

Universidade de Brasília
Departamento de Estatística

Gradiente Descendente Estocástico para otimização não convexa

Leonardo Santos da Cruz

Trabalho de conclusão de curso apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2026

Leonardo Santos da Cruz

Gradiente Descendente Estocástico para otimização não convexa

Orientador(a): Prof. Dr. Felipe Quintino
Coorientador(a): Prof. Dr. Helton Saulo Bezerra dos Santos

Trabalho de conclusão de curso apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2026

Gloria In Excelsis Deo

Agradecimentos

Gostaria de agradecer a Deus pela oportunidade de desenvolver este trabalho. Agradeço aos meus orientadores, Prof. Dr. Felipe Quintino e Prof. Dr. Helton Saulo, por me proporcionarem a oportunidade de aprender com eles. Também agradeço pela paciência e pelo suporte durante o desenvolvimento deste trabalho. Agradeço aos meus amigos e familiares pela paciência e compreensão em relação à minha ausência e falta de contato. Também agradeço ao Departamento de Estatística da Universidade de Brasília por me proporcionar a oportunidade de participar do programa graduação. Agradeço a todos os professores pelo aprendizado e pela paciência com que sempre responderam às minhas dúvidas; aos funcionários, por manterem o ambiente sempre agradável para a realização das atividades necessárias; e à equipe da secretaria, por estarem sempre disponíveis e serem pacientes com nossas demandas. Gostaria de agradecer aos Professores Dr. Donald Matthew Pianto e ao professor Dr. Luan Carlos de Sena Monteiro Ozelim pelas oportunas observações e recomendações para o desenvolvimento do trabalho e por terem aceitado participar da banca de avaliação. Agradeço ao Departamento de Matemática da Universidade de Brasília por me proporcionar uma sólida base matemática. Por fim, agradeço à CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo fomento que viabilizou o desenvolvimento deste trabalho. Obrigado a todos!

Resumo

O presente trabalho tem como objetivo apresentar, discutir e aplicar o algoritmo do gradiente descendente estocástico (GDE), com ênfase em problemas de otimização em contextos estatísticos, tais como a maximização da função de verossimilhança e a regressão logística. Inicialmente, são abordados os conceitos fundamentais necessários para a plena compreensão do tema. Em seguida, apresentam-se exemplos da aplicação do GDE na estimação dos parâmetros da regressão logística. Por fim, o algoritmo é empregado na estimação dos parâmetros de um modelo de mistura de distribuições normais. Este trabalho baseou-se em um trabalho prévio de iniciação cintífica premiado no 31º Congresso de Iniciação Científica da UnB e 22º do Distrito Federal (Exatas e Tecnológicas), Universidade de Brasília.

Palavras-chaves: Otimização, gradiente, mistura, estimação.

Lista de Tabelas

1 Iterações do método do subgradiente com $x_0 = 5$ e $\eta = 0.5$ 25

2 Estatísticas descritivas do conjunto de dados 30

3 Resultado do teste de Hartigan's Dip 32

4 Parâmetros estimados do modelo via GDE 35

5 Métricas estimadas do modelo via GDE 36

Lista de Figuras

1 $f(x, y) = x^2 + y^2$ 12

2 $f(x, y) = \frac{5}{4}(x + 6)^2 + (y - 8)^2$ 19

3 Função convexa $f(x) = \max(0, x)$ e um subgradiente no ponto não diferenciável $w = (0, 0)$. 21

4 Distribuições conjunta e marginais 32

5 Gráfico de ajuste do modelo 37

6 Ajuste dos resíduos dos modelos 37

Lista de Siglas e Notações

CV	Coeficiente de variação
DP	Desvio padrão
ERM	Empirical Risk Minimization
GD	Gradiente descendente
GDE	Gradiente descendente estocástico
MLE	Maximum Likelihood Estimator
RMSE	Root Mean Squared Error
SGD	Stochastic Gradient Descent
$\boldsymbol{\nabla} g(\mathbf{x})$	Vetor gradiente da função g em $\mathbf{x}$
$\partial f(\mathbf{w})$	Conjunto de subgradientes de f em $\mathbf{w}$
$\|\cdot\|$	Norma Euclidiana
$\langle \cdot, \cdot \rangle$	Produto interno
$\arg\min_{t \in \mathcal{T}} g(\mathbf{x}^{(t)})$	Índice(s) que minimiza(m) g sobre um conjunto $\mathcal{T}$
$\max_{i \in \mathcal{I}} g_i(\mathbf{w})$	Valor máximo de uma família de funções $\{g_i\}_{i \in \mathcal{I}}$ em $\mathbf{w}$

