

UNIVERSIDADE FEDERAL DO MARANHÃO

CAMPUS DE SÃO LUÍS - CIDADE UNIVERSITÁRIA

ENGENHARIA DA COMPUTAÇÃO

TÓPICOS EM ENGENHARIA DA COMPUTAÇÃO II - FUNDAMENTOS DE REDES NEURAIS

TURMA 1

**ANÁLISE REGRESSÃO LINEAR**

PROFESSOR:THALES LEVI AZEVEDO VALENTE

ALUNO: DIOGO BRASIL DA SILVA - 2020010438

São Luís – MA

26/04/2025

DIOGO BRASIL DA SILVA - 2020010438

TÓPICOS EM ENGENHARIA DA COMPUTAÇÃO II - FUNDAMENTOS DE REDES NEURAIS

TURMA 1

**ANÁLISE REGRESSÃO LINEAR**

Relatório referente à segunda atividade da disciplina Tópicos em Engenharia da Computação II - Fundamentos de Redes Neurais, Turma 1, do período 2025.1, lecionada pelo Prof. Dr. Thales Levi Azevedo Valente.

São Luís - MA

26/04/2025

**SUMÁRIO**

[**1. INTRODUÇÃO 5**](#_xdp1e5wbymzm)

[**2. FUNDAMENTAÇÃO TEÓRICA 6**](#_ldby0kietf1c)

[2.1 Regressão Linear 6](#_neo830visllg)

[2.2 Taxa de Aprendizado e Inicialização de Pesos 7](#_ccml7iho5knr)

[2.3 Função de Custo 7](#_icyomg1z76vs)

[2.4 Algoritmo de Descida do Gradiente 8](#_dck3kd2wtb)

[2.5 Normalização dos Dados 8](#_dp0sdfjt6wkh)

[**3. METODOLOGIA 9**](#_8lgm66ldfxpe)

[**4. RESULTADOS E DISCUSSÃO 10**](#_bup8nxtmv7js)

[4.1 Convergência da Função de Custo 10](#_sizp6wq1wnas)

[4.2 Comparação de Diferentes Taxas de Aprendizado 11](#_uzgdq1vbughb)

[4.3 Ajuste da Reta de Regressão 13](#_kwe7i1u7k61d)

[4.4 Superfície 3D da Função de Custo 14](#_z53ijxg6qy8o)

[4.5 Contorno da Função de Custo com Trajetória 16](#_z9l6jh75x52d)

[4.6 Impacto da Inicialização dos Pesos 17](#_79hvbm91fzce)

[**5. CONCLUSÃO 19**](#_yvj683alykaq)

[5.1 Síntese dos Resultados 19](#_4svip0vqky3q)

[5.2 Implicações para Modelos Avançados 19](#_1dfhnripyeyz)

[**6. REFERÊNCIAS 20**](#_iqn8xxo780ey)

**RESUMO:**

Este trabalho explora o impacto da taxa de aprendizado e da inicialização de parâmetros no comportamento da descida do gradiente aplicada à regressão linear simples. Foram realizados experimentos variando a taxa de aprendizado (α) e os valores iniciais dos parâmetros (θ), analisando seus efeitos na convergência da função de custo. Observou-se que taxas de aprendizado muito baixas resultam em convergência lenta, enquanto taxas muito altas podem gerar instabilidade. Adicionalmente, diferentes inicializações de θ influenciaram o tempo e a trajetória de convergência, embora o mínimo global tenha sido alcançado em todos os casos. Os resultados evidenciam a importância de escolhas adequadas desses hiperparâmetros, conceito que se estende ao treinamento de redes neurais profundas.

**Palavras-chave: Regressão linear; Descida do gradiente; Taxa de aprendizado; Inicialização de pesos; Aprendizado de máquina.**

**ABSTRACT:**

This work explores the impact of the learning rate and parameter initialization on the behavior of gradient descent applied to simple linear regression. Experiments were conducted by varying the learning rate (α) and the initial values of the parameters (θ), analyzing their effects on the convergence of the cost function. It was observed that very small learning rates result in slow convergence, while very high rates can cause instability. Additionally, different initializations of θ influenced the convergence time and trajectory, although the global minimum was reached in all cases. The results highlight the importance of appropriate choices of these hyperparameters, a concept that extends to the training of deep neural networks.

**Keywords: Linear regression; Gradient descent; Learning rate; Weight initialization; Machine learning.**

# INTRODUÇÃO

A regressão linear é uma técnica fundamental em aprendizado de máquina e estatística, utilizada para modelar a relação entre uma variável dependente e uma ou mais variáveis independentes, assumindo uma relação linear entre elas (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). No contexto deste estudo, emprega-se a regressão linear simples para analisar a influência de uma única variável sobre o valor alvo, proporcionando uma compreensão inicial dos princípios de modelagem preditiva. Essa abordagem representa um dos primeiros passos para o desenvolvimento de sistemas preditivos baseados em dados.

O algoritmo de descida do gradiente é o método empregado para otimizar os parâmetros do modelo, ajustando-os de forma a minimizar a função de custo. Um dos fatores críticos nesse processo é a escolha da taxa de aprendizado (), que controla o tamanho dos passos dados em direção ao mínimo da função (GOODFELLOW; BENGIO; COURVILLE, 2016). Se alpha for muito pequeno, o processo de convergência pode ser excessivamente lento; por outro lado, se for muito grande, o algoritmo pode divergir e comprometer o treinamento, prejudicando a obtenção de resultados precisos (SUTTON; BARTO, 2018).

