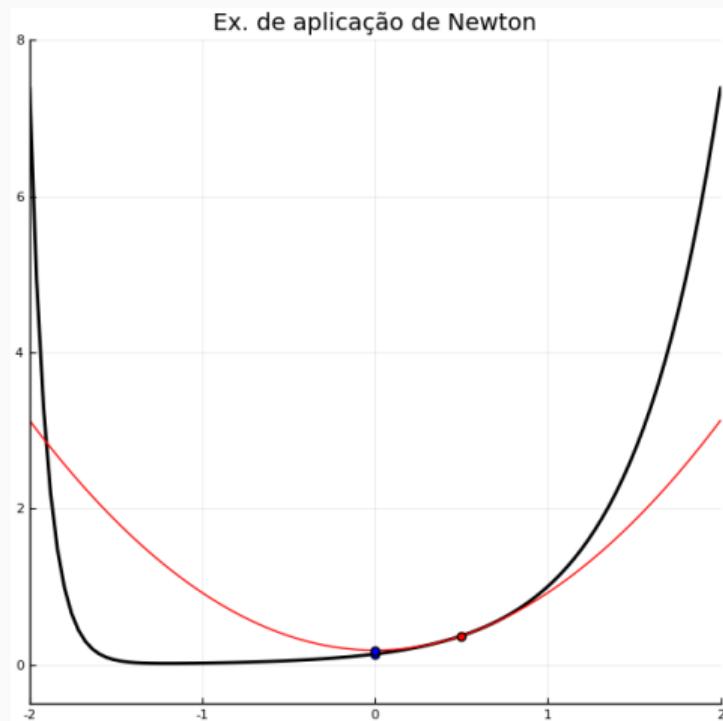
Método de Newton



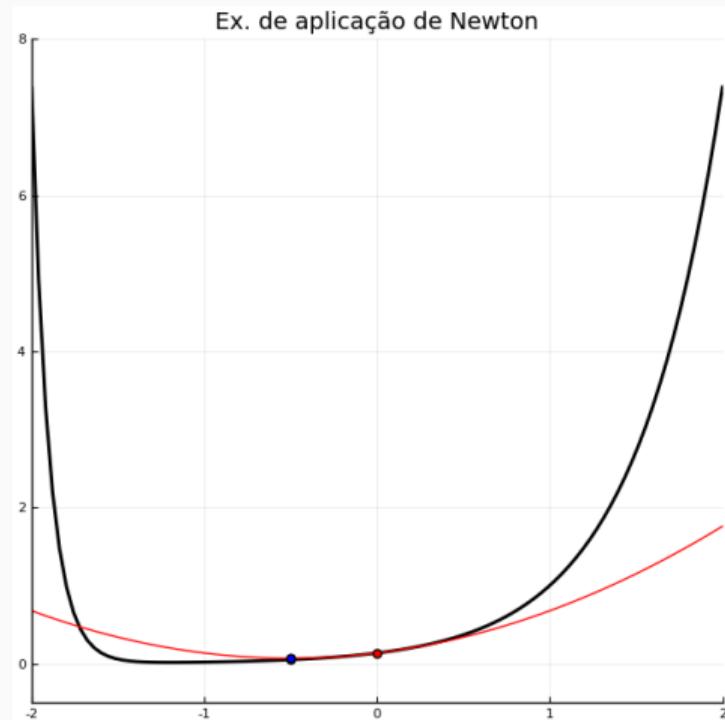
Método de Newton



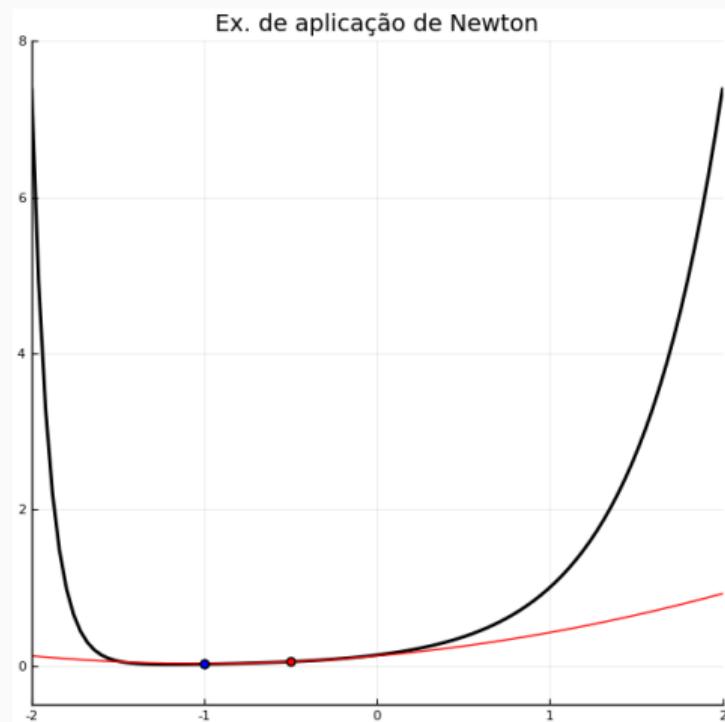
Método de Newton



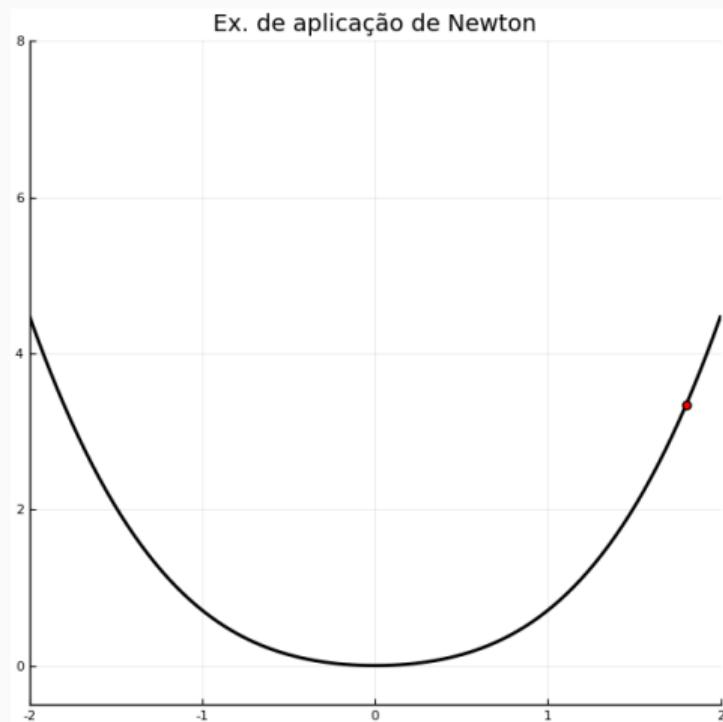
Método de Newton



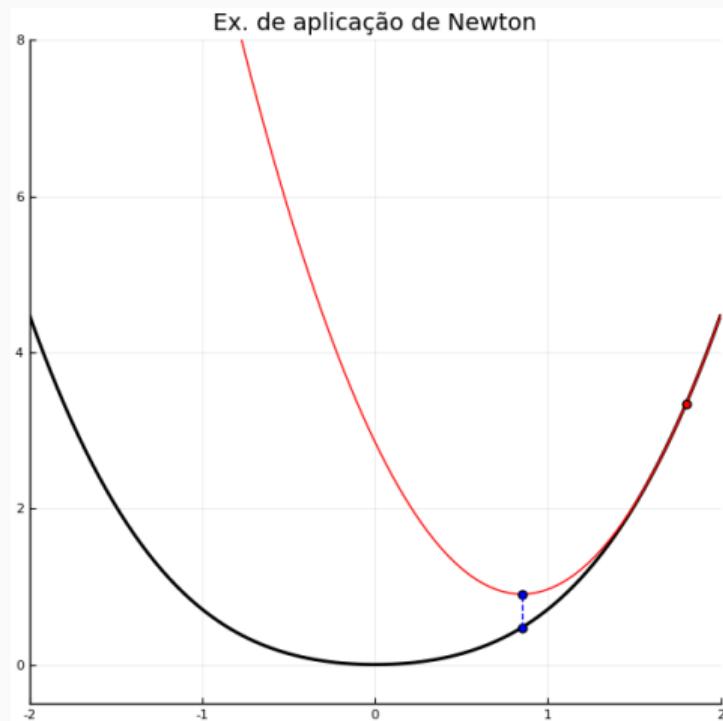
Método de Newton



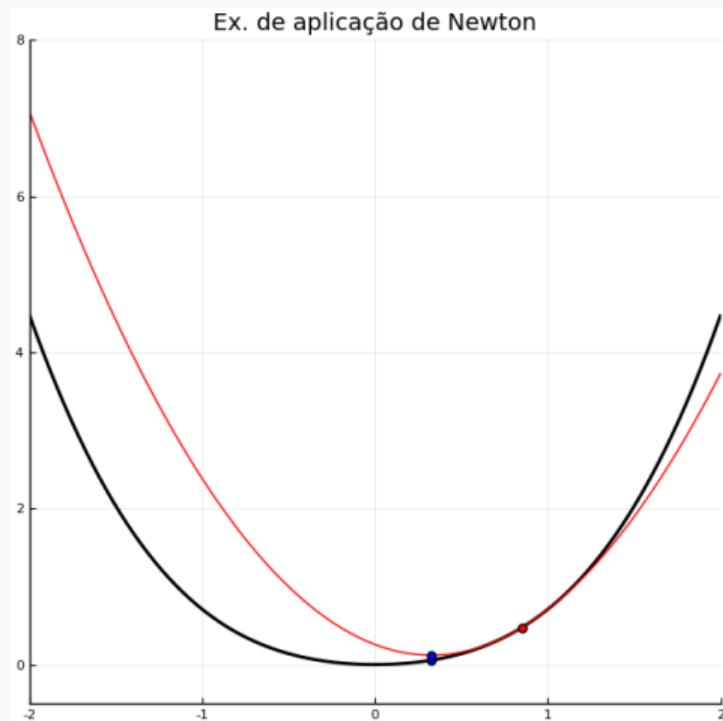
Método de Newton



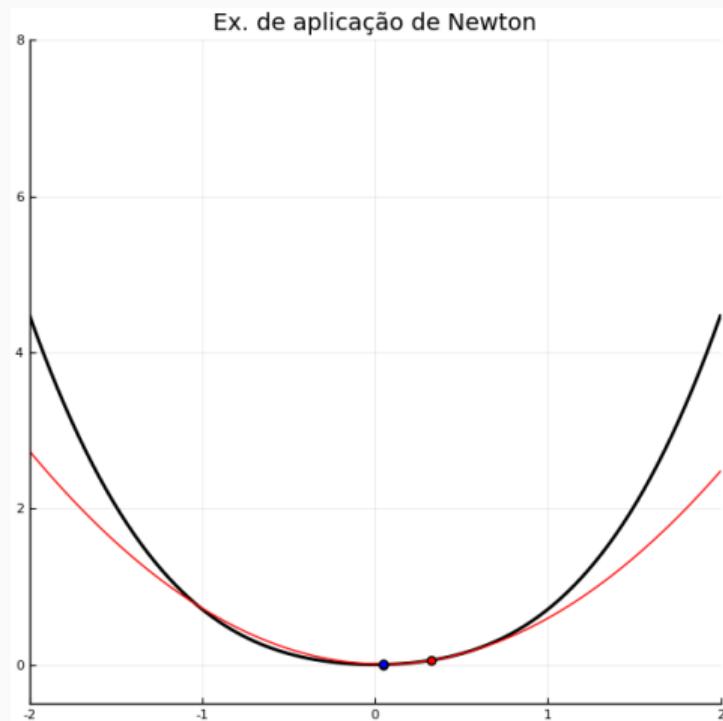
Método de Newton



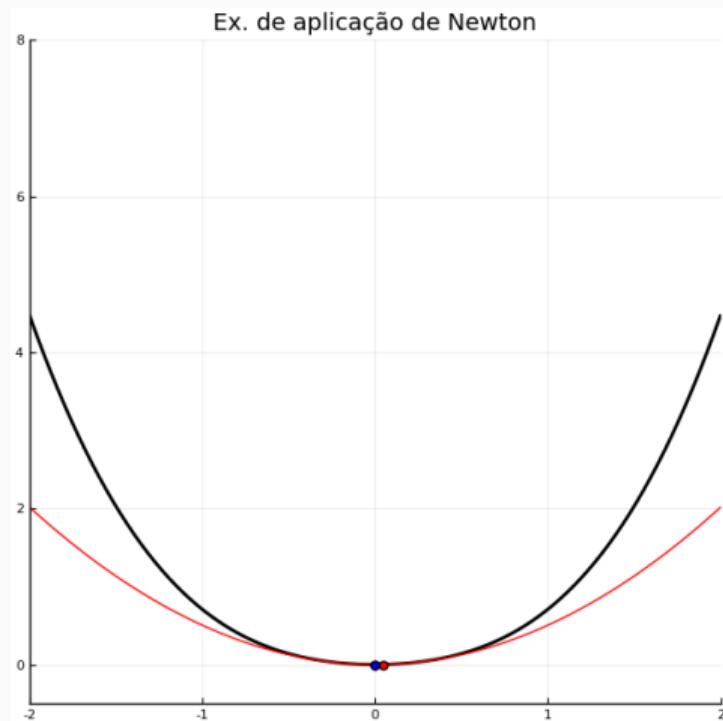
Método de Newton



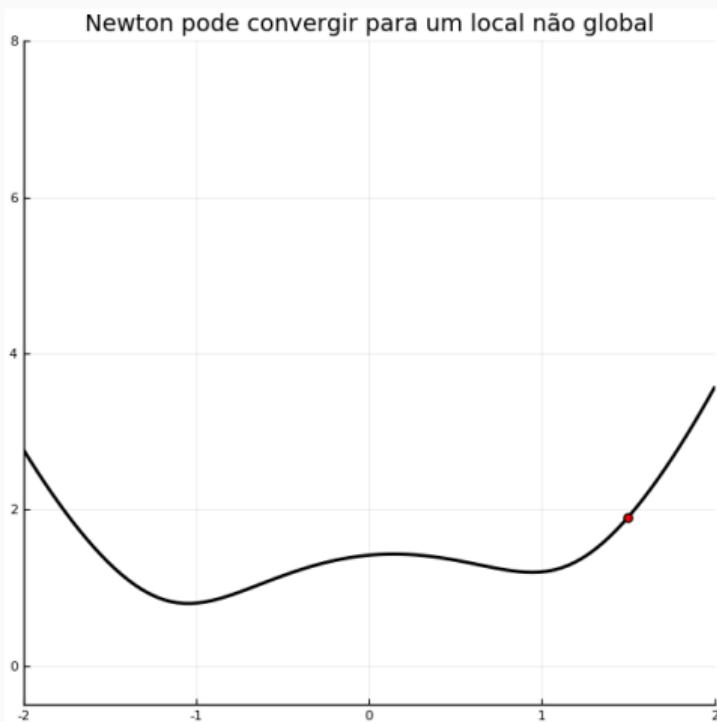
Método de Newton



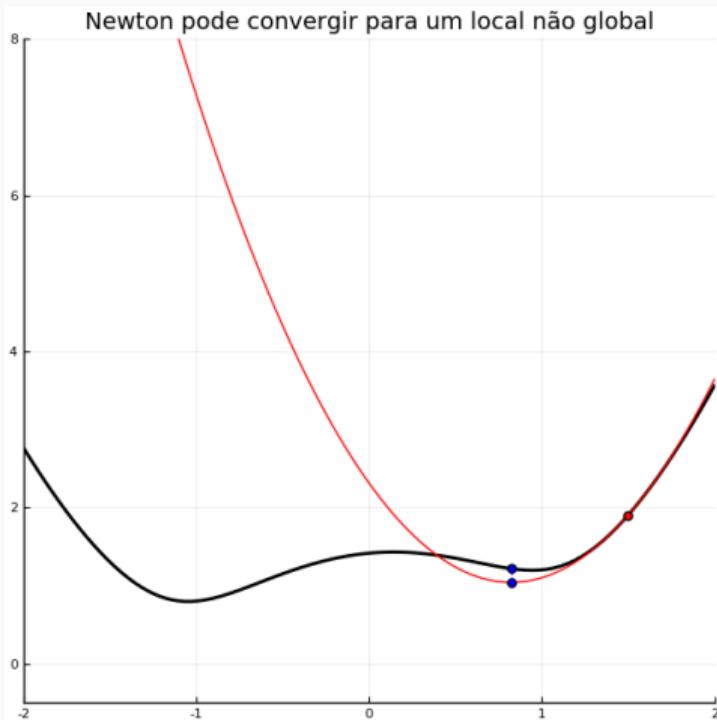
Método de Newton



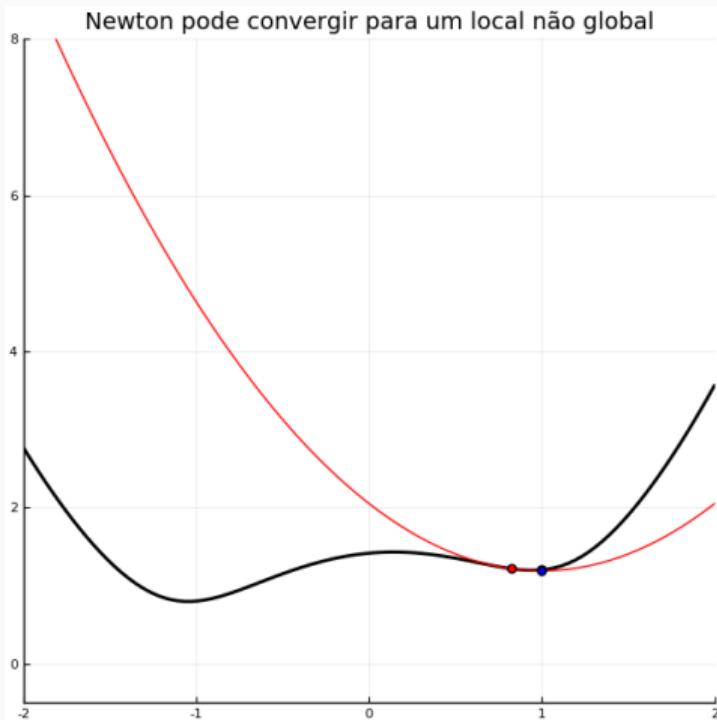
Método de Newton



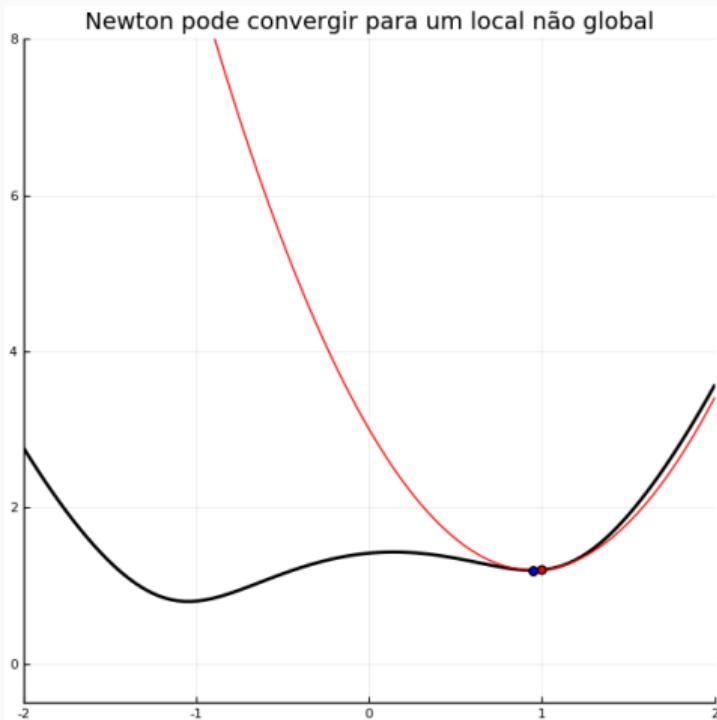
Método de Newton



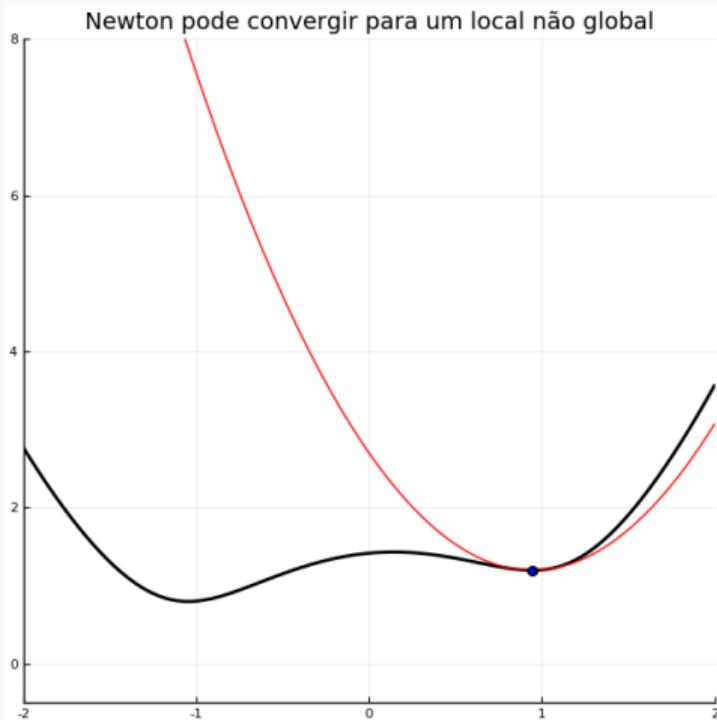
Método de Newton



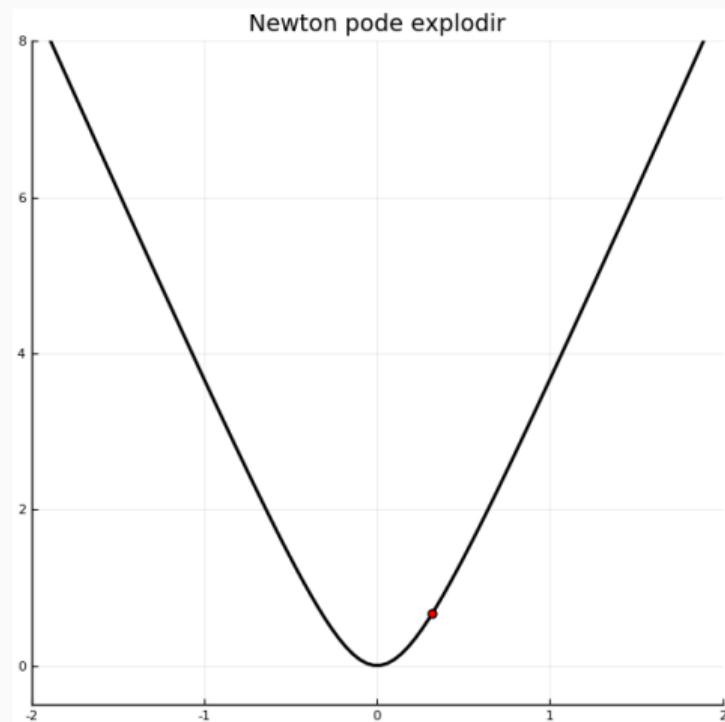
Método de Newton