Sumário

1 Introdução 8

2 Gradiente Descendente 10

2.1 Gradiente de uma Função 10

2.2 Método do Gradiente Descendente 11

2.3 Aproximação de Taylor e Convexidade 13

2.4 Análise do GD para funções convexas e Lipschitz 17

3 Subgradiente 20

3.1 Calculando subgradientes 21

3.2 Subgradientes de funções Lipschitz 22

3.2.1 O Algoritmo do subgradiente 23

4 Gradiente descendente estocástico 26

4.1 Otimização de funções limitadas convexas Lipschitzianas 26

5 Aplicações 30

5.1 Análise descritiva dos dados 30

5.2 Modelagem dos dados via misturas de normais 33

5.3 Resultados 35

6 Conclusão 38

Referências 39

1 Introdução

O método do gradiente descendente (GD) é uma das abordagens mais fundamentais para a otimização de funções, sendo amplamente utilizado em problemas de aprendizado de máquina e estatística. Ele consiste em atualizar iterativamente uma variável de decisão na direção oposta ao gradiente da função objetivo, garantindo, sob certas condições, a convergência para um ponto de mínimo. No entanto, quando se trabalha com grandes conjuntos de dados, o cálculo exato do gradiente pode ser computacionalmente inviável, motivo pelo qual surge o gradiente descendente estocástico (GDE), que utiliza amostras aleatórias da distribuição dos dados para estimar o gradiente, reduzindo o custo computacional e permitindo uma atualização mais eficiente dos parâmetros, (SHALEV-SHWARTZ; BEN-DAVID, 2014).

Para garantir a convergência do GD e do GDE a um mínimo global, uma hipótese crucial imposta sobre a função objetivo é a convexidade. Uma função convexa tem a propriedade de que qualquer mínimo local é também um mínimo global, assegurando que a sequência gerada pelo algoritmo se aproxima de uma solução ótima. Além disso, a convexidade permite a existência de garantias formais sobre a taxa de convergência, muitas vezes expressas em termos da norma do gradiente ou da subotimalidade do valor da função ao longo das iterações. Sob essa hipótese, o GD e o GDE são métodos robustos e eficientes, amplamente utilizados em regressão linear, aprendizado de máquinas e redes neurais em suas formas mais básicas (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018).

Embora o comportamento do GDE seja bem compreendido no contexto de problemas de aprendizagem convexos, os resultados teóricos disponíveis para sua aplicação a funções objetivo não convexas ainda não apresentam um progresso consistente quando considerados aspectos relevantes para aplicações práticas. Existem, por exemplo, resultados que exigem a imposição de hipóteses não triviais, como a existência de limites uniformes para os gradientes ao longo de todas as iterações do processo de aprendizagem, condição que é difícil de verificar em implementações reais (LEI et al., 2019). Nesse contexto, torna-se pertinente investigar o algoritmo de forma mais detalhada, a fim de compreender em que etapas surgem limitações técnicas associadas, por exemplo, à suposição de convexidade.

O presente trabalho tem como objetivo apresentar, discutir e aplicar o algoritmo do gradiente descendente estocástico (GDE), com ênfase em problemas de otimização em contextos estatísticos, como a maximização da função de verossimilhança e a regressão logística.

No Capítulo 2, discute-se o algoritmo do gradiente descendente determinístico, bem como suas derivações intuitivas a partir da direção do vetor gradiente e da expansão da função em série de Taylor. Em seguida, o método é analisado no contexto de funções convexas e Lipschitz-contínuas.

No Capítulo 3, o algoritmo é estendido para o caso de funções que não são diferenciáveis em certas regiões do domínio. Nesse contexto, introduz-se o conceito de subgradiente e apresenta-se a versão adaptada do algoritmo baseada nessa generalização.

No Capítulo 4, apresenta-se e discute-se a versão estocástica do gradiente descendente, com ênfase em funções convexas, limitadas e Lipschitz-contínuas. Essa classe de funções contínuas permite estabelecer condições de convergência do algoritmo, bem como controlar a taxa de aprendizagem em termos da constante de Lipschitz.

Por fim, no Capítulo 5, é apresentada uma aplicação simples com o objetivo de ilustrar o uso e a eficiência do gradiente descendente estocástico em um problema de otimização estatística.

2 Gradiente Descendente

A principal vantagem dos métodos de Newton e quasi-Newton na resolução de sistemas não lineares está em suas elevadas velocidades de convergência, desde que se disponha de uma aproximação inicial suficientemente próxima da solução. Em contrapartida, uma limitação fundamental desses métodos reside justamente na necessidade dessa boa estimativa inicial, uma vez que a convergência não é garantida quando tal condição não é satisfeita.