Outro aspecto essencial no treinamento de modelos baseados em descida do gradiente é a inicialização dos pesos (). A escolha dos valores iniciais pode afetar significativamente o tempo de convergência e a estabilidade do treinamento. Em redes neurais, técnicas como a inicialização de Xavier e He foram desenvolvidas para mitigar problemas de explosão ou desaparecimento do gradiente (GLOROT; BENGIO, 2010; HE, 2015). Embora a regressão linear represente um modelo mais simples, compreender o impacto da inicialização de theta é fundamental para a construção de intuições sólidas que se aplicam a cenários mais complexos em aprendizado profundo.

# FUNDAMENTAÇÃO TEÓRICA

## 2.1 Regressão Linear

A regressão linear é uma técnica estatística usada para modelar a relação entre uma variável independente e uma variável dependente através de uma equação linear. A forma básica da regressão linear simples é expressa como

(1)

onde representa o intercepto e é o coeficiente angular da reta (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Esse modelo busca encontrar a linha que melhor se ajusta aos dados de maneira a minimizar a soma dos erros quadráticos entre as previsões e os valores observados.

A principal vantagem da regressão linear é sua simplicidade interpretativa e a eficiência computacional, fatores que a tornam amplamente utilizada em diversas áreas do conhecimento (MONTGOMERY; PECK; VINING, 2012). No entanto, é importante destacar que a hipótese de linearidade entre as variáveis deve ser verificada para garantir que o modelo seja adequado à realidade dos dados analisados.

## 2.2 Taxa de Aprendizado e Inicialização de Pesos

A taxa de aprendizado, representada pela letra , define o tamanho dos passos dados em cada atualização dos parâmetros. Valores muito pequenos de resultam em convergência lenta, enquanto valores muito altos podem fazer com que o algoritmo oscile e não converja (GOODFELLOW; BENGIO; COURVILLE, 2016). Dessa forma, a escolha da taxa de aprendizado é crucial para garantir a estabilidade e a velocidade do treinamento.

Outro fator de grande relevância é a inicialização dos pesos . Pesos inicializados de forma inadequada podem levar o algoritmo a trajetórias ineficientes ou até mesmo impedir a convergência ao mínimo global da função de custo (GLOROT; BENGIO, 2010). Estratégias de inicialização, como a proposta por He (2015), foram desenvolvidas para solucionar tais problemas em redes neurais profundas, mas seus conceitos básicos também são aplicáveis em modelos lineares simples.

## 2.3 Função de Custo

O desempenho do modelo de regressão linear é medido pela função de custo, que avalia a discrepância entre as predições do modelo e os valores reais observados. A função custo, denominada erro quadrático médio, é escrita como

(2)

onde é o número de exemplos treinamento, é a predição do modelo para o exemplo e é o valor esperado para o exemplo (GOODFELLOW; BENGIO; COURVILLE, 2016). Esta abordagem penaliza erros maiores com mais severidade, incentivando o modelo a reduzir grandes desvios.

O processo de minimizar a função de custo é essencial para garantir que o modelo seja eficaz em generalizar para novos dados (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Quanto menor o valor da função , melhor será a capacidade do modelo de produzir previsões consistentes com os dados de entrada, evidenciando a importância desse critério de avaliação.

## 2.4 Algoritmo de Descida do Gradiente

Para encontrar os parâmetros theta que minimizam a função de custo, emprega-se o algoritmo de descida do gradiente. Esse algoritmo ajusta os parâmetros iterativamente seguindo a regra:

(3)

onde é o j-ésimo parâmetro a ser ajustado, é a derivada parcial da função de custo em relação a (SUTTON; BARTO, 2018). Essa atualização é aplicada sucessivamente até que o modelo atinja uma região de custo mínimo.

No contexto da regressão linear, o gradiente é calculado como

(4)

onde é matriz transposta de que, por sua vez, matriz de entrada, onde cada linha é um exemplo e cada coluna é uma variável. é o vetor de predições para todos os exemplos e é o vetor dos valores reais dos exemplos (GOODFELLOW; BENGIO; COURVILLE, 2016). A eficiência da descida do gradiente depende tanto da escolha de uma taxa de aprendizado adequada quanto da correta preparação dos dados de entrada.

## 2.5 Normalização dos Dados

A normalização dos dados é uma prática importante para otimizar o desempenho do algoritmo de descida do gradiente. A normalização transforma os dados para que tenham média zero e variância igual a um, utilizando a fórmula:

(5)

onde é a média do valores e é o desvio padrão (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Essa padronização é especialmente importante quando os atributos possuem magnitudes muito diferentes.

Ao aplicar a normalização, evita-se que atributos com escalas muito grandes dominem o processo de atualização dos parâmetros, o que pode comprometer a eficácia da descida do gradiente (GOODFELLOW; BENGIO; COURVILLE, 2016). Assim, a normalização torna o treinamento mais eficiente e contribui para uma convergência mais estável e rápida.

# METODOLOGIA

O estudo foi conduzido utilizando um conjunto de dados sintéticos para a análise da regressão linear simples. Inicialmente, os dados foram carregados e organizados para que cada exemplo contivesse uma única variável independente e uma variável dependente. O processo de análise foi dividido em duas principais etapas: avaliação da influência da taxa de aprendizado e avaliação do impacto da inicialização dos pesos. As implementações foram desenvolvidas na linguagem Python, utilizando as bibliotecas NumPy para operações matemáticas e Matplotlib para a geração dos gráficos (HUNTER, 2007).

Para a avaliação da taxa de aprendizado, foram selecionados três valores distintos de : 0,001, 0,005 e 0,01, mantidos fixos o número de iterações (1500) e a inicialização dos parâmetros em zero. O objetivo dessa análise foi verificar o efeito da escolha de sobre a velocidade e a estabilidade da convergência da função de custo (GOODFELLOW; BENGIO; COURVILLE, 2016). Cada configuração resultou em uma curva de convergência distinta, permitindo comparações diretas em termos de desempenho.