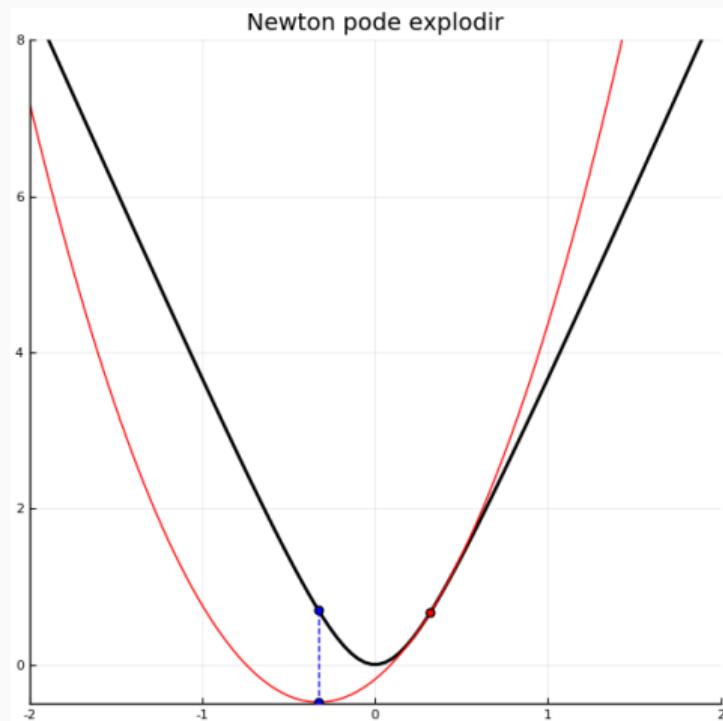
Método de Newton



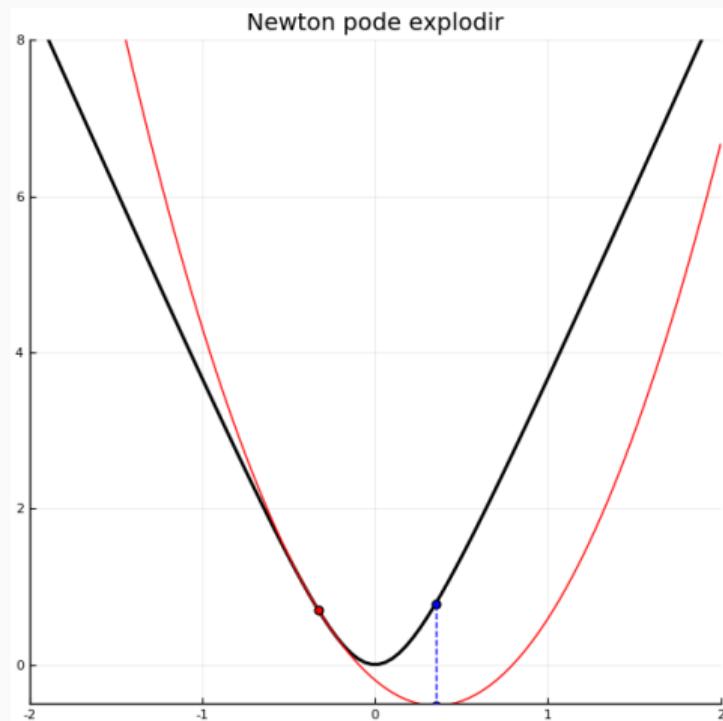
Método de Newton



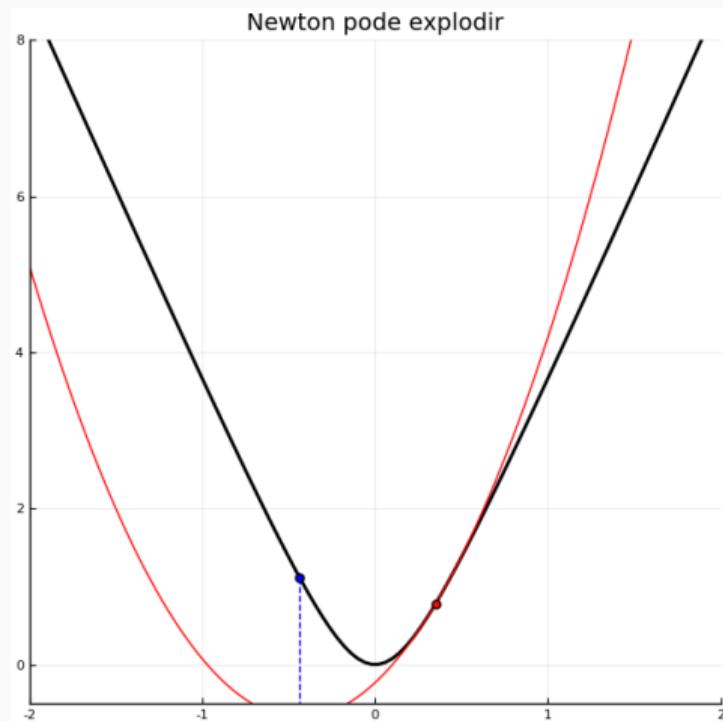
Método de Newton



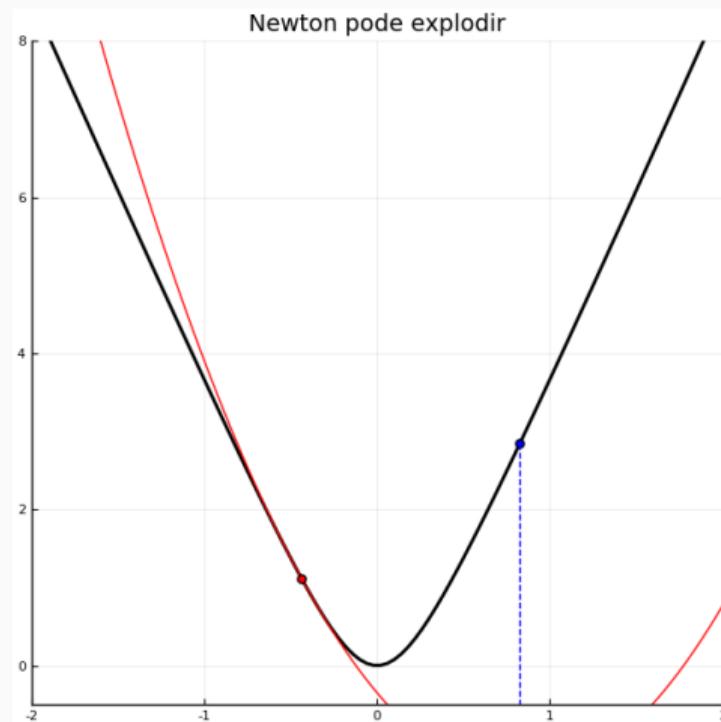
Método de Newton



Método de Newton



Método de Newton



Método de Newton



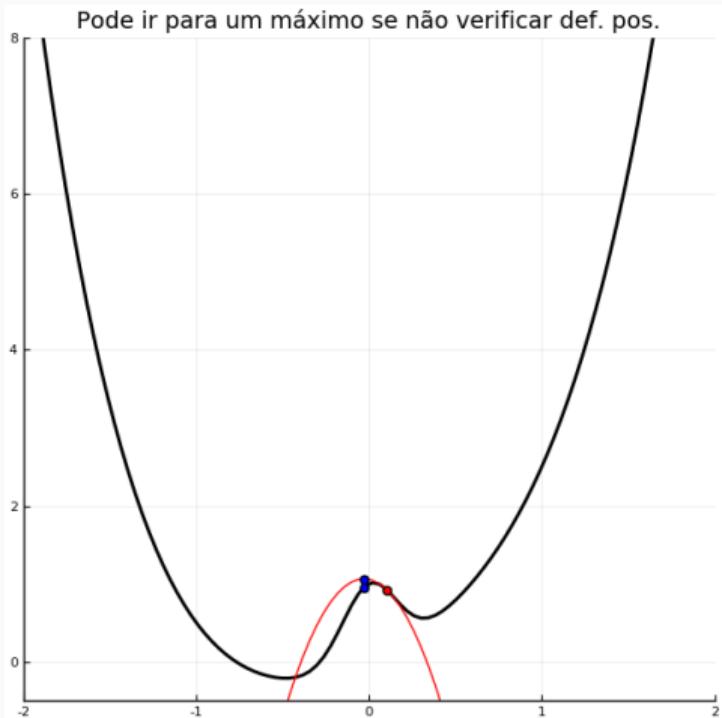
Método de Newton



Método de Newton



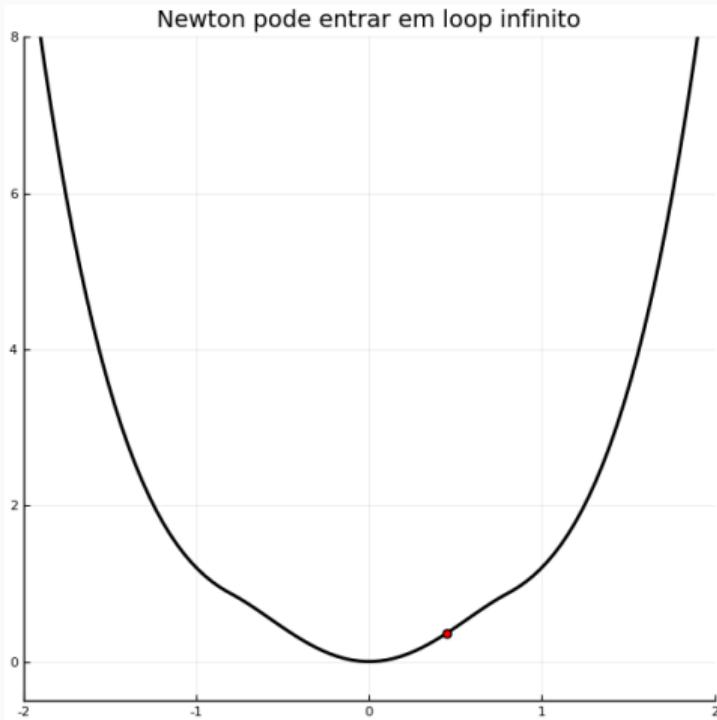
Método de Newton



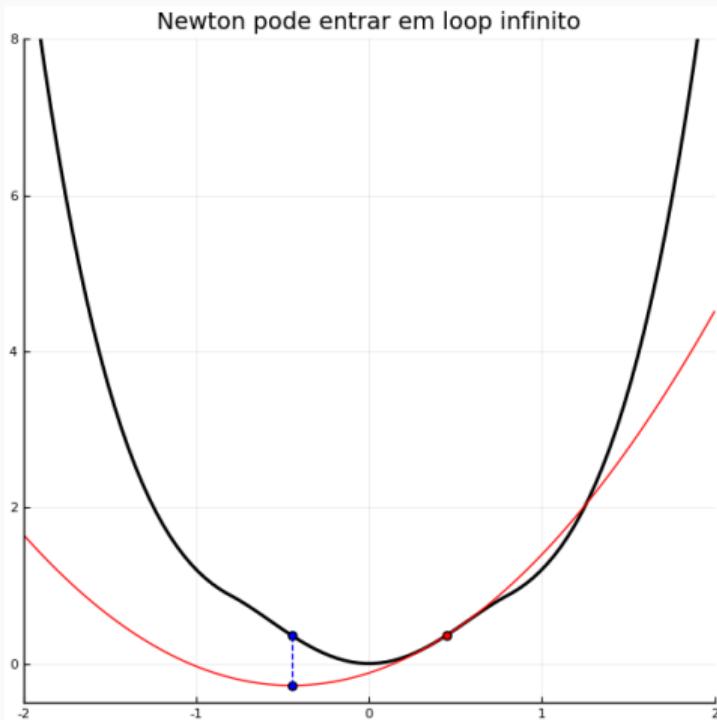
Método de Newton



Método de Newton

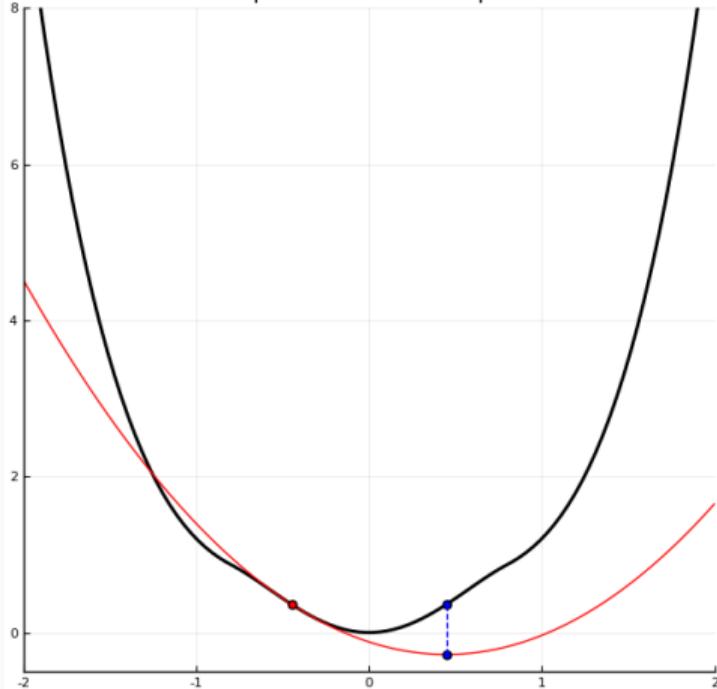


Método de Newton

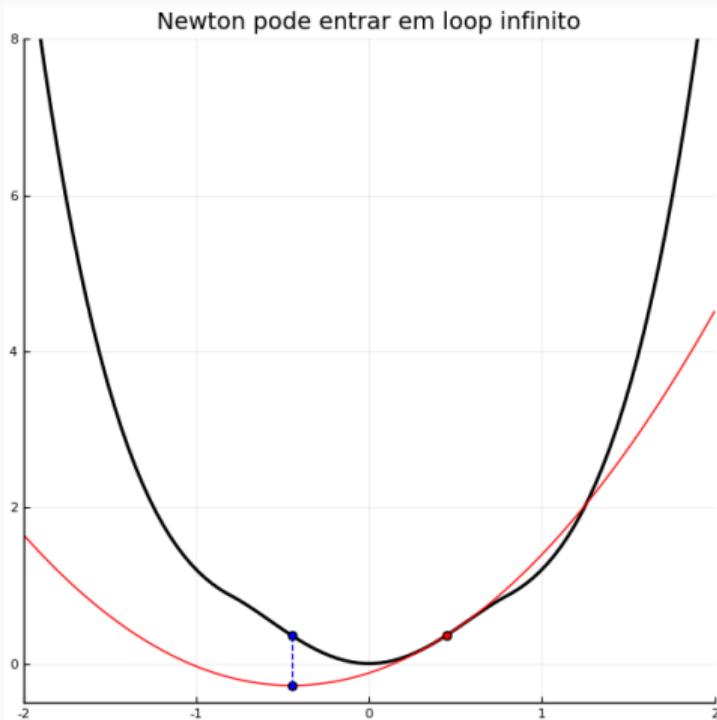


Método de Newton

Newton pode entrar em loop infinito

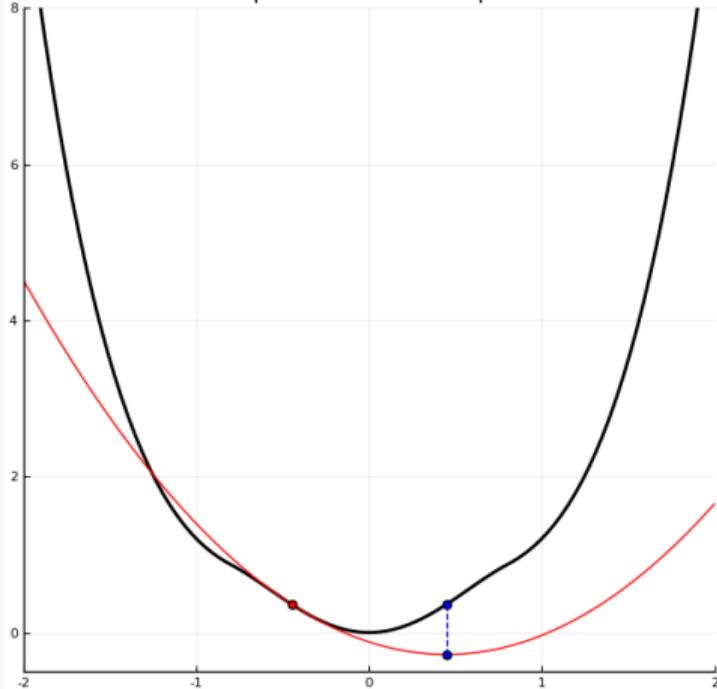


Método de Newton



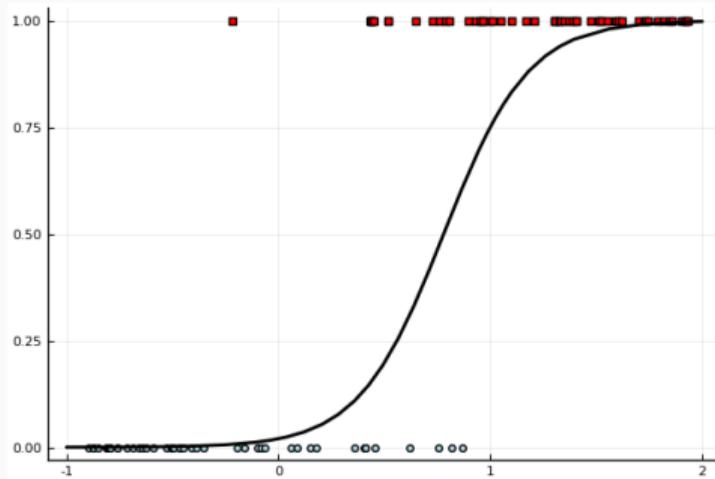
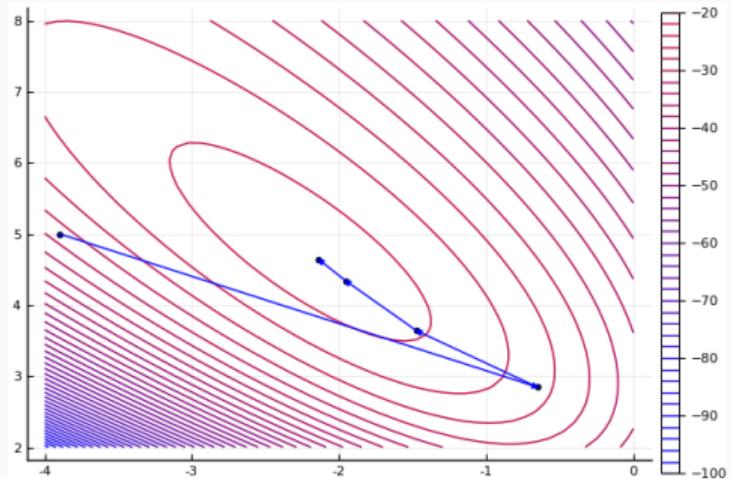
Método de Newton