Além disso, considerando a ampla aplicação desses algoritmos no contexto de redes neurais e, consequentemente, em Machine Learning, problemas de convergência podem resultar em um elevado consumo energético. Nesse cenário, o gradiente descendente (Gradient descent) apresenta-se como uma alternativa relevante, pois pode ser utilizado tanto em problemas de otimização no contexto de redes neurais quanto em aplicações mais amplas relacionadas à aprendizagem estatística (GOODFELLOW; BENGIO; COURVILLE, 2016).

De modo geral, o método do gradiente descendente tem como objetivo encontrar um mínimo local de uma função $g : \mathbb{R}^n \to \mathbb{R}$. Intuitivamente, o método baseia-se na ideia de atualizar iterativamente a solução na direção oposta ao gradiente da função, conforme os passos descritos a seguir:

1. Avaliar g em uma solução inicial $\mathbf{x}^{(0)} = (x_1^{(0)}, \ldots, x_n^{(0)})^T$.
2. Determinar uma direção Δx que resulte no decrescimento do valor de g.
3. Mover uma quantidade apropriada nessa direção e chamar esse novo valor de $x^{(1)}$.
4. Repetir os passos 1 a 3 com $x^{(1)}$ substituído por $x^{(0)}$.

Nas próximas seções detalhar-se-ão as principais características técnicas desse algorimto.

2.1 Gradiente de uma Função

Definição 1. *O **gradiente** de uma função $g : \mathbb{R}^n \to \mathbb{R}$ em $\mathbf{x} = (x_1, \ldots, x_n)$ é denotado por $\boldsymbol{\nabla} g(\mathbf{x})$ e definido por:*

$$\boldsymbol{\nabla} g(x) = \left(\frac{\partial g}{\partial x_1}(x), \ldots, \frac{\partial g}{\partial x_n}(x) \right).$$

O gradiente de uma função multivariada é análogo à derivada de uma função real, no sentido que uma função *diferenciável* possui um mínimo em $\mathbf{x}$ se, e somente se, o gradiente em $\mathbf{x}$ é o vetor nulo.

Definição 2. *Seja* $\mathbf{v} = (v_1, \ldots, v_n)$ *um vetor unitário em* $\mathbb{R}^n$*, isto é,* $\|\mathbf{v}\| = 1$*. A* ***derivada direcional*** *de* g *em* $\mathbf{x}$ *na direção* $\mathbf{v}$ *é definida por:*

$$D_{\mathbf{v}} g(\mathbf{x}) = \lim_{\epsilon \to 0} \frac{g(\mathbf{x} + \epsilon \mathbf{v}) - g(\mathbf{x})}{\epsilon} = \mathbf{v}^\top \boldsymbol{\nabla} g(\mathbf{x}).$$

Note que $D_{\mathbf{v}} g(\mathbf{x})$ mede a mudança no valor da função g relativa à mudança na variável na direção de $\mathbf{v}$. Quando g é diferenciável, a direção que produz o maior valor para a derivada direcional ocorre quando $\mathbf{v}$ é escolhido como sendo paralelo a $\boldsymbol{\nabla} g(\mathbf{x})$, desde que $\boldsymbol{\nabla} g(\mathbf{x}) \neq 0$. Como consequência, a direção de maior decrescimento no valor de g em x é a direção dada por $-\boldsymbol{\nabla} g(\mathbf{x})$.

2.2 Método do Gradiente Descendente

O gradiente descendente é um método iterativo que inicia com uma solução inicial ($\mathbf{x}^{(0)}$). Em cada iteração tomamos um *passo* na direção do gradiente negativo no ponto atual. Isto é, a atualização é feita por:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta \boldsymbol{\nabla} g(\mathbf{x}^{(t)}),$$

onde η é um parâmetro de passo a ser discutido posteriormente também denominado por taxa de aprendizagem. Após T iterações, a saída pode ser:

- $\overline{\mathbf{x}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}^{(t)}$;
- $\mathbf{x}^{(T)}$;
- $\text{argmin}_{t \in \{1,\ldots,T\}} g\big(\mathbf{x}^{(t)}\big)$.