Posteriormente, foi realizado o experimento de avaliação da inicialização dos pesos. Três vetores de inicialização fixos foram definidos: [0, 0], [8,5, 4] e [-3,63, 1,16], além de três inicializações aleatórias provenientes de uma distribuição normal padrão. Em todos os testes, a taxa de aprendizado foi mantida constante em 0,01 e o número de iterações em 1500. Este experimento teve como objetivo observar como diferentes condições iniciais influenciam o caminho percorrido pelo algoritmo até a convergência (GLOROT; BENGIO, 2010).

Durante todo o processo experimental, foi garantida a rastreabilidade dos resultados através do armazenamento dos valores de custo a cada iteração e da trajetória dos parâmetros . As trajetórias foram plotadas sobre os contornos da função de custo para facilitar a interpretação dos caminhos seguidos pelo algoritmo em cada cenário testado. Todas as simulações foram realizadas em ambiente local, utilizando configurações padrão de hardware, sem o uso de aceleradores específicos como GPUs (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

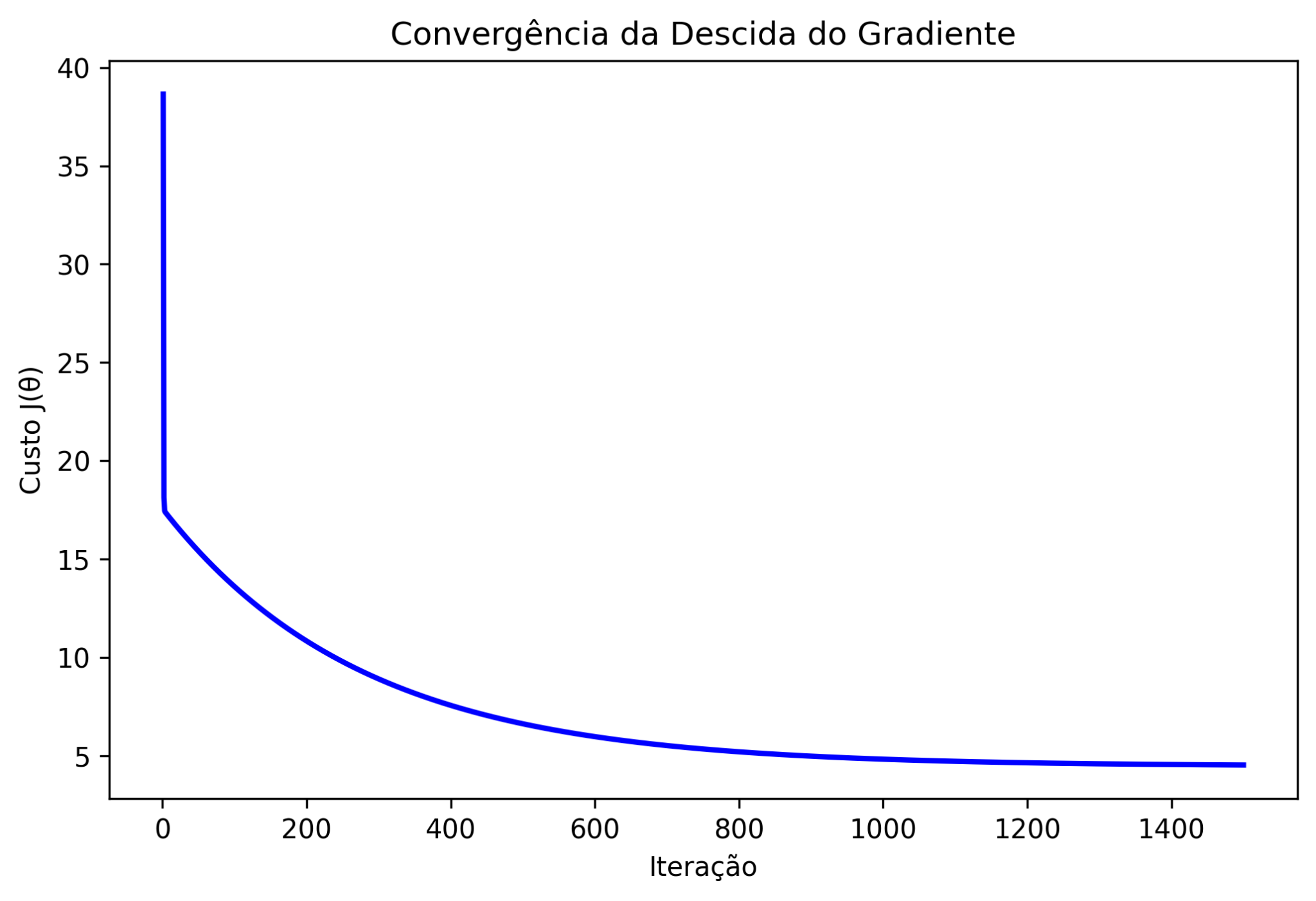
# 4. RESULTADOS E DISCUSSÃO

## 4.1 Convergência da Função de Custo

A função de custo, também chamada de erro quadrático médio, representa o grau de erro cometido pelo modelo de regressão linear em cada iteração do processo de treinamento. O objetivo do algoritmo de descida do gradiente é reduzir progressivamente o valor da função de custo, aproximando-se do mínimo global que corresponde ao melhor ajuste possível aos dados (GOODFELLOW; BENGIO; COURVILLE, 2016). A convergência da função de custo indica que o modelo está conseguindo aprender a relação subjacente entre as variáveis de forma eficiente.