Newton pode entrar em loop infinito



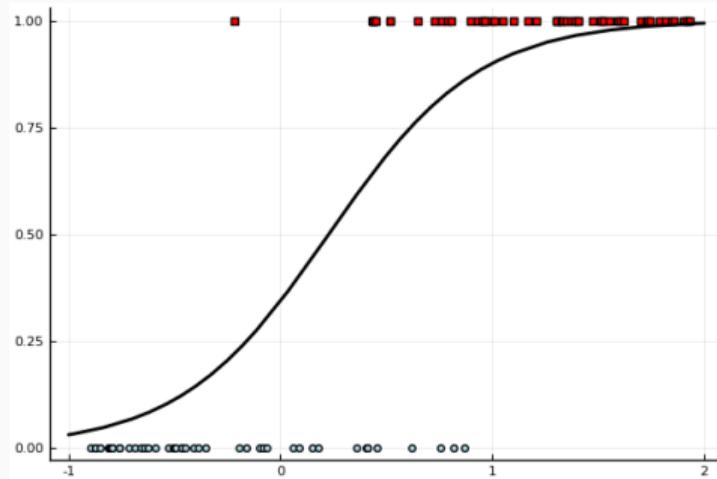
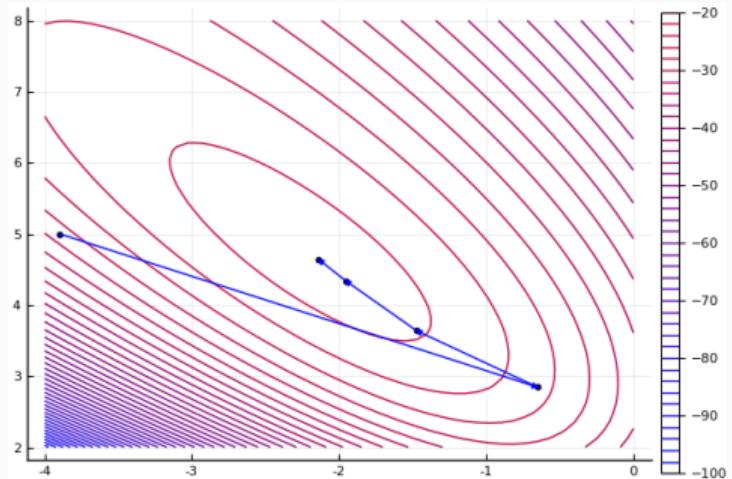
Método de Newton - Regressão Logística

Aproximação 0



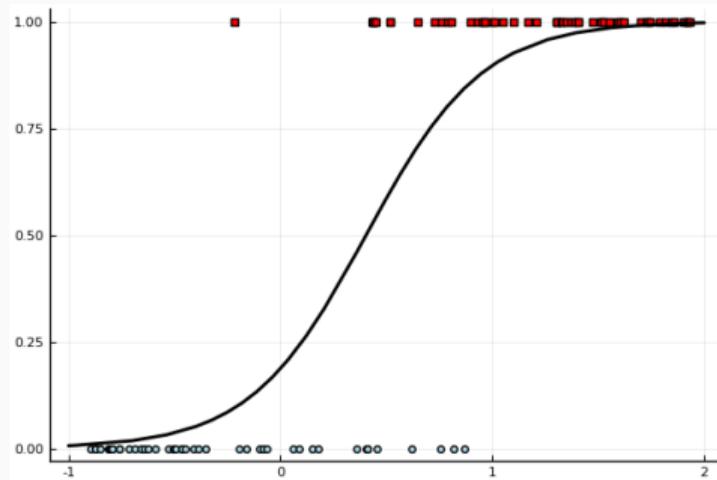
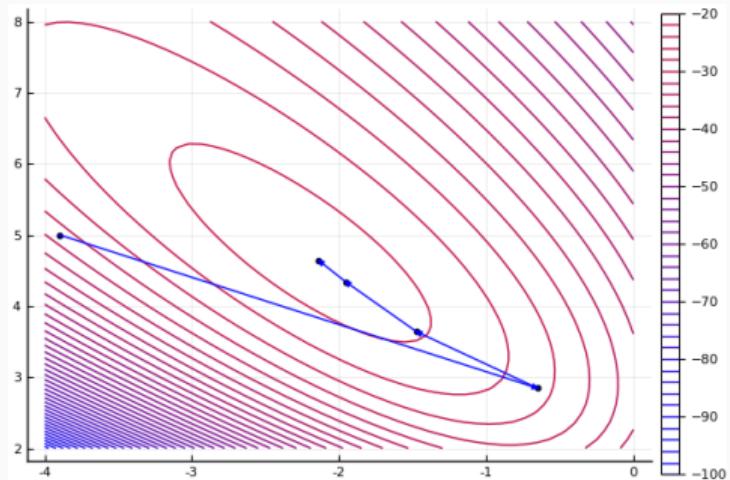
Método de Newton - Regressão Logística

Aproximação 1



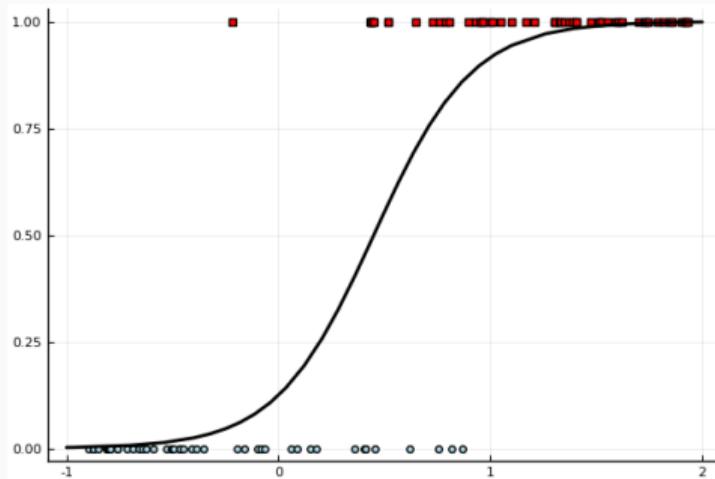
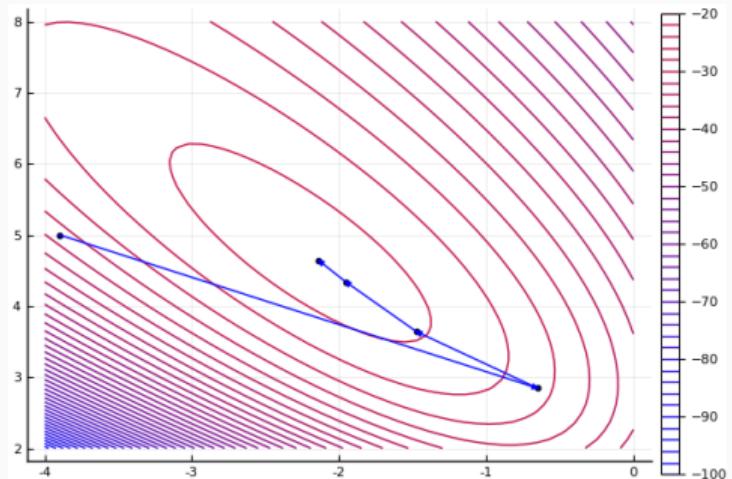
Método de Newton - Regressão Logística

Aproximação 2



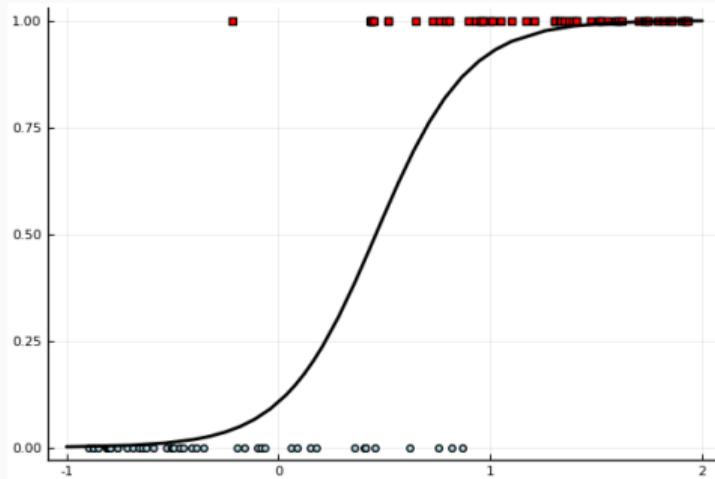
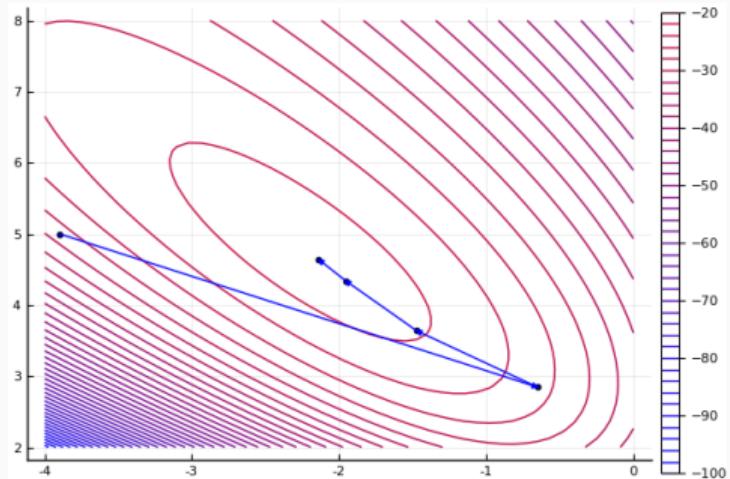
Método de Newton - Regressão Logística

Aproximação 3

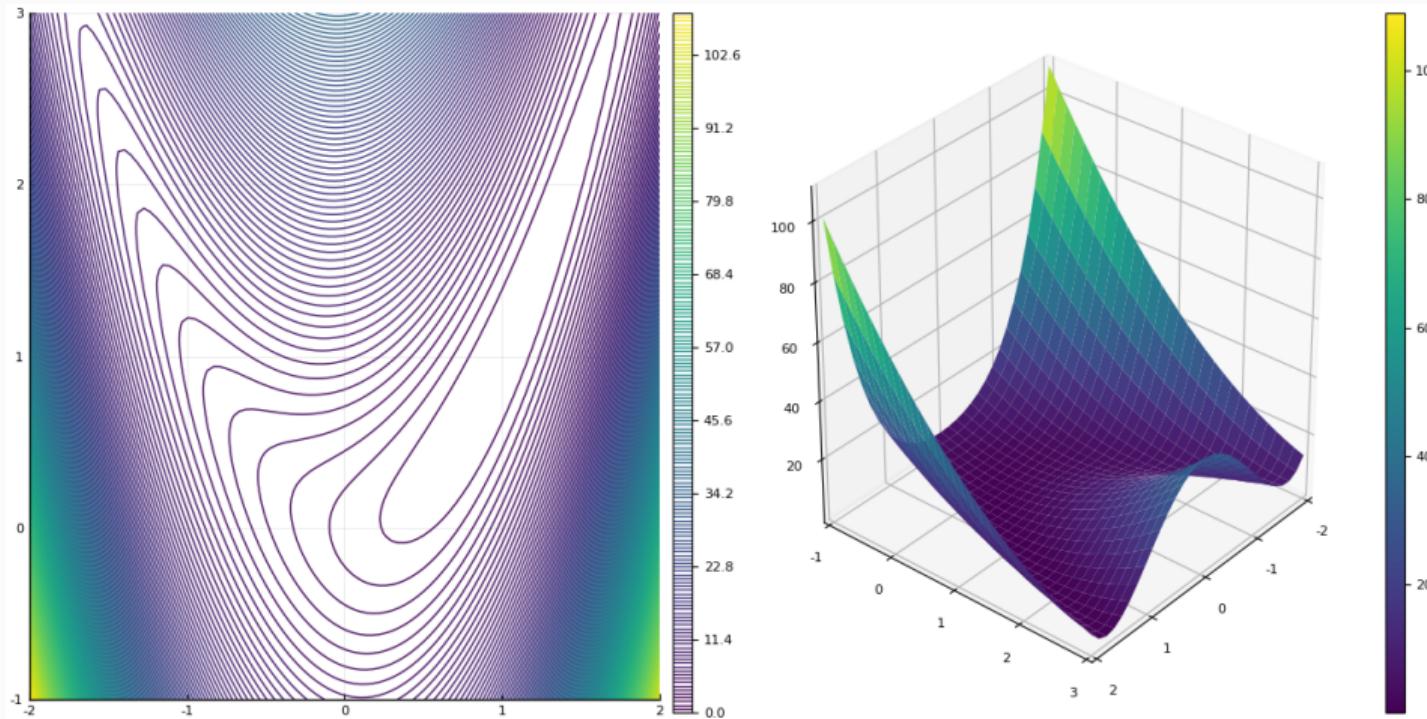


Método de Newton - Regressão Logística

Aproximação 4



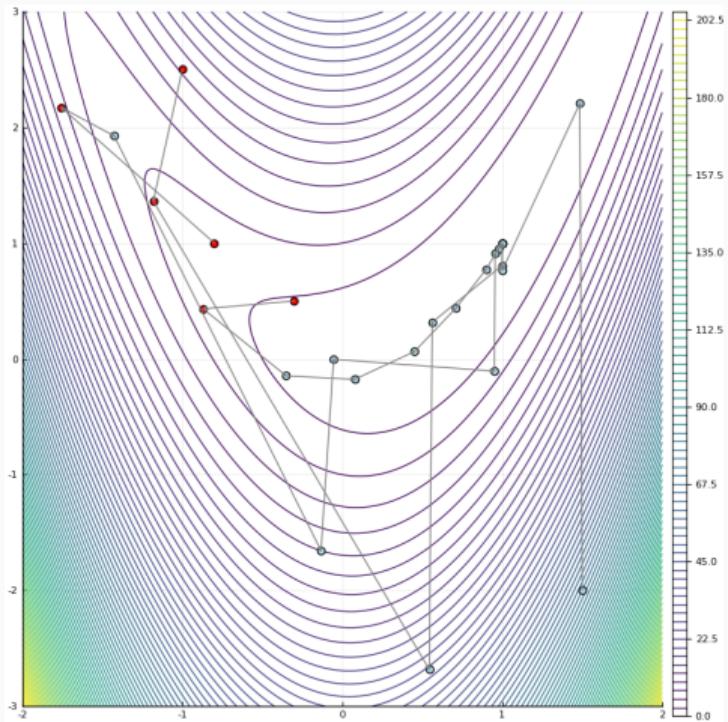
Função de Rosenbrock - $f(x) = (x_1 - 1)^2 + b^2(x_2 - x_1^2)^2$



Função de Rosenbrock - $f(x) = (x_1 - 1)^2 + b^2(x_2 - x_1^2)^2$

- Vários pontos com a matriz não definida positiva.
- Existe um único minimizador global estrito $(1, 1)$.
- Existe um vale em $x_2 = x_1^2$.

Função de Rosenbrock - $f(x) = (x_1 - 1)^2 + b^2(x_2 - x_1^2)^2$



Condições de Optimalidade

- Supondo que Newton está funcionando, ele vai parar quando?
- O passo será nulo quando $d_k = -B_k^{-1}g_k = 0$, i.e., se $g_k = 0$.
- Isso concorda com o vértice da quadrática.
- Quer dizer, $\nabla f(x_k) = 0$.
- **Def.:** Um ponto onde $\nabla f(x) = 0$ é chamado **ponto crítico** ou **estacionário** de primeira ordem.
- **Def.:** Um ponto crítico que não é minimizador nem maximizador é chamado **ponto de sela**.

Condições de Optimalidade

Teo. (C. Necessárias de 1^a ordem): Se x^* é um minimizador local de f e f é continuamente diferenciável em torno de x^* , então x^* é um ponto crítico, i.e., $\nabla f(x^*) = 0$.

Teo. (C. Necessárias de 2^a ordem): Se x^* é um minimizador local de f e f é continuamente diferenciável até segunda ordem em torno de x^* , então x^* é um ponto crítico e $\nabla^2 f(x^*)$ é semi-definida positiva.

Condições de Optimalidade

Teo. (C. Suficientes de 2^a ordem): Se f é continuamente diferenciável até segunda ordem em torno de x^* , um ponto crítico de f , e $\nabla^2 f(x^*)$ é definida positiva, então x^* é um minimizador local estrito de f .

Analogamente: Mudamos para maximizador e (semi-)definida negativa.

Teo.: Se f é continuamente diferenciável até segunda ordem em torno de x^* , um ponto crítico de f , e $\nabla^2 f(x^*)$ é indefinida, então x^* é um ponto de sela.

Método de Newton

Teo.: Seja f é continuamente diferenciável até segunda ordem em torno de x^* , um ponto crítico de x^* , e $\nabla^2 f(x^*)$ é definida positiva. Suponha que $\nabla^2 f$ é Lipschitz contínua em torno de x^* , i.e., existem $\delta > 0$ e $L > 0$ tais que

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|, \quad \forall x, y \in B(x^*, \delta).$$

Então, existe $\epsilon > 0$ tal que se $x_0 \in B(x^*, \epsilon)$, então

1. a sequência $\{x_k\}$ gerada pelo Método de Newton converge para x^* , com velocidade de convergência quadrática, i.e., existe $C > 0$ tal que

$$\|x_{k+1} - x^*\| \leq C\|x_k - x^*\|^2,$$

2. a sequência $\{f(x_k)\}$ converge para $f(x^*)$ quadraticamente.

Métodos de gradiente

Direção de descida e gradiente

- **Def.:** Uma direção d é dita **direção de descida** para a partir de x se existe $\bar{t} > 0$ tal que $f(x + tv) < f(x)$ para todo $t \in (0, \bar{t}]$.
- **Teo.:** Se $d^T \nabla f(x) < 0$, então d é uma direção de descida.
- Como $d^T \nabla f(x) = \|d\| \|\nabla f(x)\| \cos \theta$, então a direção que é o máximo de descida é $-\nabla f(x)$.
- Sendo assim, no lugar de usar o método de Newton, podemos usar a direção $-\nabla f(x)$.

Métodos de gradiente

- Os métodos de gradiente são definidos por

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k),$$

para alguma definição de $\alpha_k > 0$.

- O método de Cauchy é dado por $\alpha_k = \arg \min_{\alpha} f(x_k - \alpha \nabla f(x_k))$.
- A dificuldade do método de Cauchy é encontrar esse minizador.
- Uma estratégia é utilizar o método da seção áurea.

Método de busca unidimensional

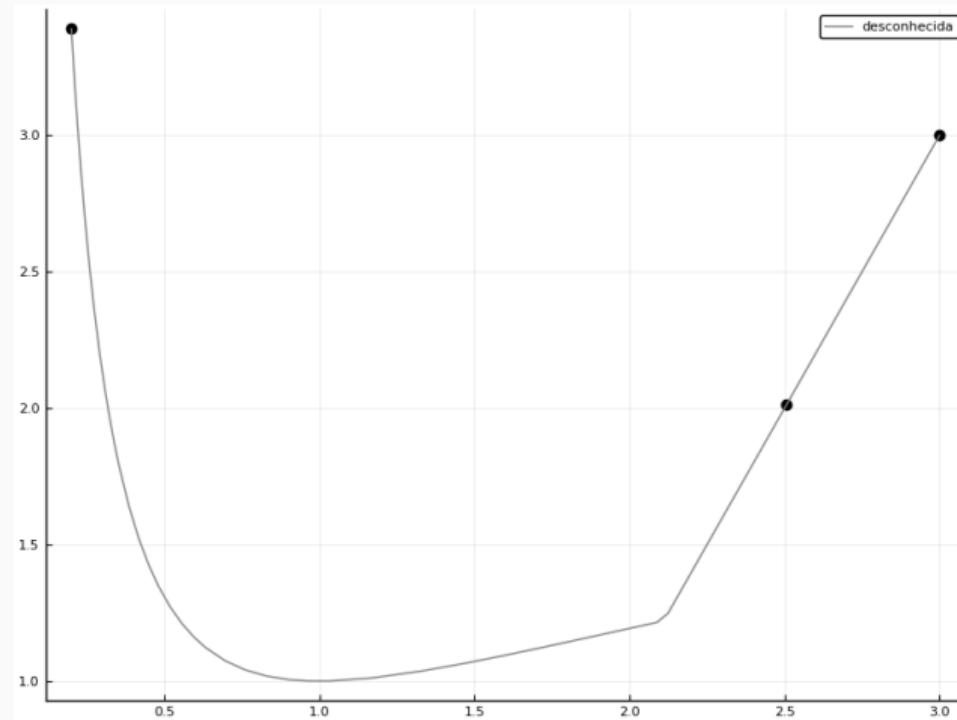
- Suponha que $f : [a, b] \rightarrow \mathbb{R}$ é contínua e que $f(c) < \min\{f(a), f(b)\}$, com $c \in (a, b)$.
- Isso quer dizer que f tem um minimizador em (a, b) .
- Tome $d \in (a, b)$, $d \neq c$. Para simplificar, assuma $d < c$.
- Se $f(d) < f(c)$, então no intervalo (a, c) tem um minimizador.
- Se $f(d) > f(c)$, então no intervalo (d, b) tem um minimizador.
- Ou seja, com 3 pontos, o do meio sendo menor, temos um minimizador. Pegamos um novo ponto e escolhemos 3 dos 4 tais que o do meio seja menor.