Exemplo 1. *Considere* $f(x, y) = x^2 + y^2$*, que pode ser observada graficamente na Figura 1*

Podemos empregar o algoritmo do gradiente descendente para determinar o seu ponto de mínimo. O gradiente é dado por:

$$\boldsymbol{\nabla} f(x, y) = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} = \begin{bmatrix} 2x \\ 2y \end{bmatrix}.$$

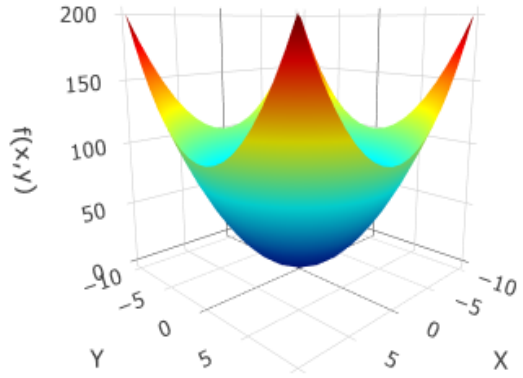


Figura 1: $f(x, y) = x^2 + y^2$

A regra de atualização dos parâmetros (x, y) *é:*

$$x_{novo} = x - \eta \frac{\partial f}{\partial x}, \quad y_{novo} = y - \eta \frac{\partial f}{\partial y}.$$

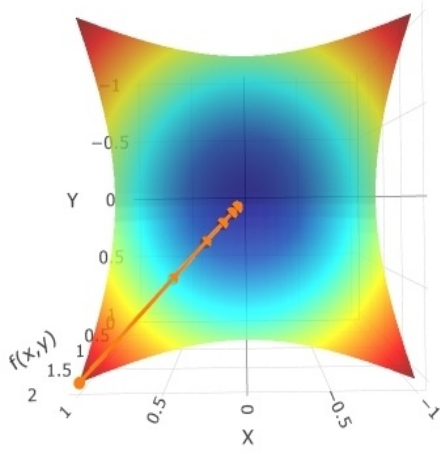
Com $\eta = 0.2$*, temos as seguintes equações de atualização:*

$$x_{novo} = x - 0.2 \cdot 2x = x(1 - 0.4)$$
$$y_{novo} = y - 0.2 \cdot 2y = y(1 - 0.4)$$

Portanto, a cada iteração, os valores de x *e* y *são multiplicados por* 0.6*. Supondo valores iniciais* $(x_0, y_0) = (1, 1)$*, calculamos as iterações abaixo.*

- *Iteração 0:* $x_0 = 1, y_0 = 1$*,* $f(x_0, y_0) = 1^2 + 1^2 = 2$*.*
- *Iteração 1:* $x_1 = 0.6 \cdot x_0 = 0.6$*,* $y_1 = 0.6 \cdot y_0 = 0.6$*,* $f(x_1, y_1) = (0.6)^2 + (0.6)^2 = 0.72$*.*
- *Iteração 2:* $x_2 = 0.6 \cdot x_1 = 0.36$*,* $y_2 = 0.6 \cdot y_1 = 0.36$*,* $f(x_2, y_2) = (0.36)^2 + (0.36)^2 = 0.2592$*.*
- *Iteração 3:* $x_3 = 0.6 \cdot x_2 = 0.216$*,* $y_3 = 0.6 \cdot y_2 = 0.216$*,* $f(x_3, y_3) = (0.216)^2 + (0.216)^2 = 0.093312$*.*

É possível continuar iterando até a convergência. Note que os valores de x *e* y *tendem a* 0*, e* $f(x, y)$ *tende ao mínimo* 0*.*



(a) Modelo 2

2.3 Aproximação de Taylor e Convexidade

Definição 3. *Seja V um espaço vetorial sobre $\mathbb{K}$, onde $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$, (HOFFMAN; KUNZE, 1971). Um* ***produto interno*** *em V é uma aplicação*

$$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{K}$$

que satisfaz, para todos $u, v, w \in V$ e $\alpha \in \mathbb{K}$:

1. *Linearidade: $\langle \alpha u + v, w \rangle = \alpha \langle u, w \rangle + \langle v, w \rangle$*
2. *Simetria (ou hermiticidade): $\langle u, v \rangle = \overline{\langle v, u \rangle}$*
3. *Positividade definida: $\langle v, v \rangle \geq 0 \quad e \quad \langle v, v \rangle = 0 \iff v = 0$*

Exemplo 2. *Para vetores $\boldsymbol{u} = (u_1, \ldots, u_n)$ e $\boldsymbol{v} = (v_1, \ldots, v_n)$ em $\mathbb{R}^n$, definimos:*

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \sum_{i=1}^{n} u_i v_i$$

Este produto interno satisfaz todas as propriedades: linearidade, simetria e positividade.