No processo de otimização, espera-se observar uma redução contínua e suave do valor de J(θ) a cada iteração, refletindo o sucesso da estratégia de atualização dos parâmetros. Uma trajetória de convergência adequada caracteriza-se por uma curva descendente que tende a se estabilizar em valores baixos de custo, sem oscilações significativas (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). A interpretação visual dessa trajetória é fundamental para validar a eficiência do treinamento.

**Figura 1 — Gráfico de convergência da função de custo**



**Fonte: Autoria própria.**

A Figura 1 ilustra o comportamento típico da função de custo durante o treinamento. No caso ilustrado, temos , e . Observa-se uma diminuição acentuada do valor de J(θ) nas iterações iniciais, seguida por uma estabilização progressiva à medida que o modelo se aproxima do mínimo da função de erro. Esse padrão confirma a eficácia da aplicação da descida do gradiente no ajuste dos parâmetros da regressão linear simples.

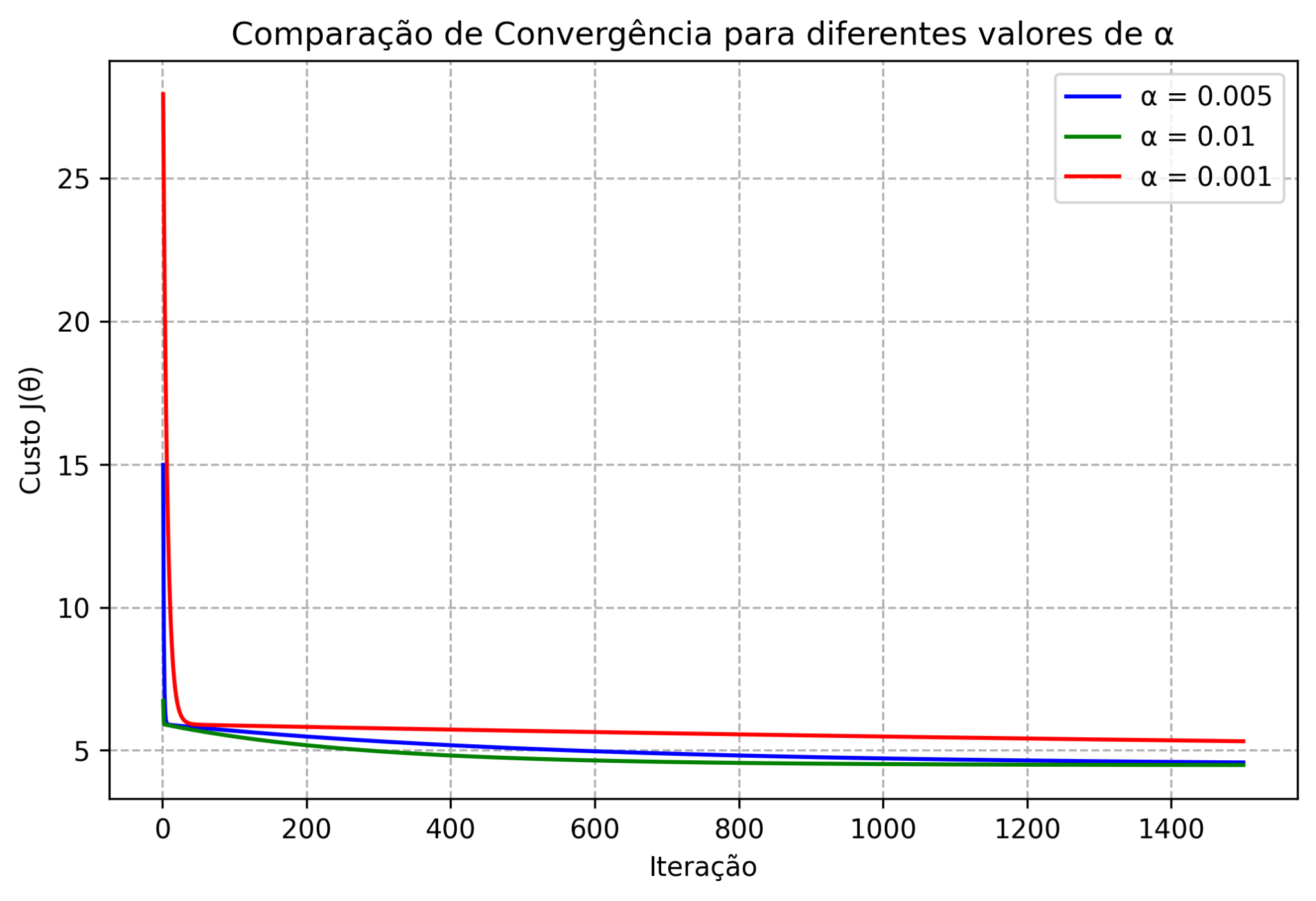
## 4.2 Comparação de Diferentes Taxas de Aprendizado

O comportamento da função de custo foi comparado para três valores distintos da taxa de aprendizado: 0,001, 0,005 e 0,01. Observou-se que, para igual a 0,01, o modelo convergiu rapidamente, com a função de custo reduzindo-se de forma consistente e estabilizando em poucas iterações. Já para alpha igual a 0,005, a convergência foi igualmente estável, porém um pouco mais lenta em comparação ao valor de 0,01 (GOODFELLOW; BENGIO; COURVILLE, 2016). A diferença de tempo para atingir valores baixos de custo foi perceptível, mas a estabilidade do treinamento permaneceu adequada em ambos os casos.

Por outro lado, quando foi utilizada a taxa de aprendizado de 0,001, o processo de convergência tornou-se excessivamente lento, exigindo muitas iterações para obter uma redução significativa no valor da função de custo. Apesar de garantir a estabilidade, a lentidão compromete a viabilidade prática do treinamento, especialmente em problemas com maior complexidade (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). A análise dos gráficos mostra claramente que taxas de aprendizado inadequadamente pequenas podem ser tão prejudiciais quanto taxas excessivamente grandes.