Método de busca aleatória

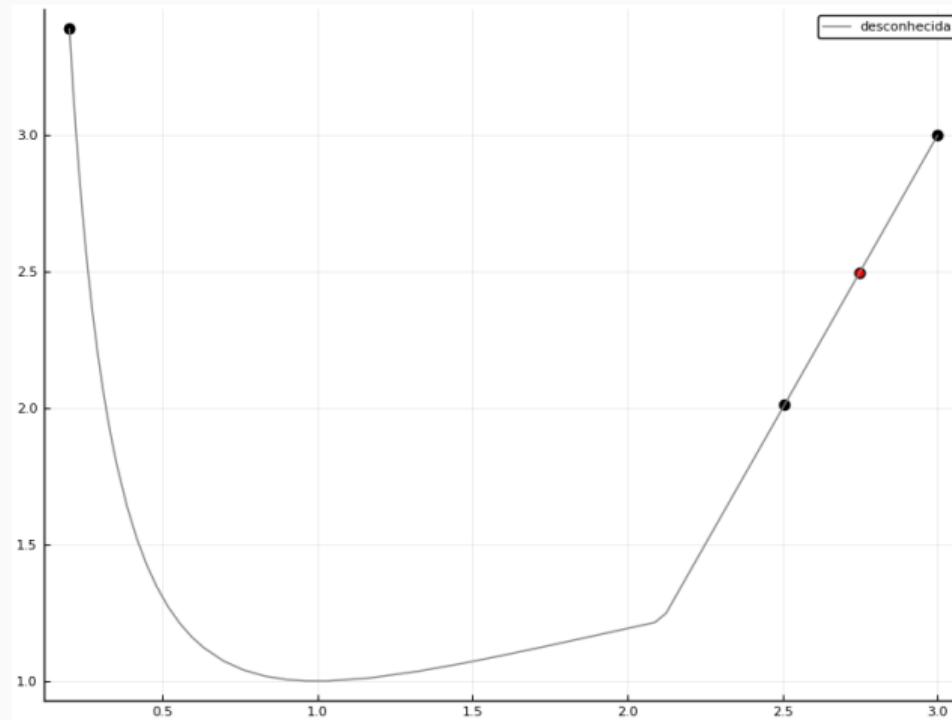
Definimos o método de busca aleatória:

1. Dados $f : [a, b] \rightarrow \mathbb{R}$ contínua, $\varepsilon > 0$.
2. Defina $a_0 = a$, $b_0 = b$, $k = 0$.
3. Escolha $c_0 \in (a, b)$ arbitrário.
4. Enquanto $b_k - a_k > \epsilon$
 - 4.1 Escolha $d_k \in (a, b)$, $d_k \neq c_k$ arbitrário.
 - 4.2 Renomeie c_k e d_k tais que $d_k < c_k$.
 - 4.3 Se $f(d_k) < f(c_k)$, renomeie $d_k \rightarrow c_k$ e $c_k \rightarrow b_k$.
 - 4.4 Se $f(d_k) \geq f(c_k)$, renomeie $d_k \rightarrow a_k$.