Exemplo 3. *Seja $V = C[a, b]$, o espaço das funções contínuas em $[a, b]$. Definimos*

$$\langle f, g \rangle = \int_a^b f(x) g(x)\, dx,$$

que satizfaz as propriedades da Definição 3. De fato

- *Linearidade:*

$$\langle \alpha f + h, g \rangle = \int_a^b (\alpha f(x) + h(x)) g(x)\, dx = \alpha \int_a^b f(x) g(x)\, dx + \int_a^b h(x) g(x)\, dx$$

- *Simetria:*

$$\langle f, g \rangle = \int_a^b f(x) g(x)\, dx = \int_a^b g(x) f(x)\, dx = \langle g, f \rangle$$

- *Positividade:*

$$\langle f, f \rangle = \int_a^b f(x)^2\, dx \geq 0$$

  e

$$\langle f, f \rangle = 0 \iff f(x) = 0 \text{ para todo } x \in [a, b]$$

Logo, esta integral define um produto interno em $C[a, b]$.

Seja $f : \mathbb{R}^d \to \mathbb{R}$ uma função $(k+1)$-vezes continuamente diferenciável em $\mathbf{a} \in \mathbb{R}^d$; a expansão de Taylor com resto integral fornece a identidade

$$\begin{aligned} f(\mathbf{x}) = f(\mathbf{a}) &+ \sum_{\substack{\boldsymbol{\alpha} \in \mathbb{N}^d \\ 1 \leq |\boldsymbol{\alpha}| \leq k}} \frac{1}{\boldsymbol{\alpha}!} D^{\boldsymbol{\alpha}} f(\mathbf{a}) (\mathbf{x} - \mathbf{a})^{\boldsymbol{\alpha}} \\ &+ \sum_{\substack{\boldsymbol{\alpha} \in \mathbb{N}^d \\ |\boldsymbol{\alpha}| = k+1}} (\mathbf{x} - \mathbf{a})^{\boldsymbol{\alpha}} \int_0^1 (1 - s)^k D^{\boldsymbol{\alpha}} f\big(\mathbf{a} + s(\mathbf{x} - \mathbf{a})\big)\, ds. \end{aligned}$$

Ao tomar $k = 1$, os termos de primeira ordem correspondem à diferencial e obtém-se a forma explícita

$$f(\mathbf{x}) = f(\mathbf{a}) + \langle \mathbf{x} - \mathbf{a}, \boldsymbol{\nabla} f(\mathbf{a}) \rangle + R_2(\mathbf{x}; \mathbf{a}),$$

onde o termo $R_2(\mathbf{x}; \mathbf{a})$ é o resto de ordem dois, expresso pela integral dos segundos derivados ao longo do segmento entre $\mathbf{a}$ e $\mathbf{x}$. Desprezando esse resto obtém-se a aproximação linear

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \langle \mathbf{x} - \mathbf{a}, \boldsymbol{\nabla} f(\mathbf{a}) \rangle,$$

a qual fornece um modelo local de primeira ordem válido quando $\mathbf{x}$ está suficientemente próximo de $\mathbf{a}$. É precisamente a partir desta aproximação linear que se pode motivar o passo elementar do gradiente descendente: para reduzir o valor aproximado de f a partir do ponto $\mathbf{a}$, basta mover-se na direção que minimiza a aproximação linear, ou seja, na direção oposta ao gradiente. Assim, o método itera a regra

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \boldsymbol{\nabla} f(\mathbf{x}_t),$$

onde $\eta_t > 0$ é o parâmetro de passo ou taxa de aprendizado. A escolha desta regra decorre diretamente da otimização do modelo linear unidimensional em torno de $\mathbf{x}_t$: tomando $\mathbf{x} = \mathbf{x}_t + s\mathbf{d}$ com $\mathbf{d}$ de norma unitária, a aproximação linear

$$f(\mathbf{x}_t + s\mathbf{d}) \approx f(\mathbf{x}_t) + s\langle \mathbf{d}, \boldsymbol{\nabla} f(\mathbf{x}_t)\rangle$$

é minimizada quando $\mathbf{d} = -\boldsymbol{\nabla} f(\mathbf{x}_t)/\|\boldsymbol{\nabla} f(\mathbf{x}_t)\|$, justificando a escolha da direção de descida. A escala do deslocamento é então regulada por η_t.