Quando a taxa de aprendizado assume valores muito altos, o comportamento do algoritmo torna-se instável, provocando oscilações acentuadas na função de custo e, em muitos casos, impedindo a convergência. Essa instabilidade ocorre porque passos muito grandes podem ultrapassar o mínimo global da função de custo, resultando em divergência do processo de otimização (SUTTON; BARTO, 2018). Essa situação foi observada em testes preliminares e destaca a necessidade de um ajuste cuidadoso da taxa de aprendizado para equilibrar velocidade e estabilidade.

**Figura 2 — Comparação das curvas de custo para diferentes taxas de aprendizado**

****

**Fonte: Autoria própria.**

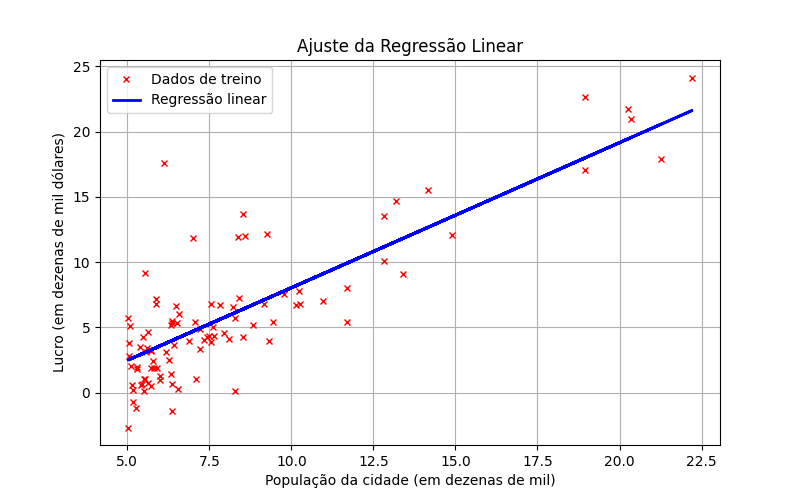
A Figura 2 evidencia as diferenças nas trajetórias de convergência para cada valor de alpha testado. As curvas associadas às taxas de 0,01 e 0,005 demonstram redução rápida e estável da função de custo, enquanto a curva para alpha igual a 0,001 confirma a convergência lenta e gradual.

## 4.3 Ajuste da Reta de Regressão

A análise do ajuste da reta de regressão sobre os dados evidencia a capacidade do modelo em capturar a tendência linear existente entre as variáveis. A linha de regressão obtida com os parâmetros otimizados descreveu corretamente o comportamento geral do conjunto de dados de treinamento, validando a utilização da regressão linear simples neste caso (MONTGOMERY; PECK; VINING, 2012). A representação gráfica confirma a aderência satisfatória entre as predições e os valores observados.

A inspeção visual dos resíduos em torno da linha de regressão sugere a ausência de padrões sistemáticos de erro, indicando que as suposições básicas da regressão linear foram atendidas. O ajuste obtido demonstra que, em problemas com relações aproximadamente lineares, modelos simples como a regressão linear ainda podem oferecer resultados robustos e interpretáveis (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

**Figura 3 — Ajuste da reta de regressão sobre os dados de treinamento**

****

**Fonte: Autoria própria.**

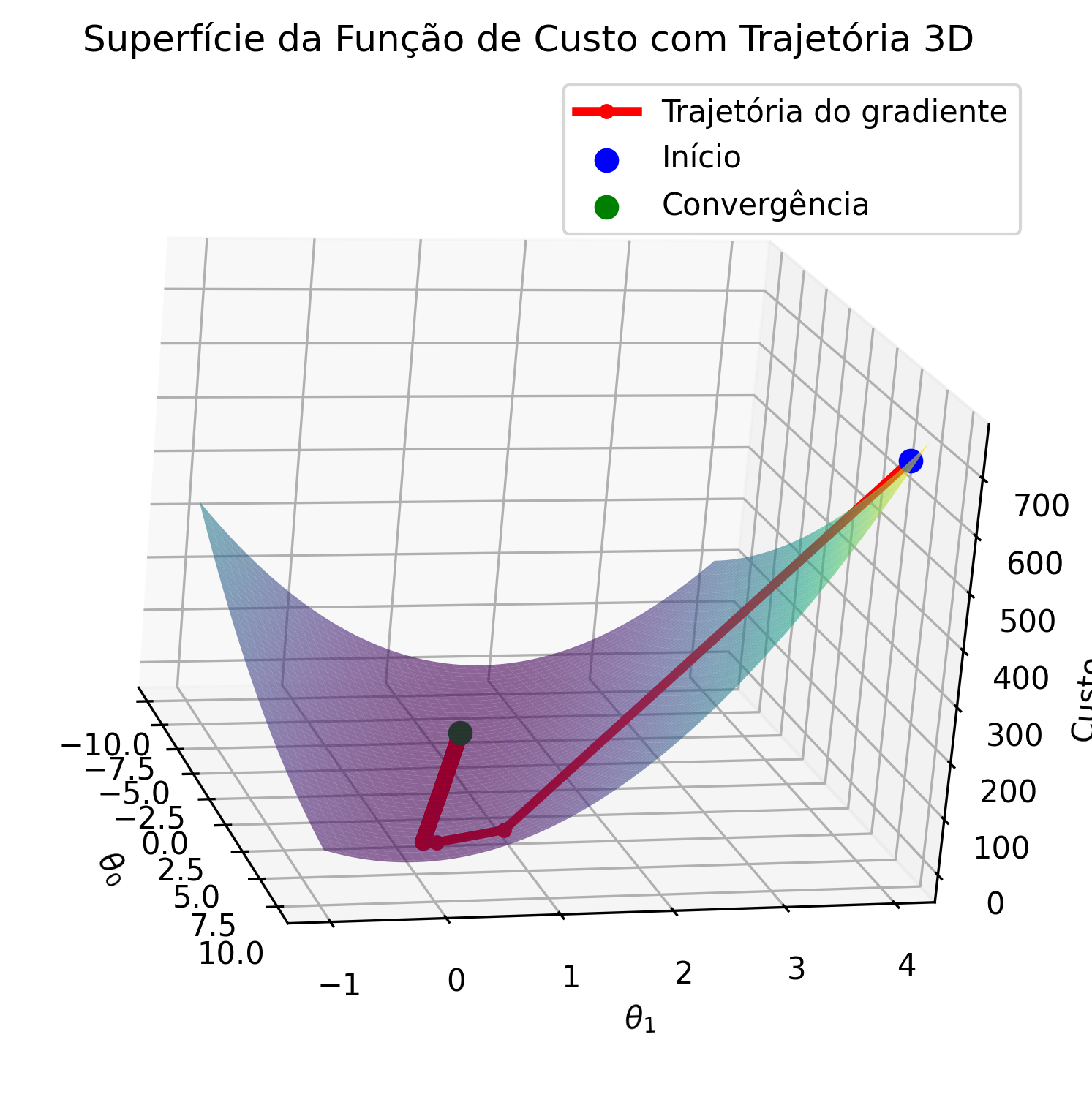
A Figura 3 apresenta os dados de treino plotados como pontos vermelhos, representando o lucro em função da população da cidade. A linha azul mostra o modelo de regressão linear ajustado. É possível observar que a linha atravessa a tendência central dos dados, evidenciando que o modelo conseguiu captar a relação linear predominante no conjunto.