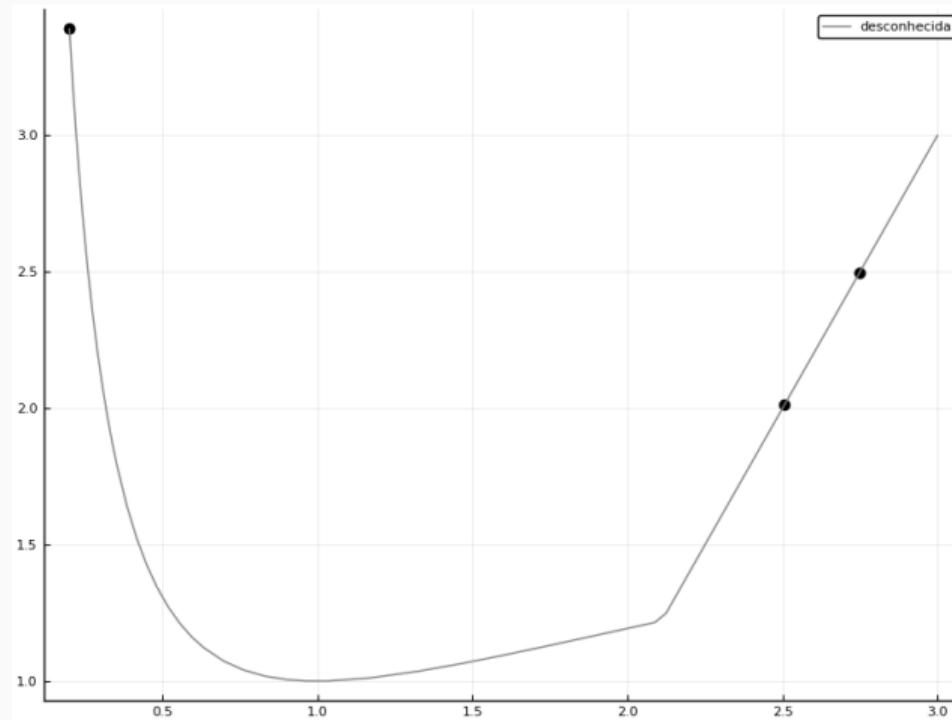
Método de busca aleatória



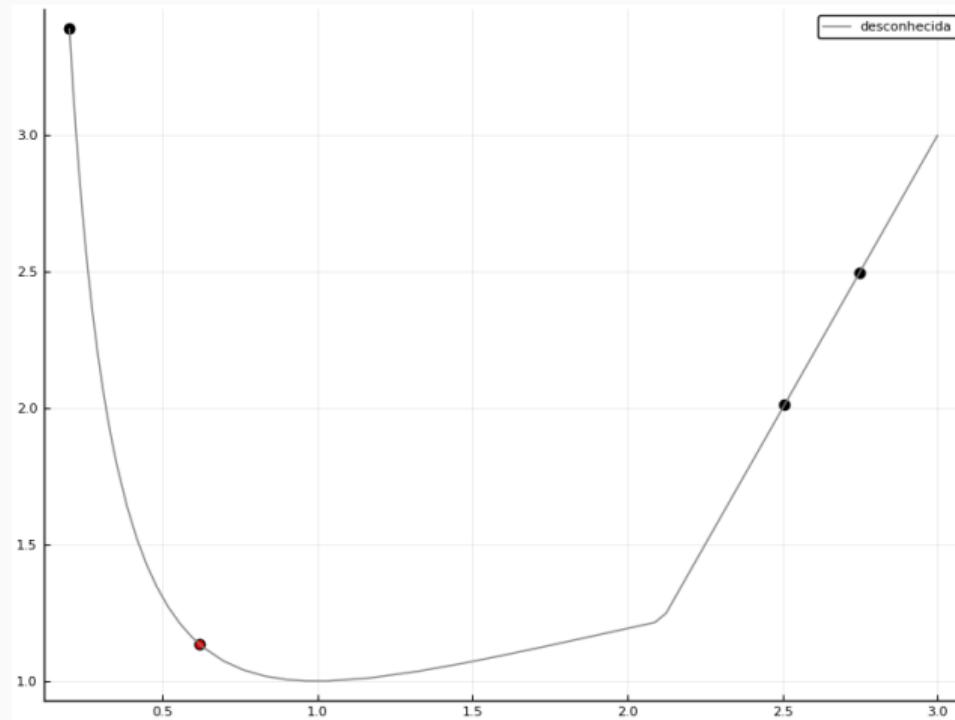
Método de busca aleatória



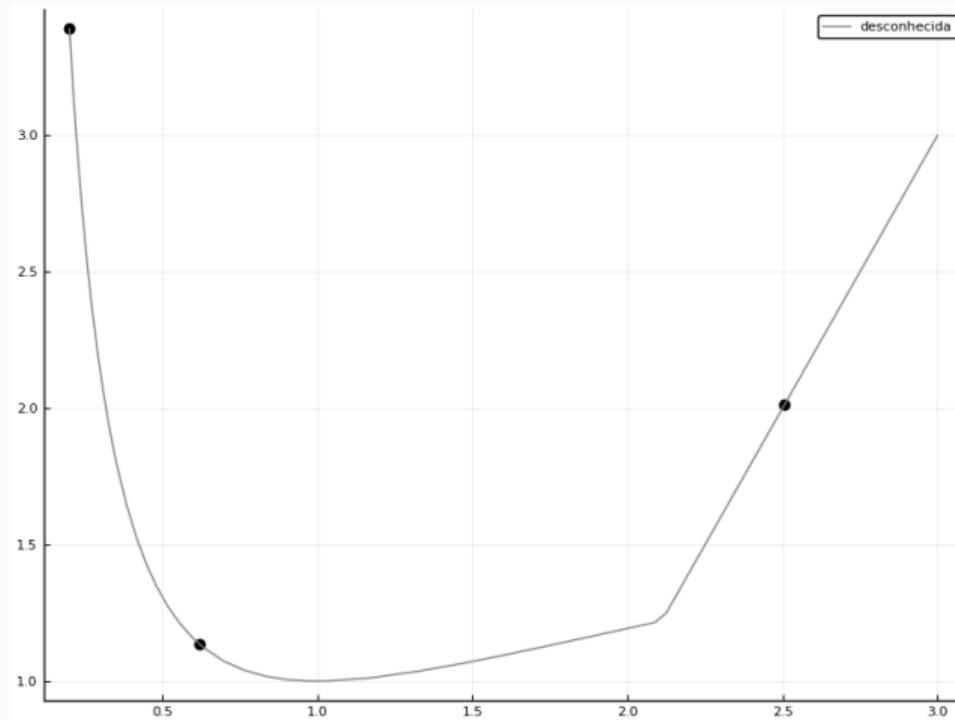
Método de busca aleatória



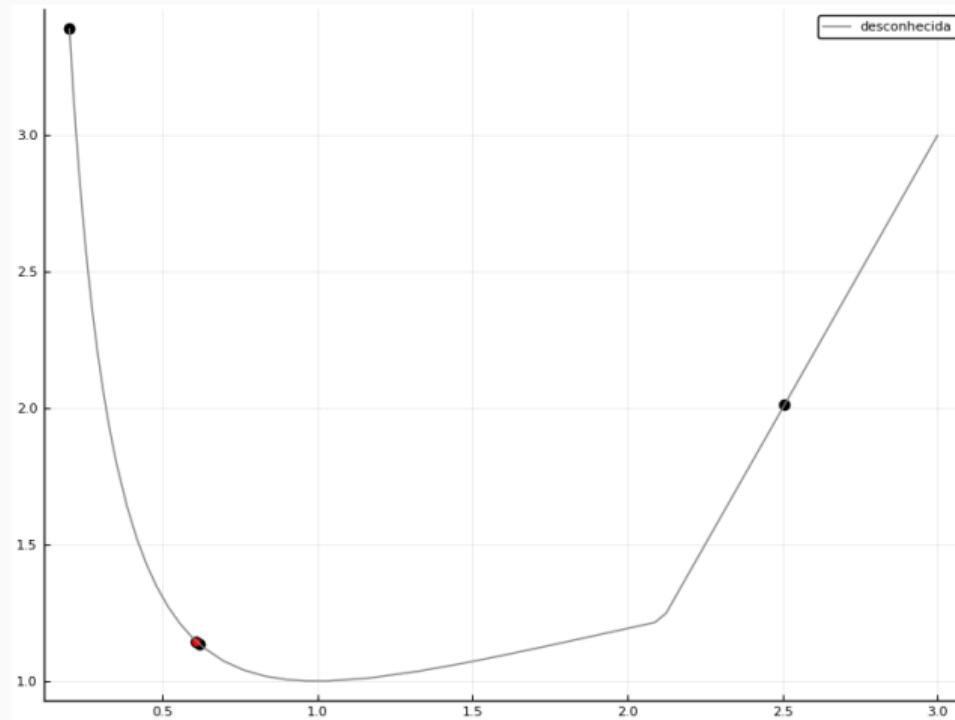
Método de busca aleatória



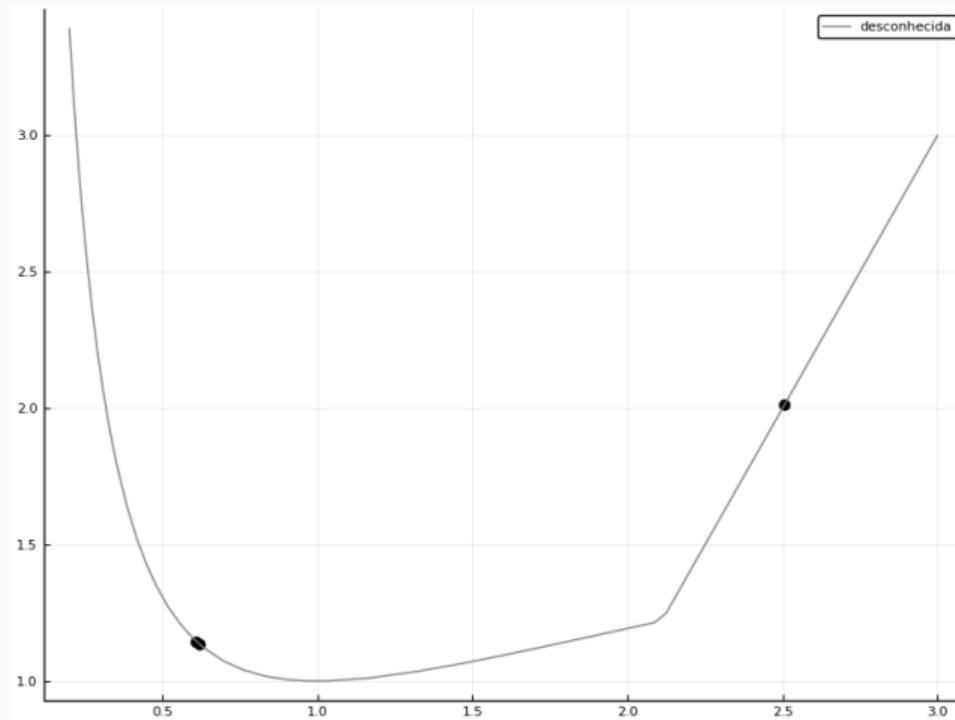
Método de busca aleatória



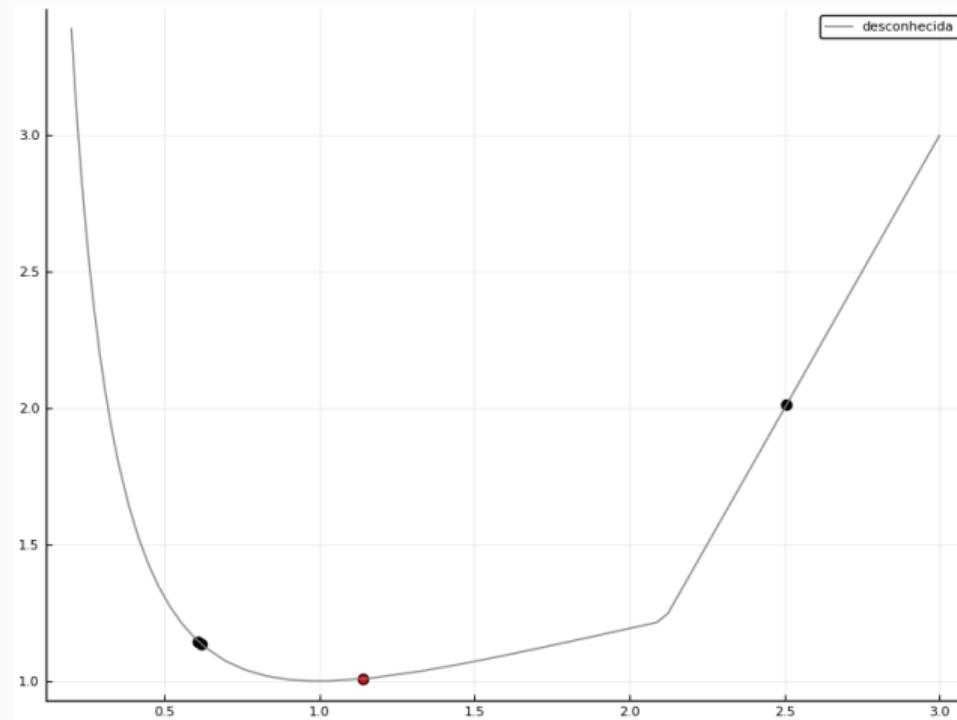
Método de busca aleatória



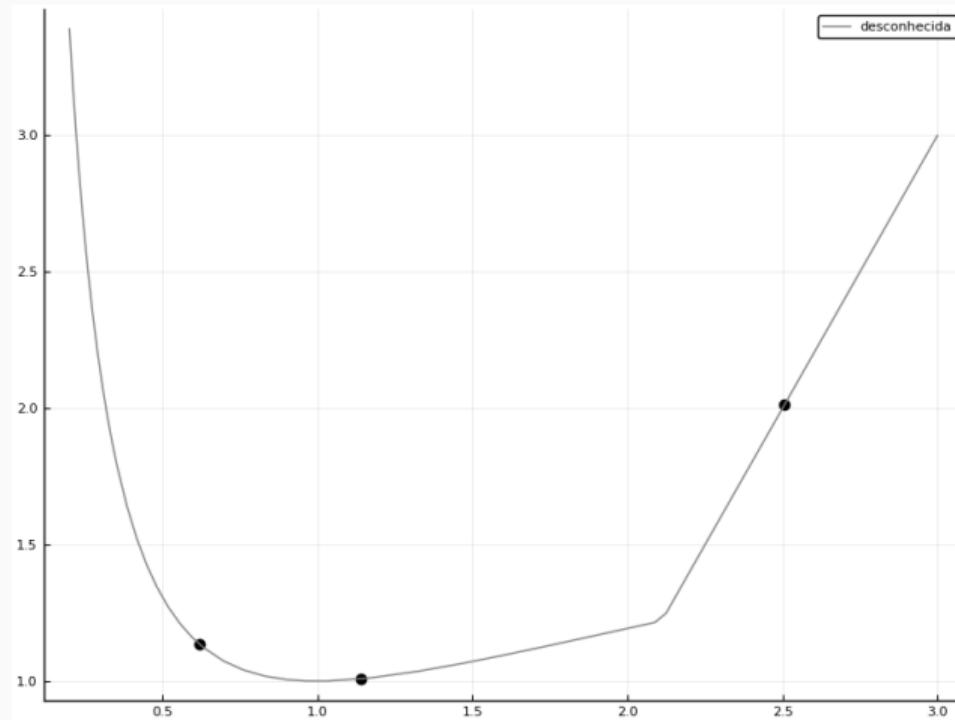
Método de busca aleatória



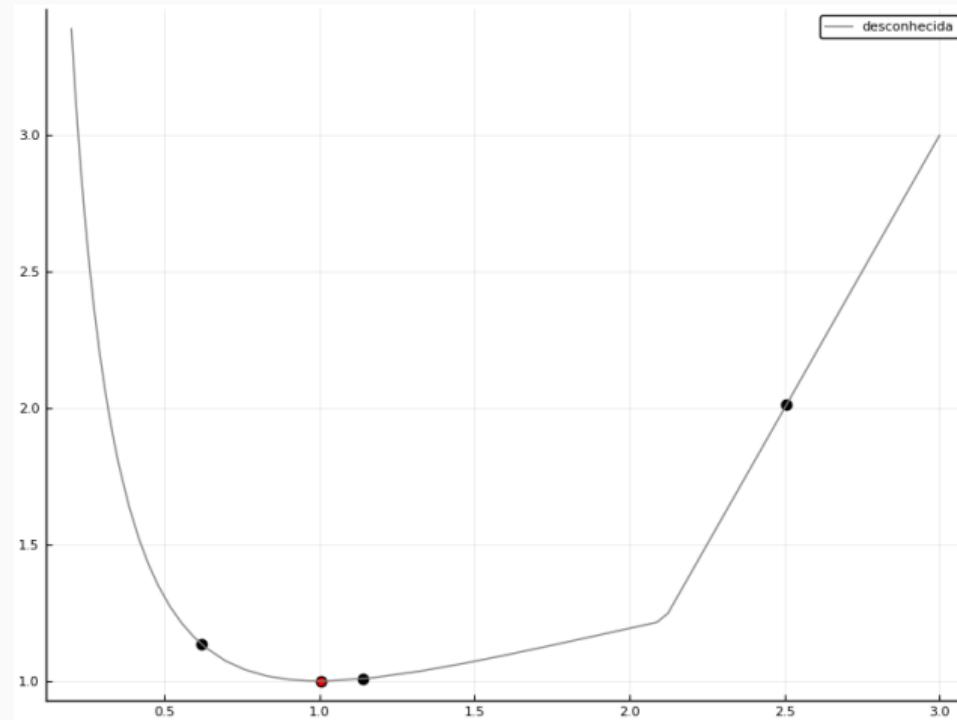
Método de busca aleatória



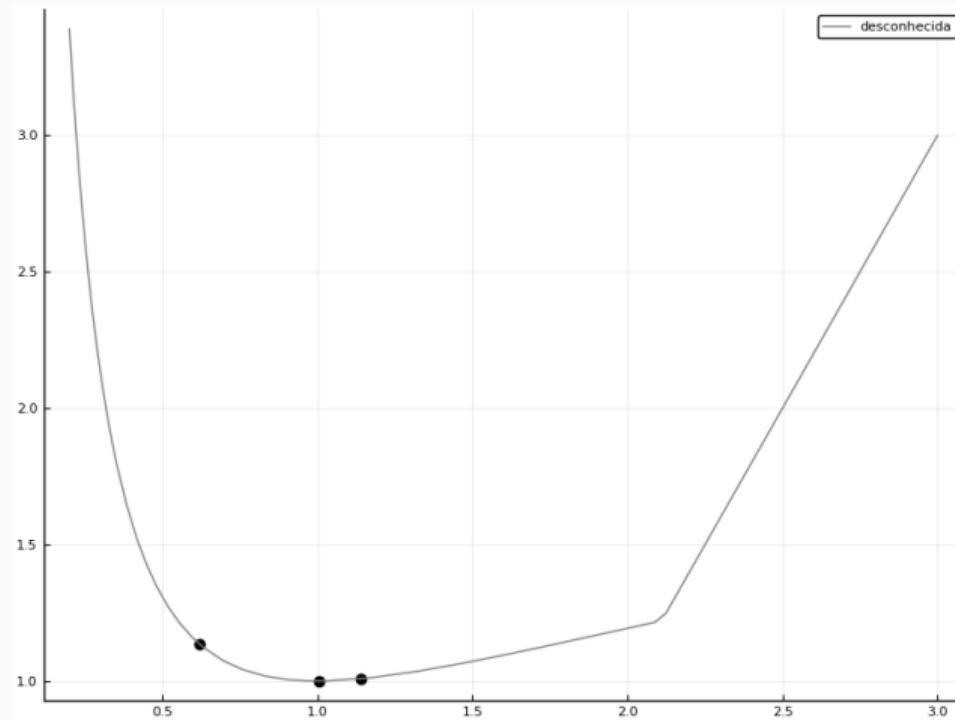
Método de busca aleatória



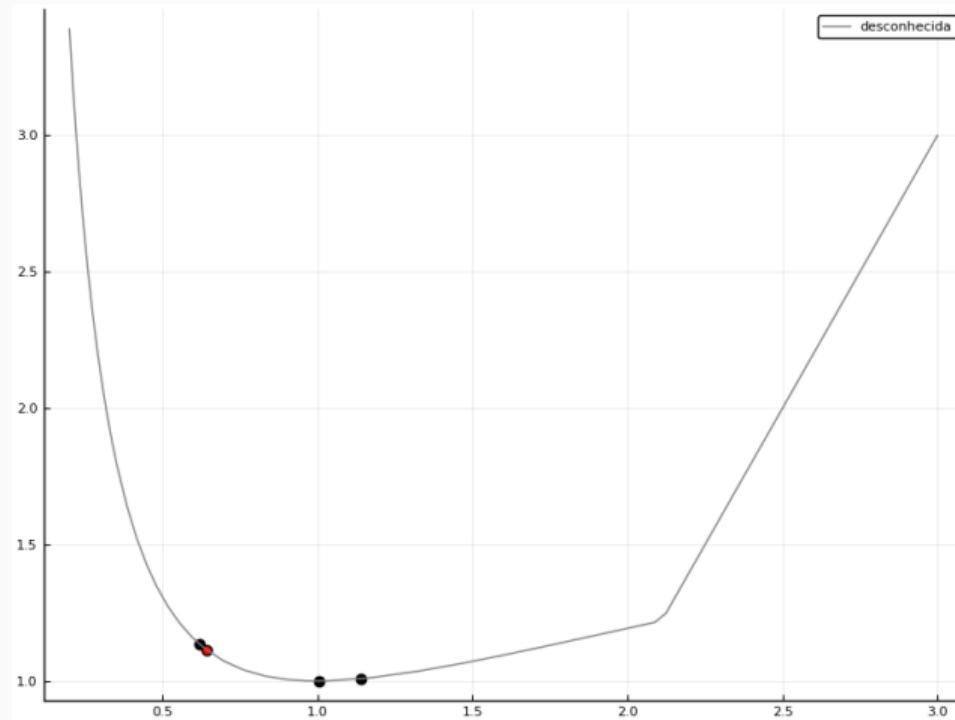
Método de busca aleatória



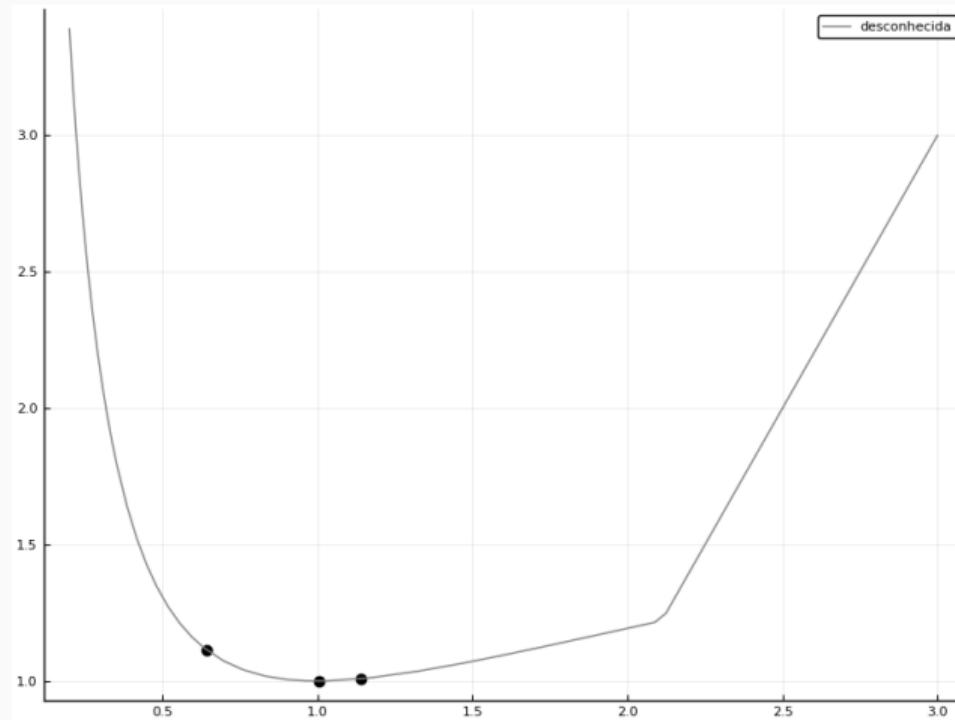
Método de busca aleatória



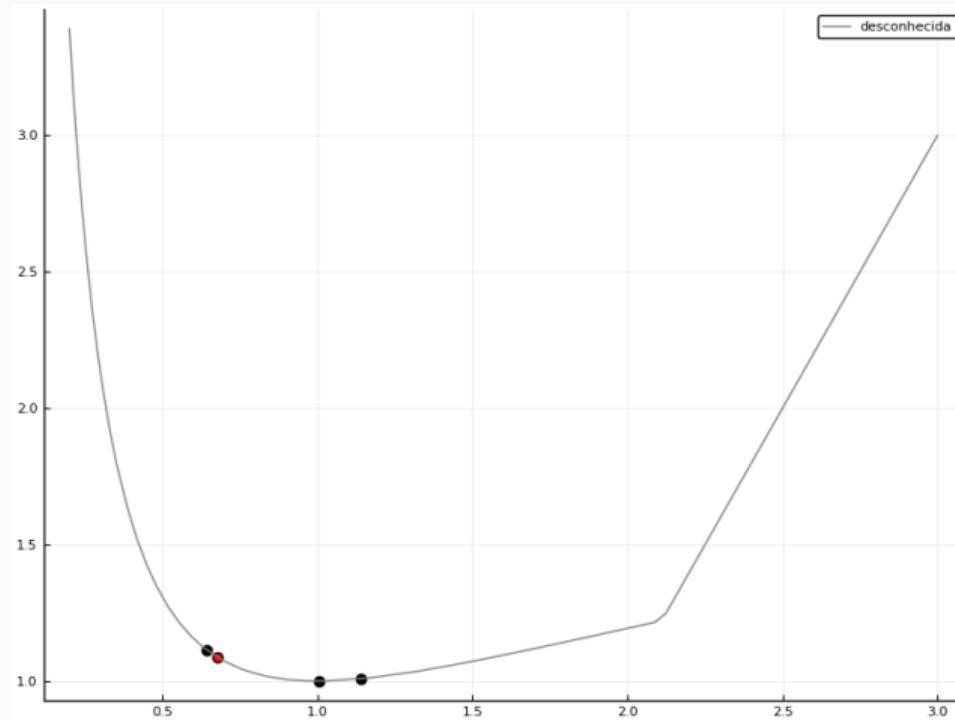
Método de busca aleatória



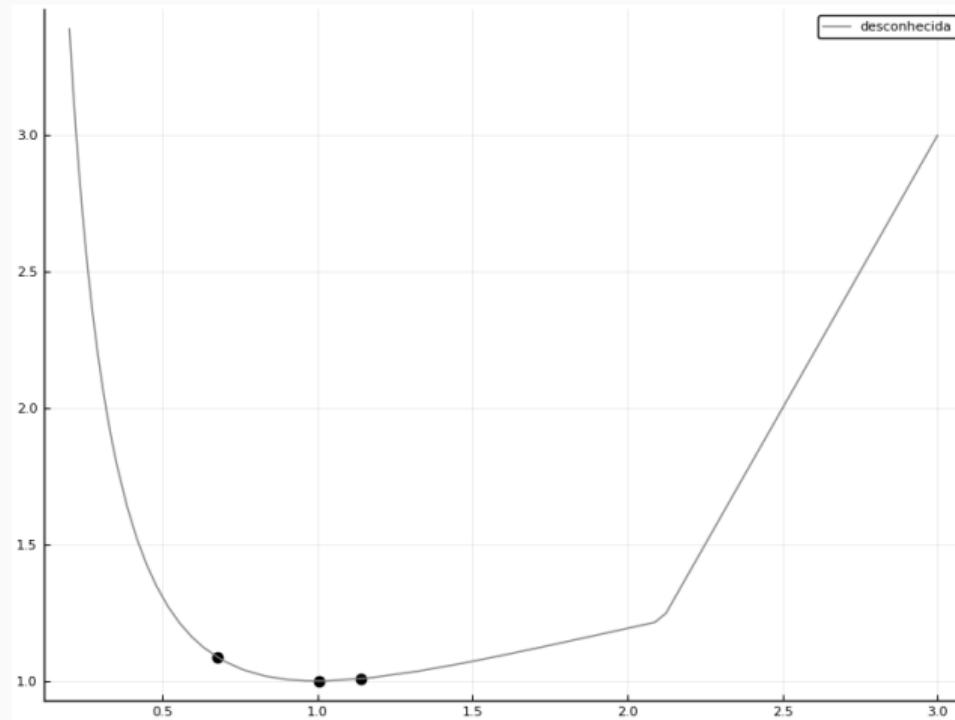
Método de busca aleatória



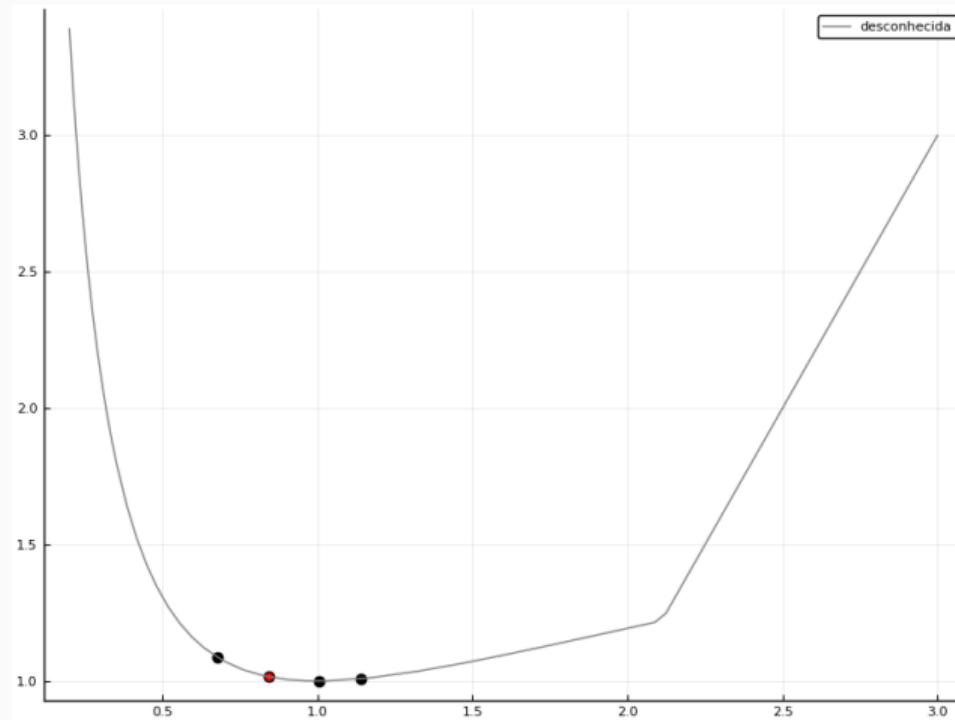
Método de busca aleatória



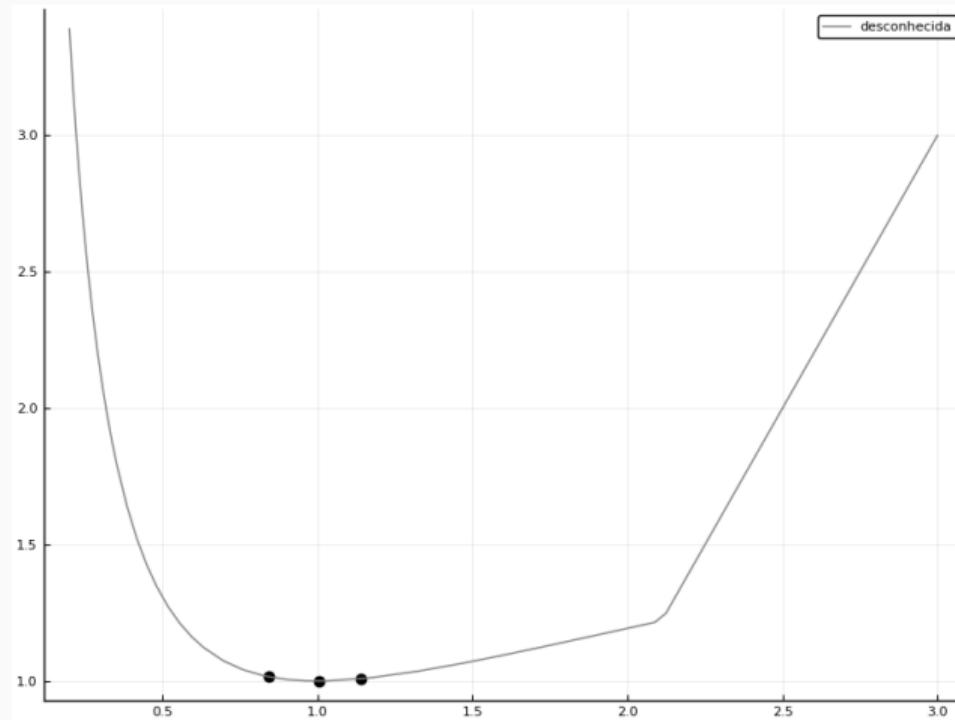
Método de busca aleatória



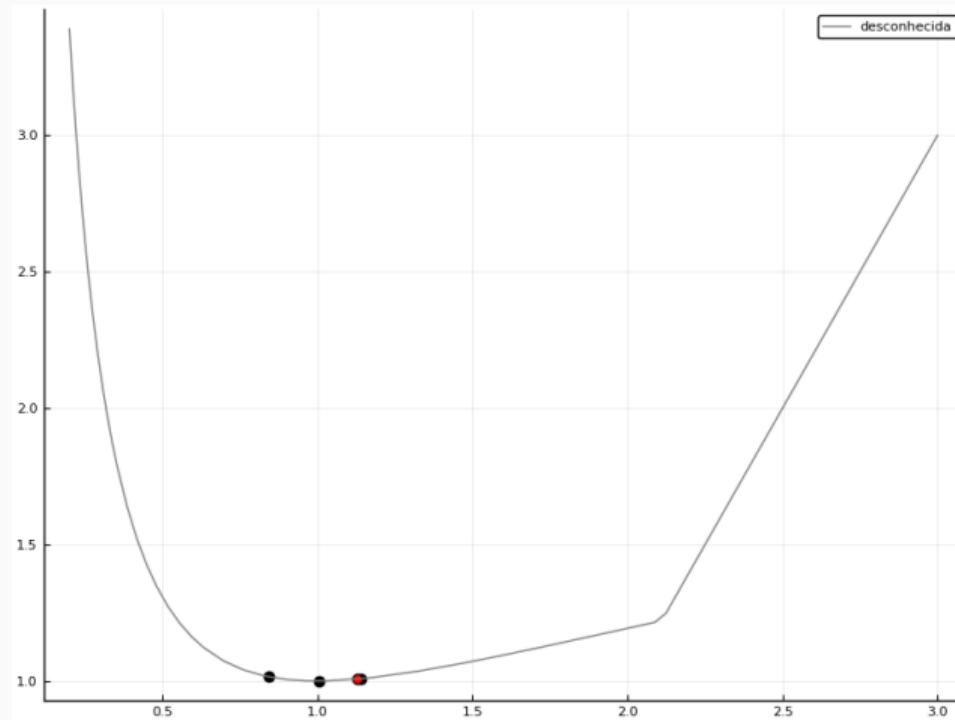
Método de busca aleatória



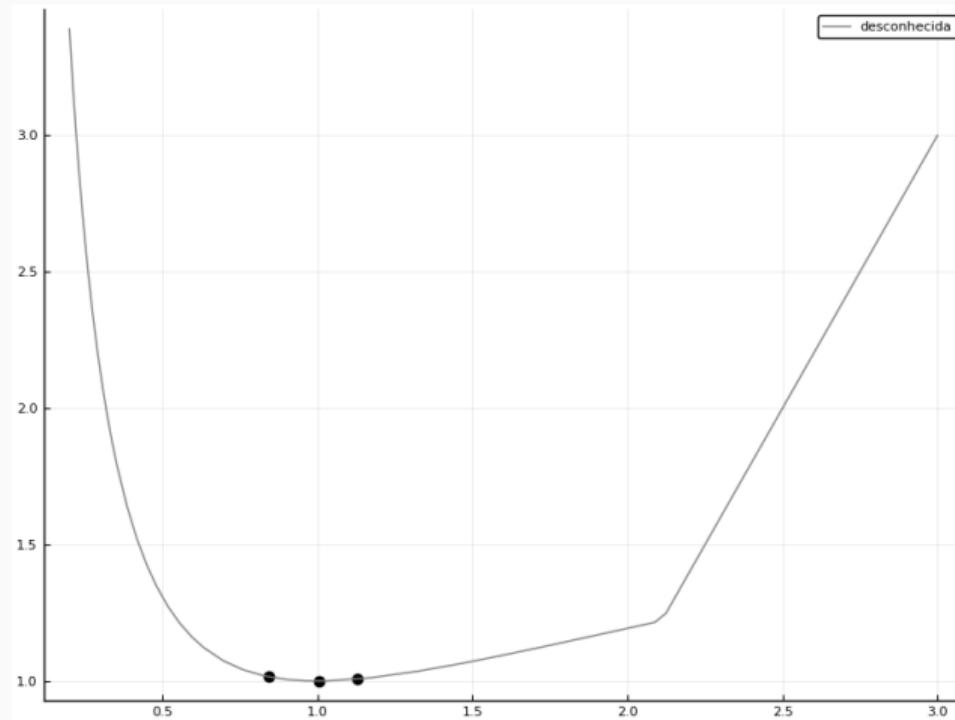
Método de busca aleatória



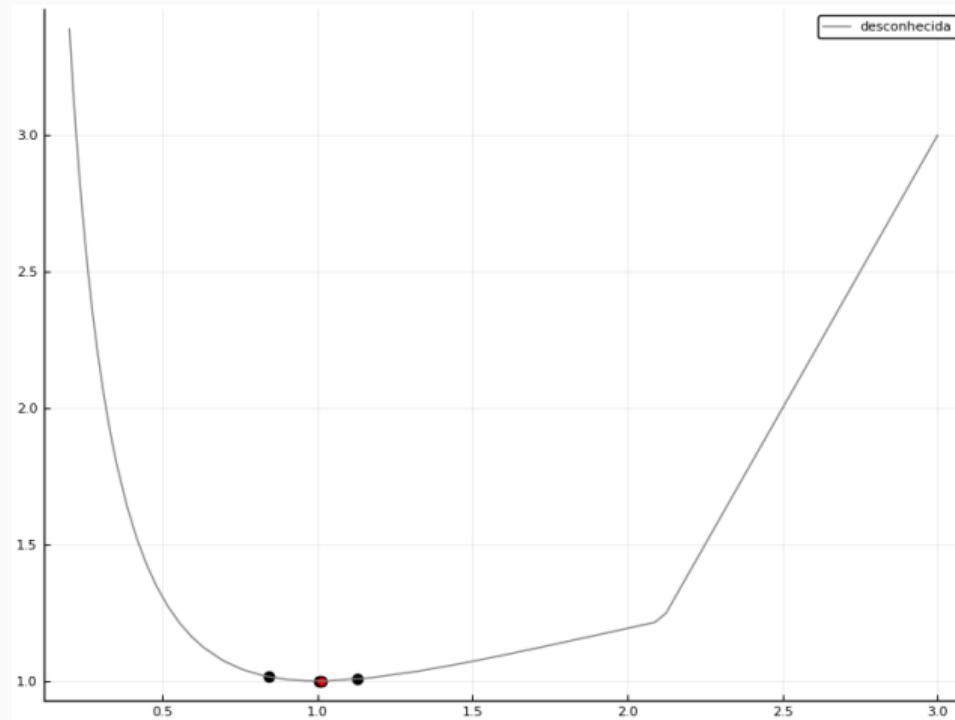
Método de busca aleatória



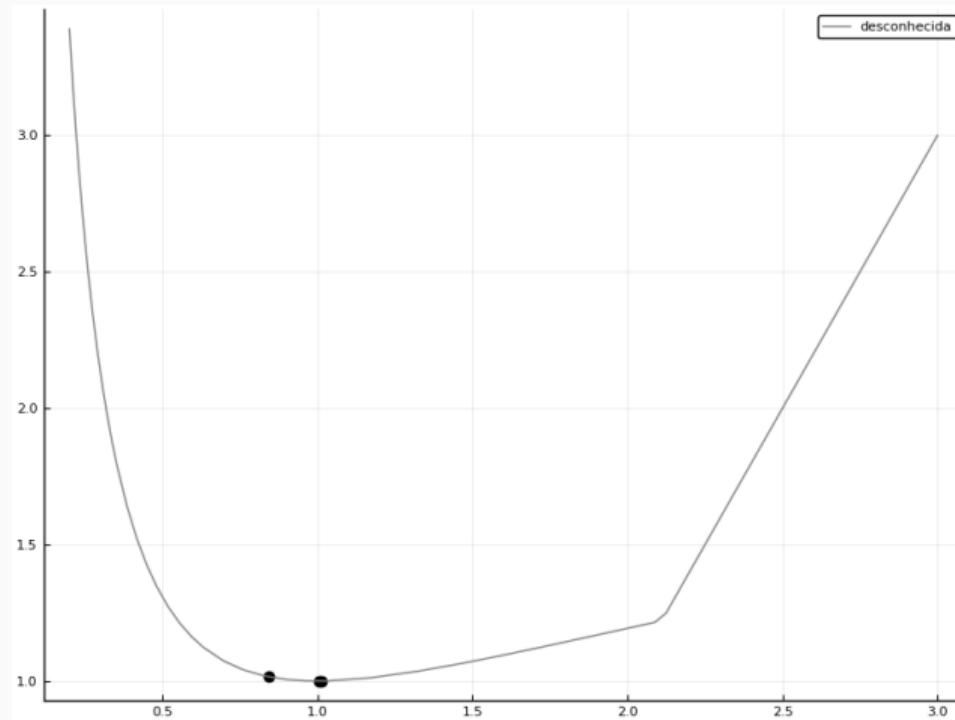
Método de busca aleatória



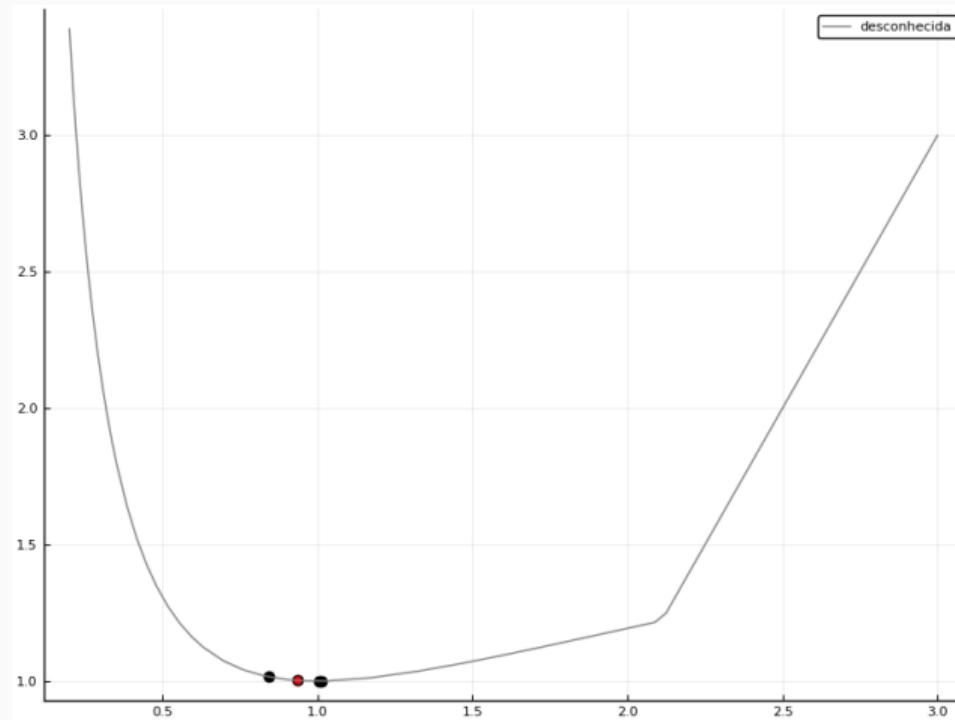
Método de busca aleatória



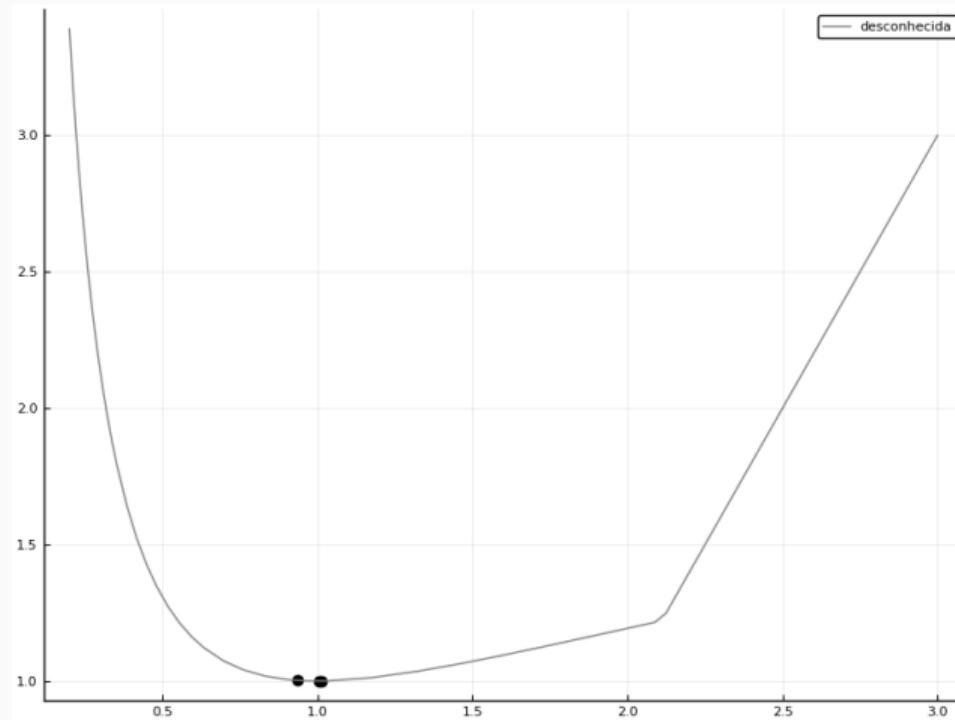
Método de busca aleatória



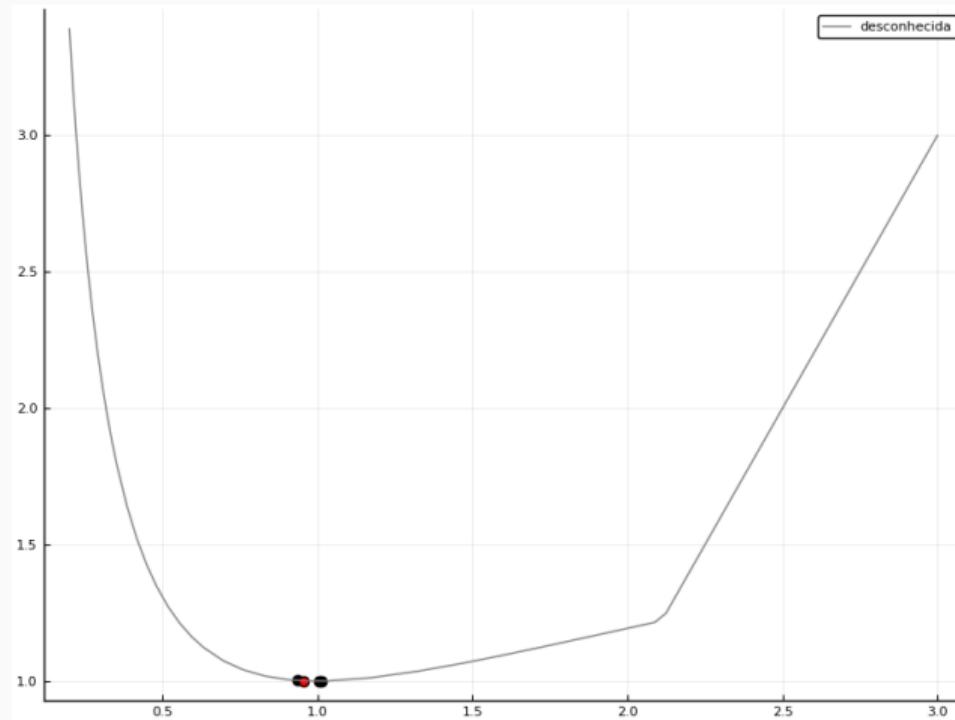
Método de busca aleatória



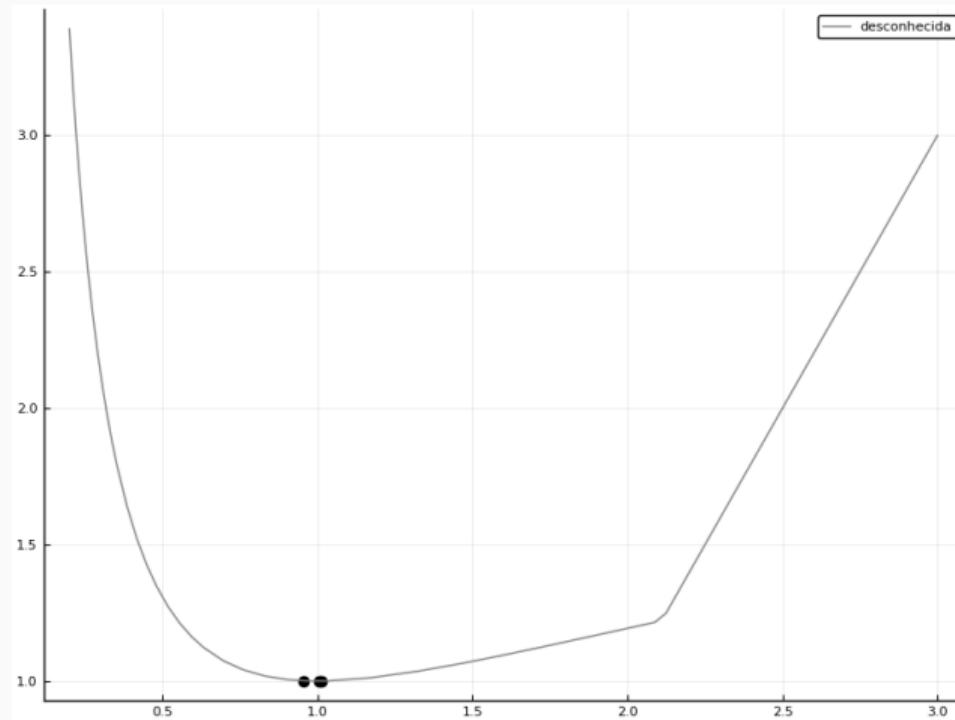
Método de busca aleatória