Para além da motivação heurística via Taylor de primeira ordem, a análise da convergência exige controlar o efeito do resto de ordem superior, (SHALEV-SHWARTZ; BEN-DAVID, 2014). Uma hipótese comum e conveniente é que f possua gradiente Lipschitz contínuo (LIMA, 2025) com constante $L > 0$, isto é,

$$\|\boldsymbol{\nabla} f(\mathbf{y}) - \boldsymbol{\nabla} f(\mathbf{x})\| \leq L\|\mathbf{y} - \mathbf{x}\| \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

Sob essa hipótese a expansão de Taylor com estimativa de segundo ordem implica a desigualdade de suavidade

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \boldsymbol{\nabla} f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

Aplicando-a com $\mathbf{y} = \mathbf{x} - \eta\boldsymbol{\nabla} f(\mathbf{x})$ obtém-se

$$f(\mathbf{x} - \eta\boldsymbol{\nabla} f(\mathbf{x})) \leq f(\mathbf{x}) - \eta\|\boldsymbol{\nabla} f(\mathbf{x})\|^2 + \frac{L\eta^2}{2}\|\boldsymbol{\nabla} f(\mathbf{x})\|^2 = f(\mathbf{x}) - \eta\Big(1 - \frac{L\eta}{2}\Big)\|\boldsymbol{\nabla} f(\mathbf{x})\|^2.$$

Dessa desigualdade segue-se imediatamente uma condição suficiente para que o passo reduza o valor da função, isto é, qualquer η com $0 < \eta < 2/L$ garante que o fator entre parênteses seja positivo e, portanto, $f(\mathbf{x}_{t+1}) < f(\mathbf{x}_t)$ sempre que $\boldsymbol{\nabla} f(\mathbf{x}_t) \neq 0$, (SHALEV-SHWARTZ; BEN-DAVID, 2014). Esta estimativa explicita como a informação da segunda ordem controla, através de L, o comprimento admissível do passo para que a aproximação linear permaneça representativa.

A convexidade tem papel central aqui, pois transforma garantias locais em garantias globais. Uma função $f : I \to \mathbb{R}$ é convexa se, para quaisquer $x, y \in I$,

$$f(x) \geq f(y) + f'(y)(x - y)$$

no caso unidimensional, ou, em várias dimensões, se definida em um conjunto convexo C satisfaz

$$f\big(t\mathbf{x} + (1-t)\mathbf{y}\big) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y}), \qquad \forall\, t \in [0,1].$$

Quando f é convexa, a desigualdade de primeira ordem é generalizada para

$$f(\mathbf{x}) \geq f(\mathbf{a}) + \langle \mathbf{x} - \mathbf{a}, \boldsymbol{\nabla} f(\mathbf{a}) \rangle.$$

Tal expressão indica que o plano tangente em qualquer ponto subestima globalmente a função. Em termos de otimização, isso implica que qualquer ponto estacionário[1] é um mínimo global, ou seja, não existem outros mínimos locais que não sejam globais. Consequentemente, para funções convexas e suaves a propriedade de monotonicidade do método aliada à não existência de mínimos locais indesejáveis traduz-se em convergência do método para um mínimo global, sob condições adequadas sobre a escolha de η_t.

Quando, além da convexidade, se assume *forte convexidade* com parâmetro $\mu > 0$, i.e. para todo $\mathbf{x}, \mathbf{y}$

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \boldsymbol{\nabla} f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2,$$

obtém-se uma desigualdade que relaciona o desvio funcional e a norma do gradiente,

$$\|\boldsymbol{\nabla} f(\mathbf{x})\|^2 \geq 2\mu \big(f(\mathbf{x}) - f(\mathbf{x}^\star) \big), \tag{2.3.1}$$

onde $\mathbf{x}^\star$ denota o minimizador único de f. Combinando esta desigualdade (2.3.1) com a estimativa de decaimento já obtida para o passo do gradiente, tem-se

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) \leq \Big(1 - 2\eta\mu\big(1 - \tfrac{L\eta}{2}\big) \Big) \big(f(\mathbf{x}_t) - f(\mathbf{x}^\star) \big).$$

Escolhendo, por exemplo, $0 < \eta \leq 1/L$ garante-se $(1 - \frac{L\eta}{2}) \geq 1/2$ e, portanto, um fator de contração simples

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) \leq (1 - \eta\mu) \big(f(\mathbf{x}_t) - f(\mathbf{x}^\star) \big).$$

Isto demonstra taxa de convergência linear do erro em função do número de iterações: a forte convexidade transforma as garantias de descida em uma lei de decaimento exponencial do excesso de custo, (BRINK; RICHARDS; TETHEROLF, 2016).