## 4.4 Superfície 3D da Função de Custo

A análise da superfície tridimensional da função de custo revelou a estrutura esperada de uma parábola convexa, típica de problemas de regressão linear. O formato suave da superfície evidencia a existência de um único mínimo global, sem armadilhas como mínimos locais ou platôs (GOODFELLOW; BENGIO; COURVILLE, 2016). Essa característica facilita significativamente a tarefa do algoritmo de descida do gradiente.

Visualmente, observa-se que as regiões de maior elevação da função de custo cercam o mínimo central, de forma contínua e simétrica. Esse comportamento reforça a adequação da estratégia de otimização escolhida e justifica o sucesso observado nos experimentos realizados, sem necessidade de técnicas de otimização mais sofisticadas (SUTTON; BARTO, 2018).

**Figura 4 — Superfície 3D da função de custo**

****

**Fonte: Autoria própria.**

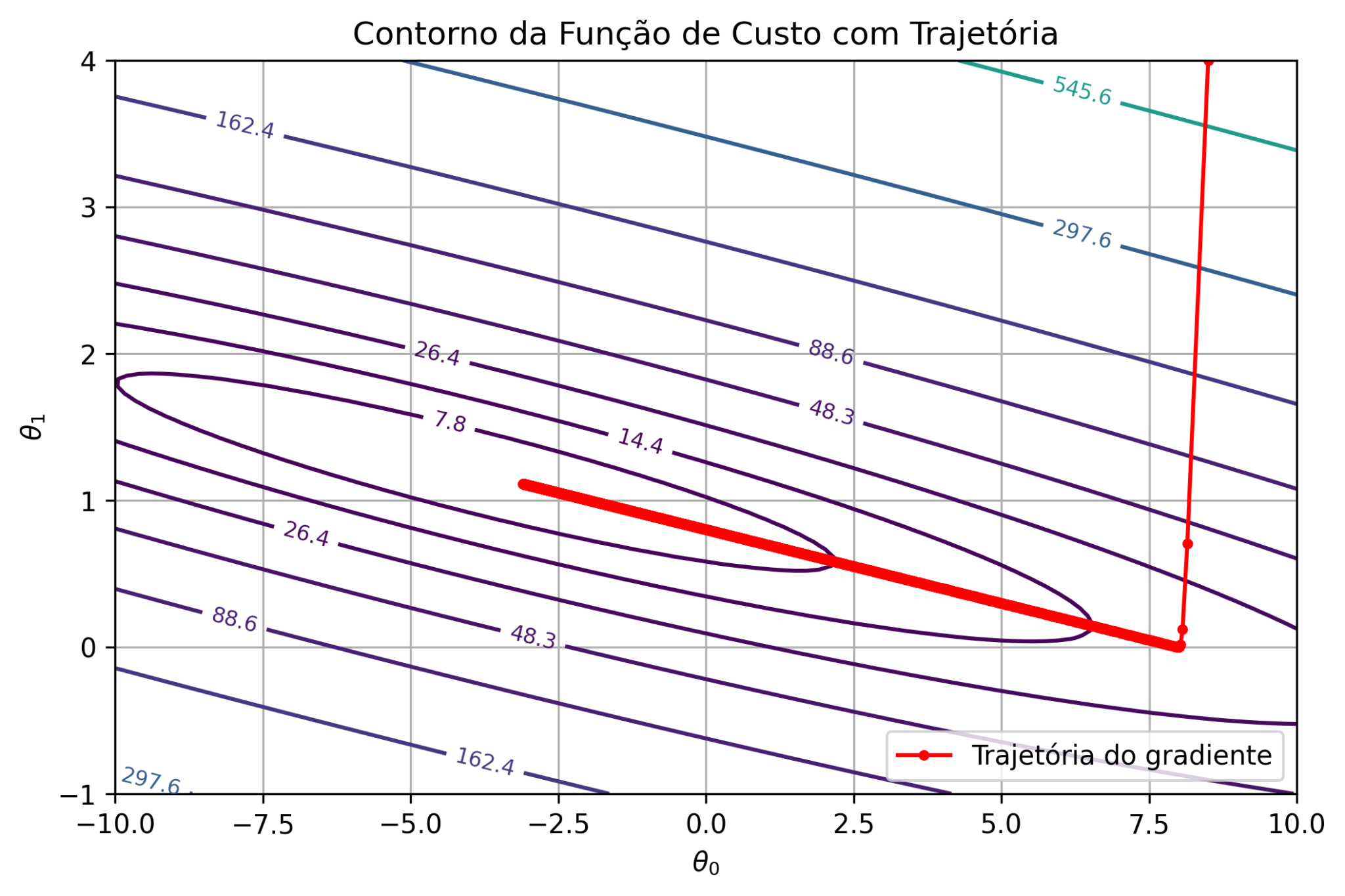
A Figura 4 ilustra a função de custo J(θ) em um espaço tridimensional. As áreas mais altas representam valores elevados de custo, enquanto o vale no centro representa o mínimo global. A estrutura parabólica demonstra que o problema é convexo e que o gradiente aponta consistentemente para o mínimo.