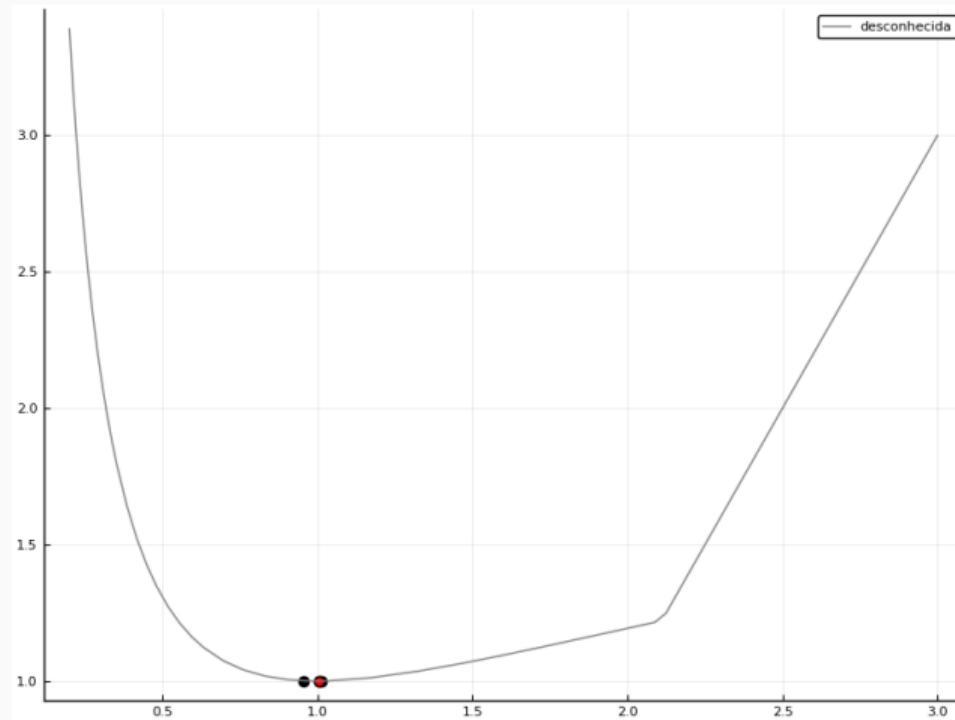
Método de busca aleatória



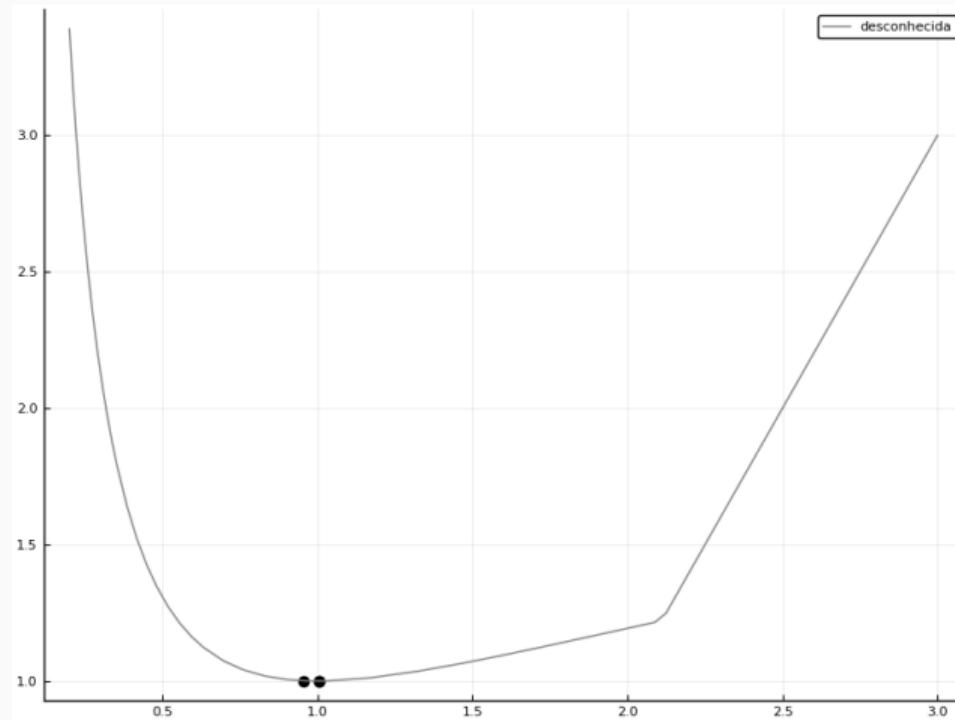
Método de busca aleatória



Método de busca aleatória



Método de busca aleatória



Método da seção áurea

- A cada iteração, um dos intervalos $[a_k, d_k]$ ou $(c_k, b_k]$ é descartado.
- Com escolhas aleatórias, podemos dar sorte e eliminar uma grande parte do intervalo, ou um pedaço muito pequeno.
- Uma alternativa é não escolher aleatoriamente. Por exemplo, se escolhemos $d_k = \frac{1}{3}(b_k - a_k)$ e $c_k = \frac{2}{3}(b_k - a_k)$, sempre removemos um pedaço razoável.
- No entanto, a escolha acima não permite aproveitar d_k e c_k da iteração anterior.
- O método da razão áurea escolhe d_k e c_k com uma proporção tal que na próxima iteração eles podem ser reutilizados.

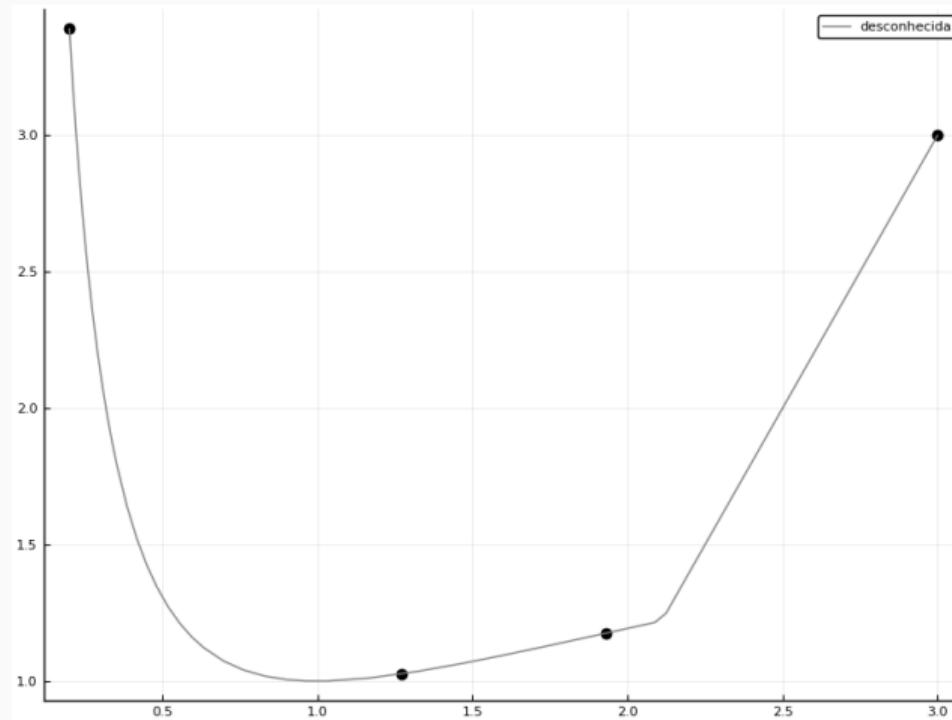
Método da seção áurea

- Temos $\frac{b_k - a_k}{b_k - d_k} = \alpha = \frac{b_k - a_k}{c_k - a_k}$, com a ordem a_k, d_k, c_k, b_k .
- Na próxima iteração, se ficarmos com a_k, d_k e c_k , definimos $(a_{k+1}, c_{k+1}, b_{k+1}) = (a_k, d_k, c_k)$.
- Com $\frac{b_{k+1} - a_{k+1}}{c_{k+1} - a_{k+1}} = \alpha$, temos

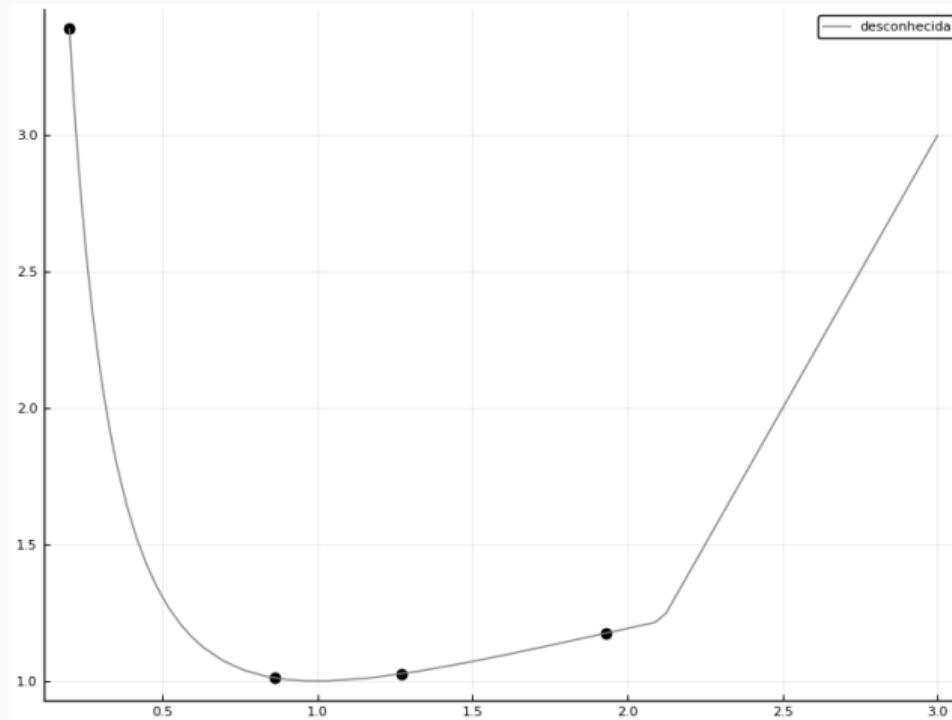
$$\alpha = \frac{b_{k+1} - a_{k+1}}{c_{k+1} - a_{k+1}} = \frac{c_k - a_k}{d_k - a_k} = \frac{a_k + \alpha^{-1}(b_k - a_k) - a_k}{b_k - \alpha^{-1}(b_k - a_k) - a_k} = \frac{\alpha^{-1}}{1 - \alpha^{-1}} = \frac{1}{\alpha - 1}.$$

- Daí, $\alpha^2 - \alpha - 1 = 0$, isto é, $\alpha = \frac{1 \pm \sqrt{5}}{2}$. Como $\alpha > 0$ para que os valores caiam dentro do intervalo, temos $\alpha = (1 + \sqrt{5})/2$, i.e., a **razão áurea**.