Assim, o gradiente descendente derivado a partir da série de Taylor explica intuitivamente por que o negativo do gradiente é a direção de descida local, assim como, a incorporação do termo quadrático por meio da constante de Lipschitz do gradiente fornece condições práticas sobre o tamanho do passo. Paralelamente a convexidade assegura que

[1] $\boldsymbol{\nabla} f(\mathbf{x}) = 0$.

a diminuição local se traduza em convergência global, frequentemente com taxa linear quando se dispõe de hipóteses necessárias, (SHALEV-SHWARTZ; BEN-DAVID, 2014).

2.4 Análise do GD para funções convexas e Lipschitz

Para analisar a taxa de convergência do algoritmo de Gradiente Descendente (GD), restringimo-nos ao caso de funções convexas e Lipschitz. Seja $\mathbf{w}^\star$ um vetor arbitrário e seja B uma cota superior para $\|\mathbf{w}^\star\|$. É conveniente interpretar $\mathbf{w}^\star$ como um minimizador de $f(\mathbf{w})$, embora a análise a seguir seja válida para qualquer $\mathbf{w}^\star$.

Nosso objetivo é obter uma cota superior para a suboptimalidade da solução em relação a $\mathbf{w}^\star$, isto é,

$$f(\overline{\mathbf{w}}) - f(\mathbf{w}^\star),$$

onde

$$\overline{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}^{(t)}.$$

Pela definição de $\overline{\mathbf{w}}$ e pela desigualdade de Jensen, temos

$$\begin{aligned} f(\overline{\mathbf{w}}) - f(\mathbf{w}^\star) &= f\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{w}^{(t)}\right) - f(\mathbf{w}^\star) \\ &\leq \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{w}^{(t)}) - f(\mathbf{w}^\star) \\ &= \frac{1}{T}\sum_{t=1}^{T} \left(f(\mathbf{w}^{(t)}) - f(\mathbf{w}^\star)\right). \end{aligned} \tag{2.4.1}$$

Para todo t, pela convexidade de f, vale

$$f(\mathbf{w}^{(t)}) - f(\mathbf{w}^\star) \leq \left\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \nabla f(\mathbf{w}^{(t)}) \right\rangle. \tag{2.4.2}$$

Combinando as desigualdades anteriores, obtemos

$$f(\overline{\mathbf{w}}) - f(\mathbf{w}^\star) \leq \frac{1}{T}\sum_{t=1}^{T} \left\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \nabla f(\mathbf{w}^{(t)}) \right\rangle.$$

Para limitar o lado direito da desigualdade, utilizamos o lema a seguir.

Lema 2.1. *Sejam $\mathbf{v}_1, \ldots, \mathbf{v}_T$ vetores arbitrários. Qualquer algoritmo inicializado em*

$\mathbf{w}^{(1)} = 0$ *e com regra de atualização*

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t \tag{2.4.3}$$

satisfaz

$$\sum_{t=1}^{T} \left\langle \mathbf{w}^{(t)} - \mathbf{w}^{\star}, \mathbf{v}_t \right\rangle \leq \frac{\|\mathbf{w}^{\star}\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2. \tag{2.4.4}$$

Em particular, para quaisquer $B, \rho > 0$, se $\|\mathbf{v}_t\| \leq \rho$ para todo t e se

$$\eta = \sqrt{\frac{B^2}{\rho^2 T}},$$

então, para todo $\mathbf{w}^{\star}$ tal que $\|\mathbf{w}^{\star}\| \leq B$, vale

$$\frac{1}{T} \sum_{t=1}^{T} \left\langle \mathbf{w}^{(t)} - \mathbf{w}^{\star}, \mathbf{v}_t \right\rangle \leq \frac{B\rho}{\sqrt{T}}.$$

Demonstração. Utilizando manipulações algébricas, obtem-se

$$\begin{aligned}
\left\langle \mathbf{w}^{(t)} - \mathbf{w}^{\star}, \mathbf{v}_t \right\rangle &= \frac{1}{\eta} \left\langle \mathbf{w}^{(t)} - \mathbf{w}^{\star}, \eta \mathbf{v}_t \right\rangle \\
&= \frac{1}{2\eta} \Big(- \|\mathbf{w}^{(t)} - \mathbf{w}^{\star} - \eta \mathbf{v}_t\|^2 + \|\mathbf{w}^{(t)} - \mathbf{w}^{\star}\|^2 + \eta^2 \|\mathbf{v}_t\|^2 \Big) \\
&= \frac{1}{2\eta} \Big(- \|\mathbf{w}^{(t+1)} - \mathbf{w}^{\star}\|^2 + \|\mathbf{w}^{(t)} - \mathbf{w}^{\star}\|^2 \Big) + \frac{\eta}{2} \|\mathbf{v}_t\|^2.
\end{aligned}$$