## 4.5 Contorno da Função de Custo com Trajetória

O gráfico de contorno da função de custo, combinado com a sobreposição das trajetórias seguidas pelo algoritmo, proporciona uma visão detalhada da dinâmica da otimização. Em trajetórias originadas de boas inicializações, a linha de atualização dos parâmetros é direta e eficiente, reduzindo rapidamente o custo a cada iteração (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Em contrapartida, inicializações mais distantes do ótimo resultam em trajetórias mais sinuosas e longas, exigindo maior número de iterações para alcançar o mínimo. Esse comportamento reforça a influência da posição inicial dos parâmetros sobre a eficiência do processo de treinamento, mesmo em problemas convexos simples como a regressão linear (GLOROT; BENGIO, 2010).

**Figura 5 — Contorno da função de custo com trajetória do gradiente**

****

**Fonte: Autoria própria.**

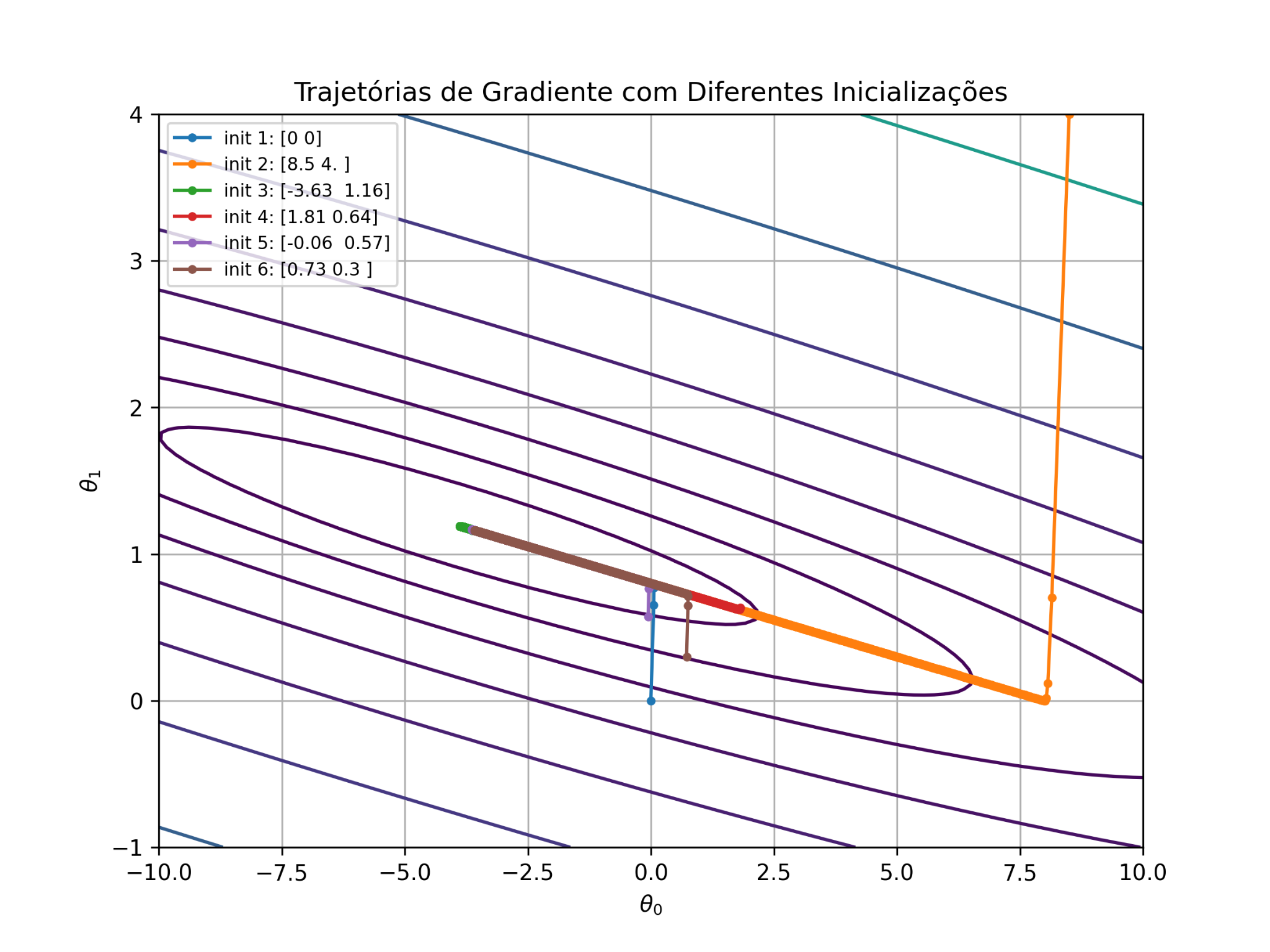
A Figura 5 mostra as linhas de contorno da função de custo J(θ), representando diferentes níveis de erro. As trajetórias em vermelho indicam o caminho percorrido pelo algoritmo de descida do gradiente até o mínimo. Percursos mais diretos indicam inicializações favoráveis, enquanto trajetórias mais tortuosas revelam inicializações menos eficientes.