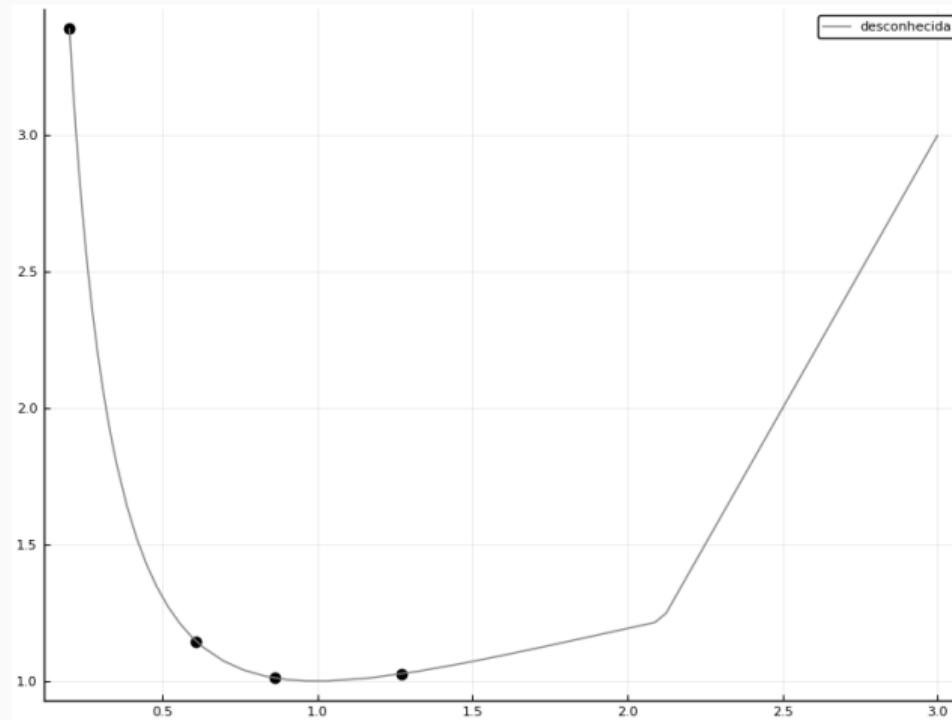
Método da seção áurea



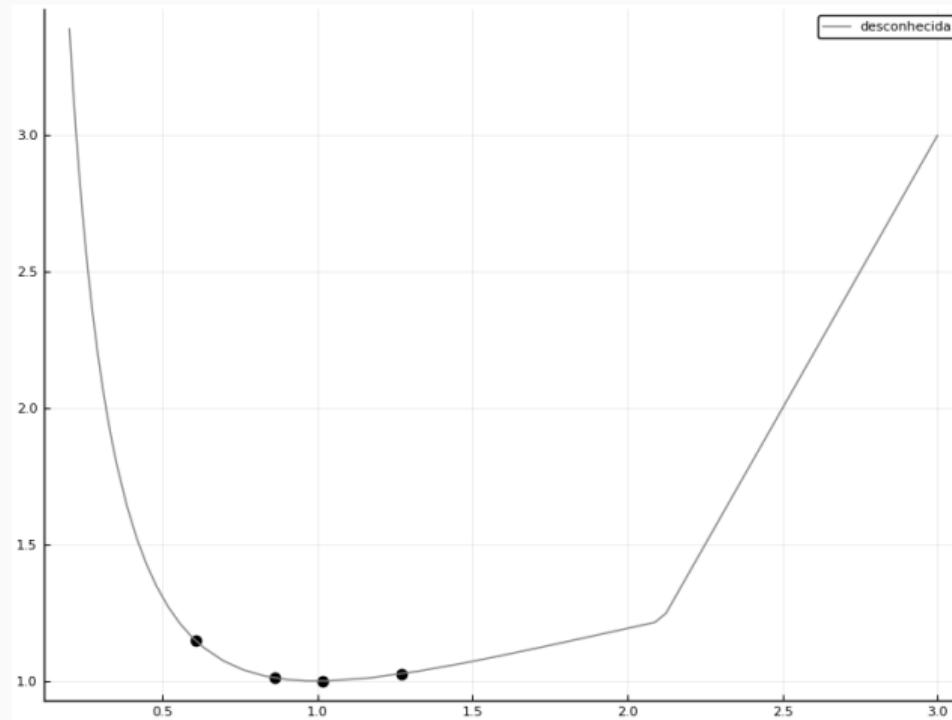
Método da seção áurea



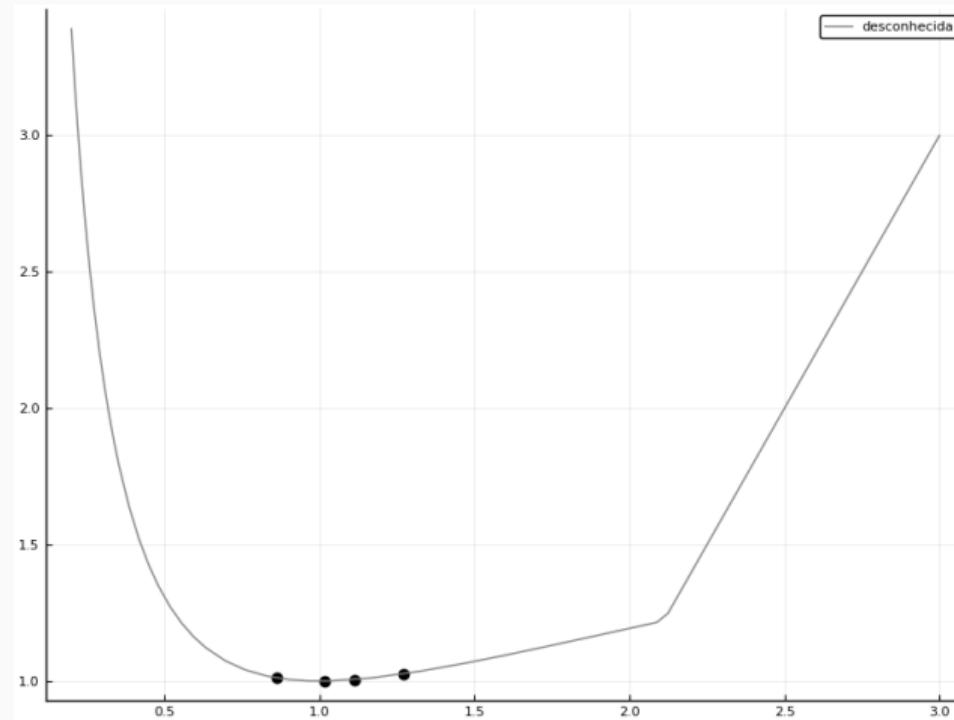
Método da seção áurea



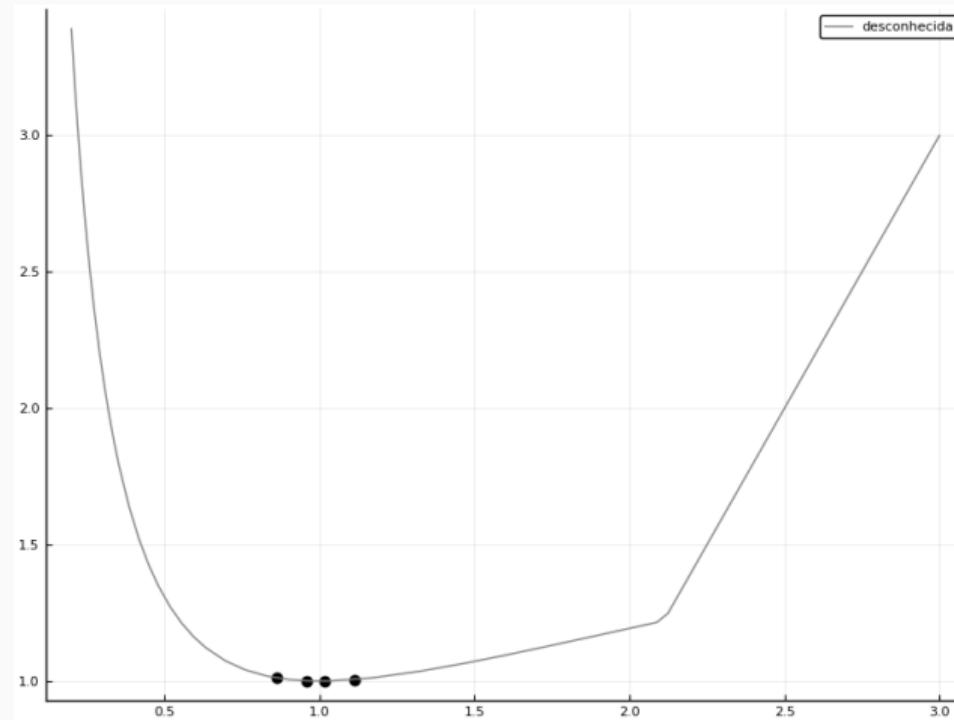
Método da seção áurea



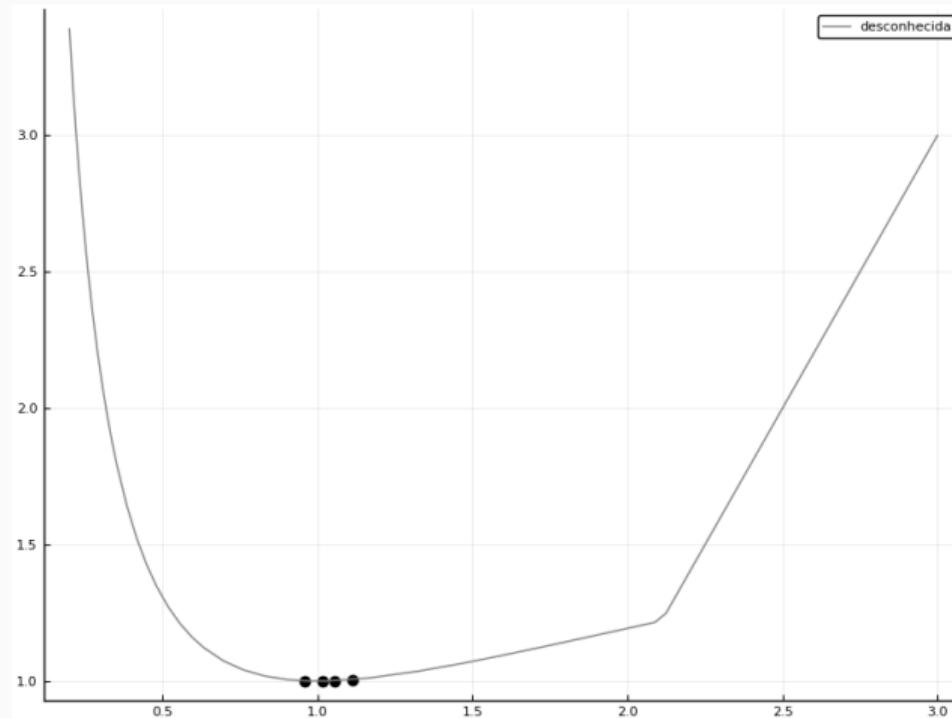
Método da seção áurea



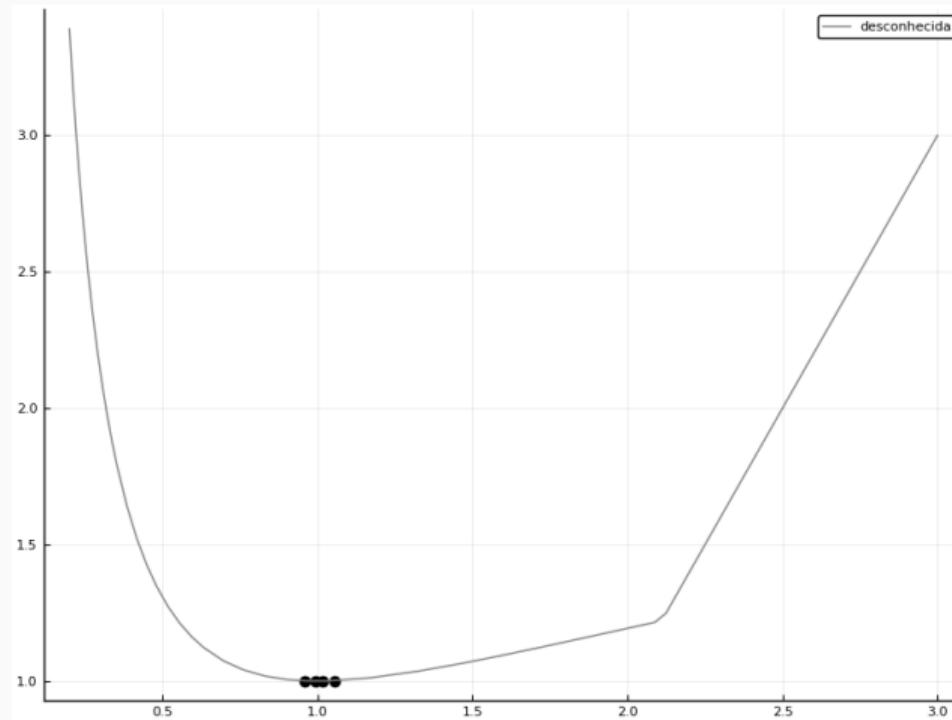
Método da seção áurea



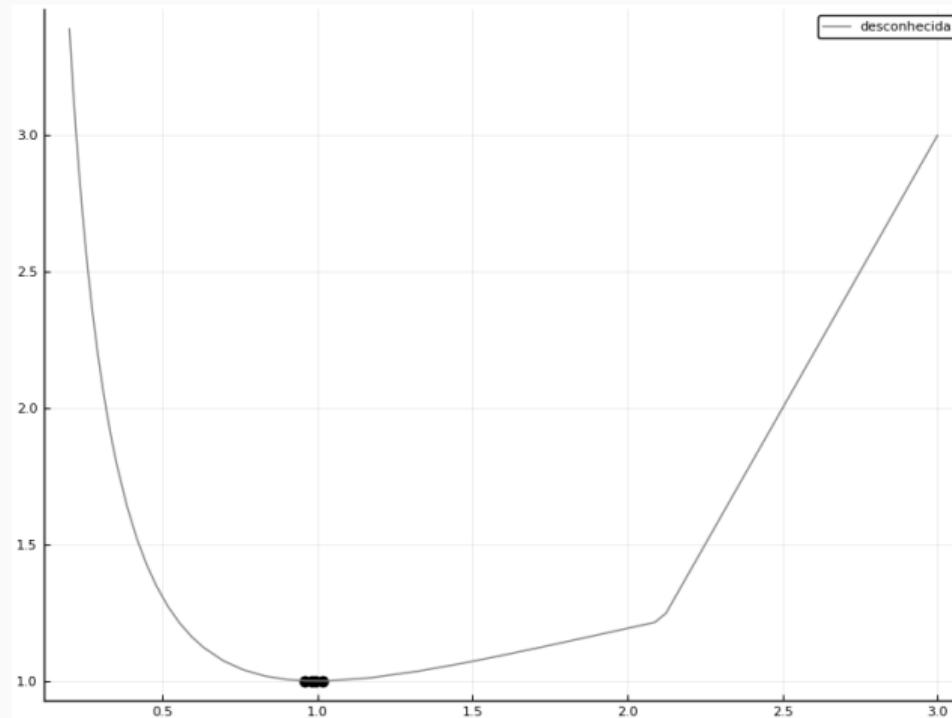
Método da seção áurea



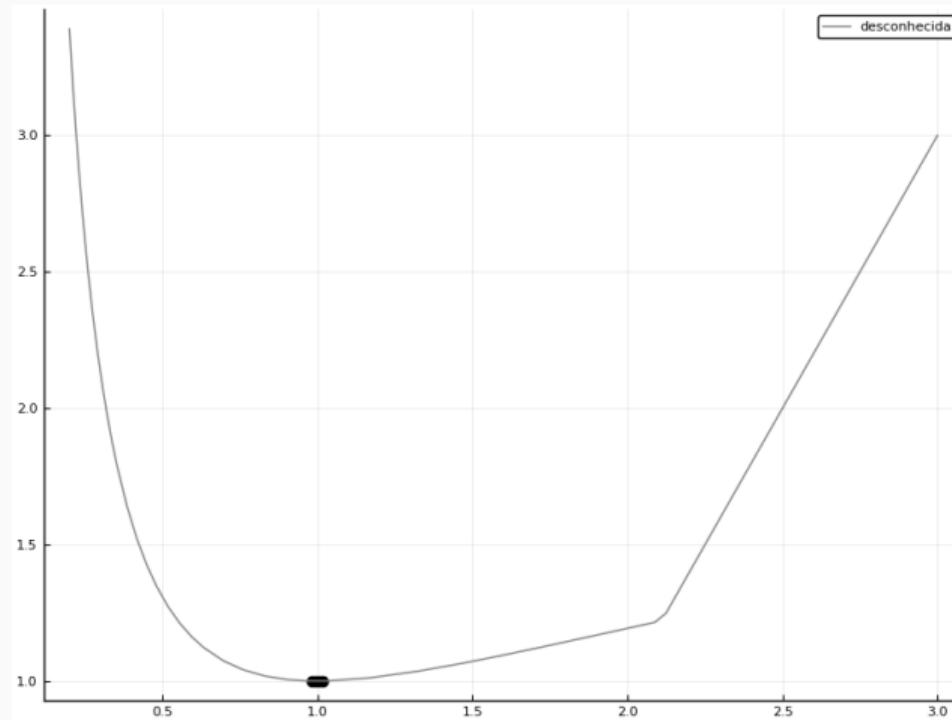
Método da seção áurea



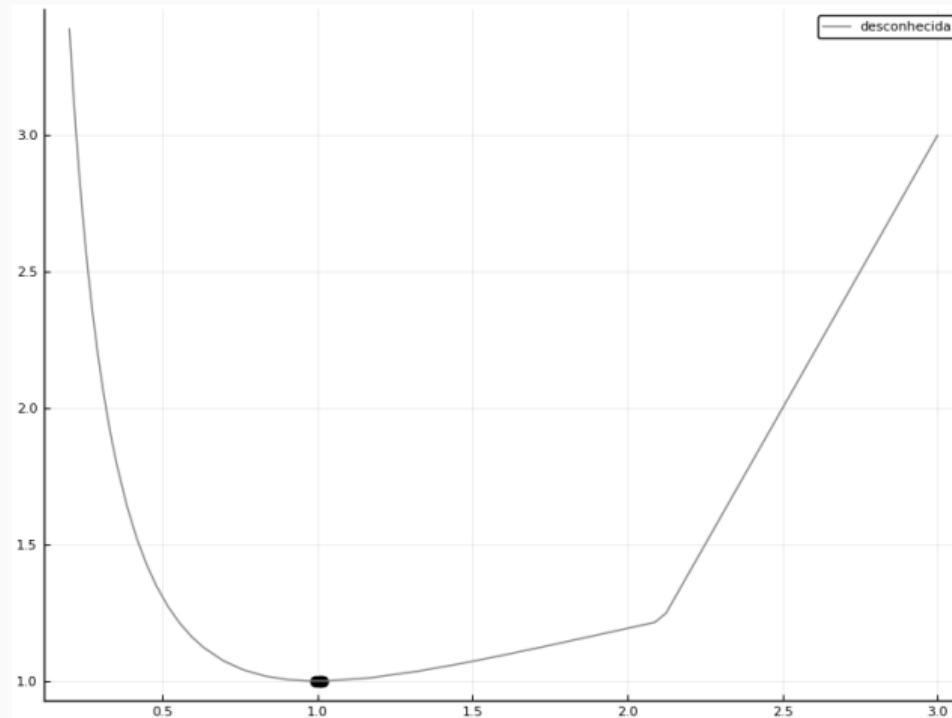
Método da seção áurea



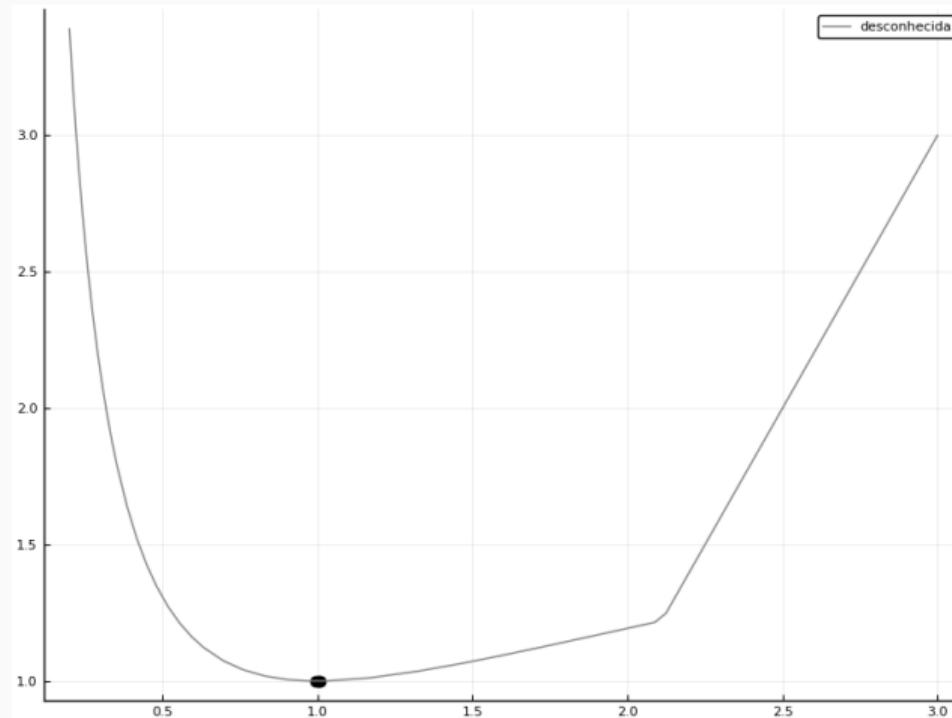
Método da seção áurea



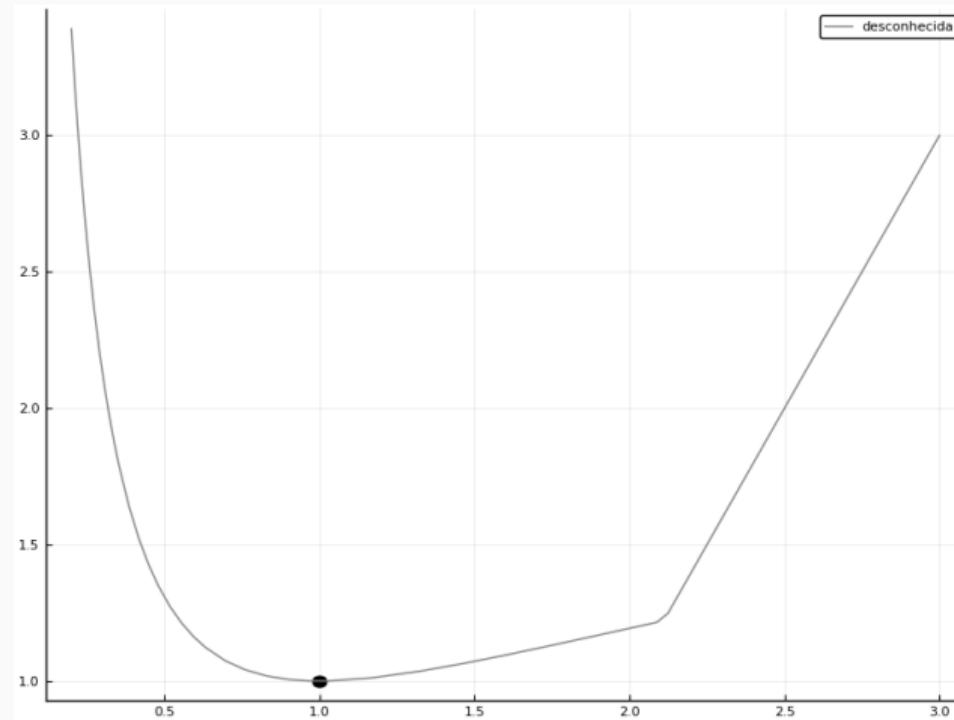
Método da seção áurea



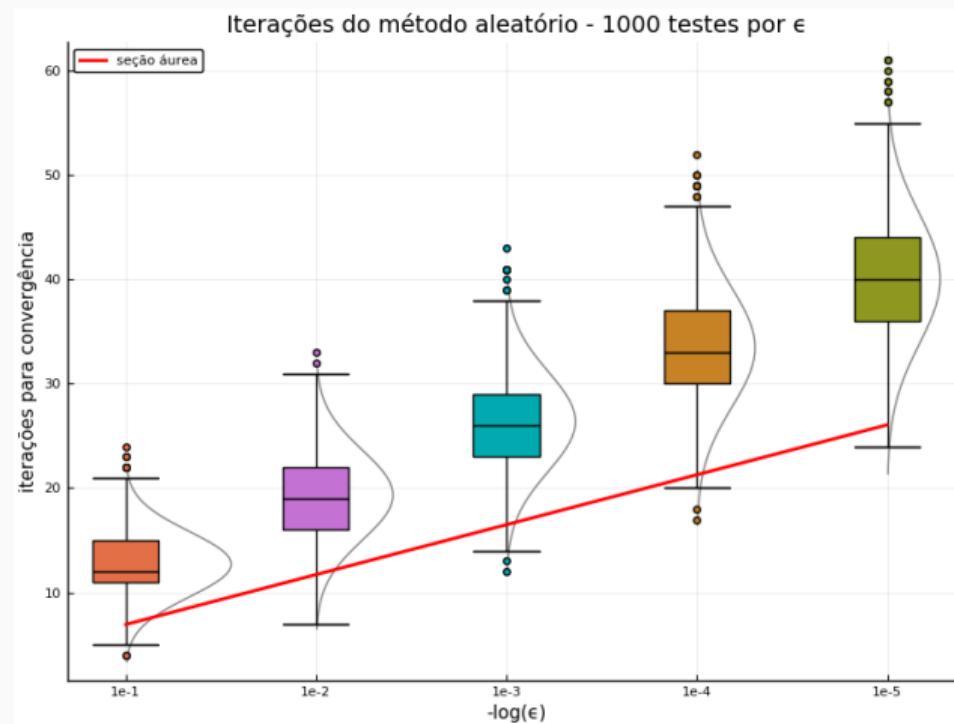
Método da seção áurea



Método da seção áurea



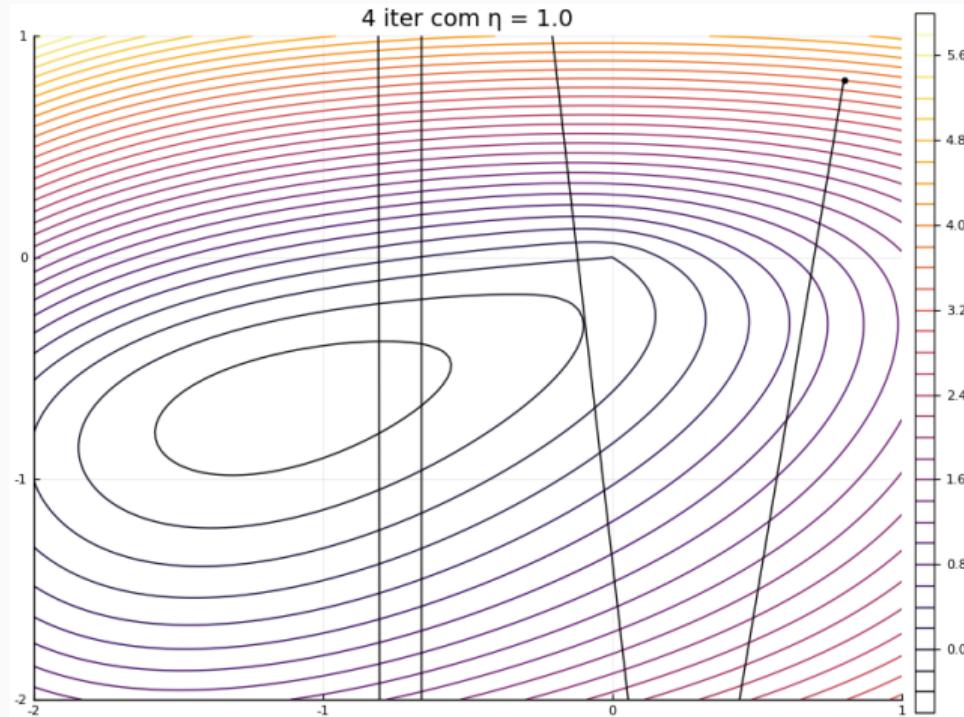
Método da seção áurea



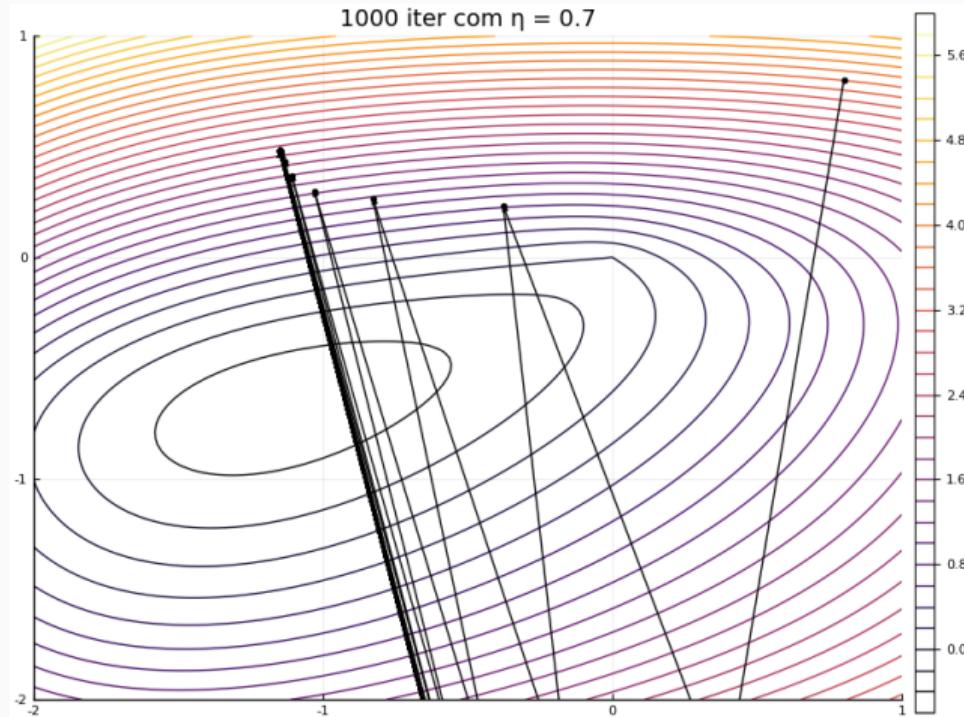
Método do gradiente com passo constante

- Se f é convexa, também podemos usar $\alpha_k = \eta$ constante, desde que pequeno o suficiente.
- Esse método tem garantia de convergência para f convexa, por isso é considerado em Machine Learning.
- Esse método dá origem ao gradiente estocástico.
- A escolha do η não é trivial. Na prática, se testa um valor de η , e se não funcionar, diminui.
- Uma variante é o passo decrescente.

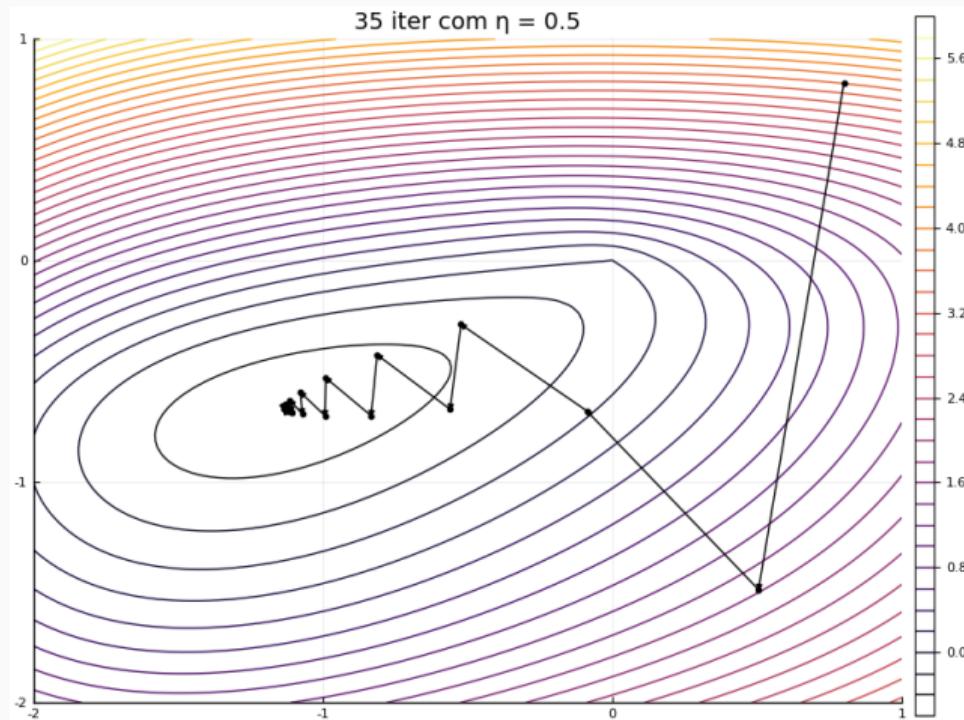
Métodos de gradiente



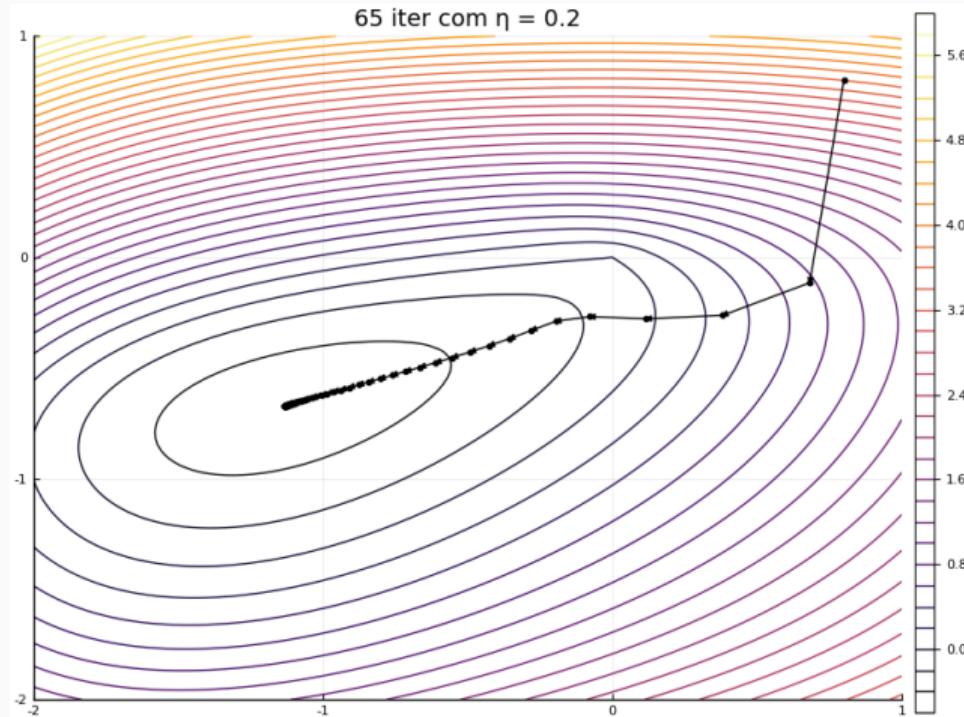
Métodos de gradiente



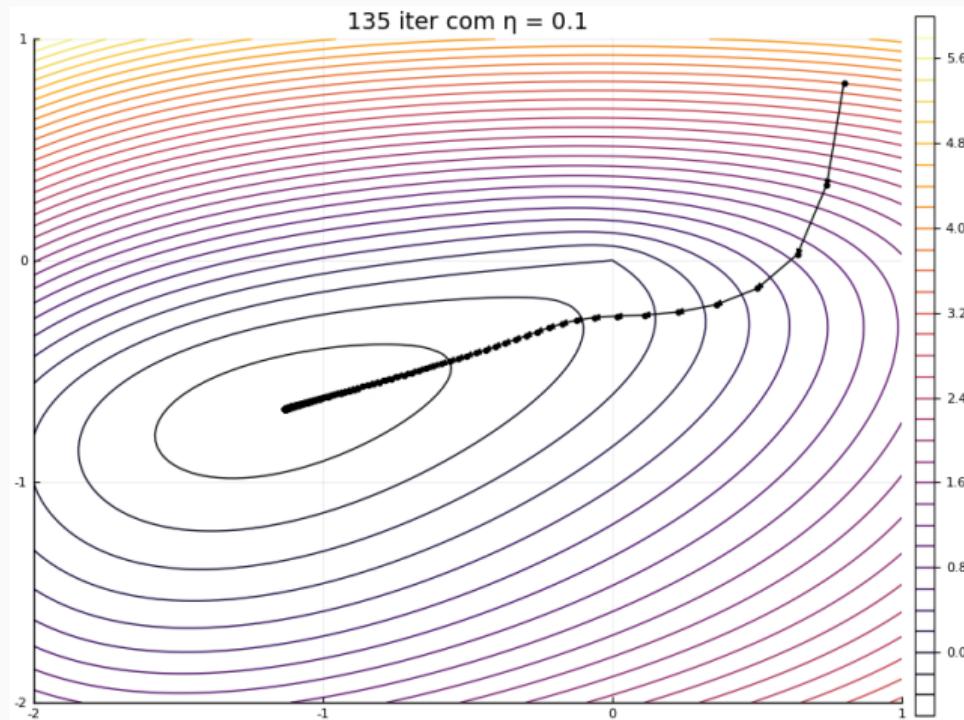
Métodos de gradiente



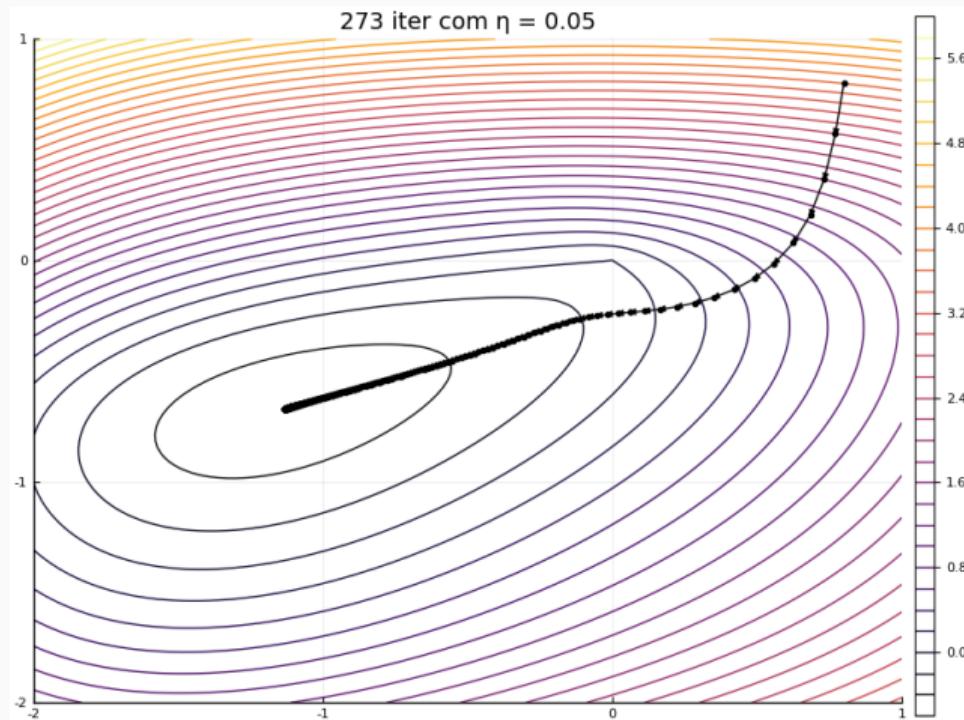
Métodos de gradiente



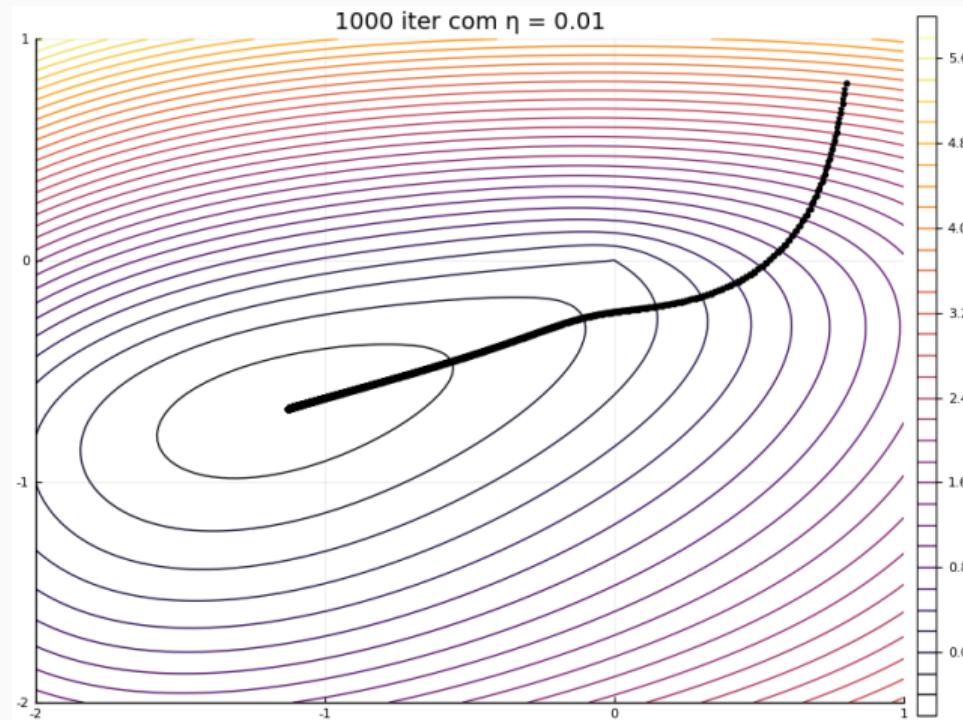
Métodos de gradiente



Métodos de gradiente



Métodos de gradiente



Sumário

- Funções C^2 podem ser aproximadas bem por quadráticas.
- As condições de otimalidade vêm da aproximação quadrática.
- O método de Newton é a minimização sucessiva de aproximações quadráticas de Taylor.
- O método de Newton não tem garantia de convergência longe do minimizador.
- Se perto o suficiente do minimizador, numa região estritamente convexa, a convergência é quadrática.
- Métodos tipo gradiente precisam de controle de passo.
- Um tipo de controle é a busca exata (Cauchy), outro é um passo constante.

FIM