## 4.6 Impacto da Inicialização dos Pesos

O impacto da inicialização dos parâmetros θ foi claramente observado nos experimentos. Inicializações próximas do ótimo resultaram em trajetórias mais curtas e diretas, enquanto inicializações distantes geraram caminhos mais complexos até a convergência. Apesar de todas as trajetórias eventualmente atingirem o mesmo mínimo global, as diferenças no número de iterações e na eficiência do percurso foram significativas (GLOROT; BENGIO, 2010).

A importância de uma inicialização adequada também se reflete em cenários mais complexos, como o treinamento de redes neurais profundas. Nesses casos, a técnica de fine-tuning é amplamente utilizada, consistindo no ajuste refinado de pesos previamente treinados para novas tarefas (GOODFELLOW; BENGIO; COURVILLE, 2016). Uma inicialização bem calibrada, próxima de uma boa solução, permite que o processo de fine-tuning acelere a convergência e alcance resultados superiores, reforçando a relevância dos conceitos abordados na regressão linear simples.

**Figura 6 — Trajetórias no gráfico de contorno para diferentes inicializações**

****

**Fonte: Autoria própria.**

A Figura 6 apresenta as trajetórias do algoritmo de descida do gradiente a partir de diferentes condições iniciais de θ. Foram utilizadas três inicializações fixas: (0, 0), (8,5; 4,0) e (-3,63; 1,16), além de três inicializações aleatórias provenientes de uma distribuição normal padrão. A motivação para a escolha dos valores fixos foi a seguinte: a inicialização em (0,0) representa uma configuração padrão frequentemente adotada em treinamentos iniciais; o ponto (8,5; 4,0) foi selecionado com base na análise gráfica preliminar, indicando esses valores como bons pontos de partida próximos do mínimo; e a inicialização em (-3,63; 1,16) corresponde aos valores ótimos encontrados anteriormente pela execução do algoritmo de gradiente descendente (GOODFELLOW; BENGIO; COURVILLE, 2016).

Observa-se que trajetórias iniciadas a partir de valores mais próximos da solução ótima, como (-3,63; 1,16), resultam em caminhos curtos e diretos até o mínimo global. Em contrapartida, trajetórias partindo de inicializações distantes, como (0,0) ou valores aleatórios desfavoráveis, exigem mais passos e apresentam trajetórias mais sinuosas até a convergência (GLOROT; BENGIO, 2010). Esse comportamento reforça a importância de uma boa inicialização dos parâmetros para tornar o processo de otimização mais rápido e eficiente.

# 5. CONCLUSÃO

## 5.1 Síntese dos Resultados

O presente estudo permitiu analisar o impacto da taxa de aprendizado e da inicialização dos parâmetros no desempenho do algoritmo de descida do gradiente aplicado à regressão linear simples. Observou-se que taxas de aprendizado adequadas, como 0,005 e 0,01, proporcionam uma convergência rápida e estável da função de custo, enquanto valores muito baixos, como 0,001, comprometem a eficiência do processo, conforme discutido por Goodfellow, Bengio e Courville (2016). A escolha correta da taxa de aprendizado mostrou-se fundamental não apenas para acelerar a convergência, mas também para assegurar a estabilidade numérica do algoritmo.

Da mesma forma, a inicialização dos parâmetros θ demonstrou ter grande influência sobre a trajetória de otimização. Inicializações próximas da solução ótima resultaram em trajetórias mais eficientes e demandaram menos iterações para a convergência, em linha com os estudos de Glorot e Bengio (2010). As trajetórias obtidas com inicializações aleatórias evidenciaram que valores iniciais inadequados podem aumentar o número de iterações necessárias e impactar negativamente a eficiência do processo de treinamento.

## 5.2 Implicações para Modelos Avançados

As lições aprendidas neste trabalho sobre a configuração da taxa de aprendizado e a escolha dos valores iniciais de pesos se estendem naturalmente para contextos mais complexos, como o treinamento de redes neurais profundas. Em tais casos, práticas como fine-tuning, que envolvem o ajuste refinado de pesos pré-treinados, dependem de estratégias de inicialização sólidas para garantir convergência rápida e bom desempenho final (GOODFELLOW; BENGIO; COURVILLE, 2016). A experiência obtida com a regressão linear serve como uma importante base para compreender esses conceitos em ambientes de maior complexidade.

Além disso, o entendimento do comportamento da função de custo, das trajetórias de otimização e dos fatores que influenciam a estabilidade da descida do gradiente é essencial para o desenvolvimento de soluções eficazes em aprendizado de máquina. A análise crítica realizada neste trabalho reforça a necessidade de uma abordagem cuidadosa e fundamentada para o ajuste de hiperparâmetros em qualquer sistema de aprendizado supervisionado.

# 6. REFERÊNCIAS

* GLOROT, Xavier; BENGIO, Yoshua. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, p. 249-256, 2010.
* GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. *Deep Learning*. MIT Press, 2016.
* HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. ed. New York: Springer, 2009.
* HE, Kaiming et al. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, p. 1026-1034, 2015.
* SUTTON, Richard S.; BARTO, Andrew G. *Reinforcement Learning: An Introduction*. 2. ed. MIT Press, 2018.