Somando para $t = 1, \ldots, T$, temos

$$\sum_{t=1}^{T} \left\langle \mathbf{w}^{(t)} - \mathbf{w}^{\star}, \mathbf{v}_t \right\rangle = \frac{1}{2\eta} \sum_{t=1}^{T} \Big(- \|\mathbf{w}^{(t+1)} - \mathbf{w}^{\star}\|^2 + \|\mathbf{w}^{(t)} - \mathbf{w}^{\star}\|^2 \Big) + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2. \tag{2.4.5}$$

A soma telescópica resulta em

$$\|\mathbf{w}^{(1)} - \mathbf{w}^{\star}\|^2 - \|\mathbf{w}^{(T+1)} - \mathbf{w}^{\star}\|^2.$$

Como $\mathbf{w}^{(1)} = 0$, obtemos

$$\sum_{t=1}^{T} \left\langle \mathbf{w}^{(t)} - \mathbf{w}^{\star}, \mathbf{v}_t \right\rangle \leq \frac{\|\mathbf{w}^{\star}\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2,$$

concluindo a prova. □

O Lema 2.1 aplica-se ao algoritmo GD tomando $\mathbf{v}_t = \nabla f(\mathbf{w}^{(t)})$. Se f é ρ-

Lipschitz, então $\|\nabla f(\mathbf{w}^{(t)})\| \leq \rho$. Assim, obtem-se o seguinte resultado.

Corolário 2.2. *Seja f uma função convexa e ρ-Lipschitz, e seja $\mathbf{w}^\star \in \arg\min_{\|\mathbf{w}\|\leq B} f(\mathbf{w})$.*

Se o algoritmo GD for executado por T iterações com

$$\eta = \sqrt{\frac{B^2}{\rho^2 T}},$$

então o vetor médio $\overline{\mathbf{w}}$ satisfaz

$$f(\overline{\mathbf{w}}) - f(\mathbf{w}^\star) \leq \frac{B\rho}{\sqrt{T}}.$$

Além disso, para garantir $f(\overline{\mathbf{w}}) - f(\mathbf{w}^\star) \leq \epsilon$, basta escolher

$$T \geq \frac{B^2\rho^2}{\epsilon^2}.$$

Exemplo 4. *Considere a função $f(x,y) = \frac{5}{4}(x+6)^2+(y-8)^2$, e o domínio $D = \mathrm{Dom}(f) = B((0,0),10) = \{(x,y) \in \mathbb{R}^2;\ \|(x,y)\| \leq 10\}$.*

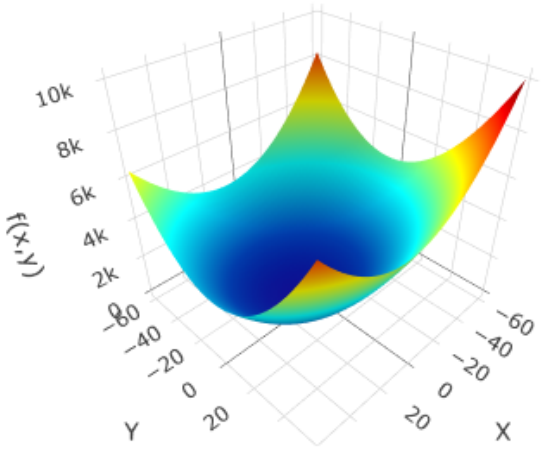


Figura 2: $f(x,y) = \frac{5}{4}(x+6)^2 + (y-8)^2$

Em particular, para todo $(x,y) \in D$, temos $|x| \leq 10$ e $|y| \leq 10$. Tomemos $(x_1,y_1),(x_2,y_2) \in D$. Pela desigualdade triangular,

$$|f(x_1,y_1) - f(x_2,y_2)| \leq \frac{5}{4}\left|(x_1+6)^2 - (x_2+6)^2\right| + \left|(y_1-8)^2 - (y_2-8)^2\right|.$$

Usando a identidade da diferença de quadrados,

$$(x_1+6)^2-(x_2+6)^2=(x_1+x_2+12)(x_1-x_2),$$

$$(y_1-8)^2-(y_2-8)^2=(y_1+y_2-16)(y_1-y_2).$$

Como $(x_1,y_1),(x_2,y_2)\in D$*, segue que*

$$|x_1+x_2+12|\le|x_1|+|x_2|+12\le 32,\qquad |y_1+y_2-16|\le|y_1|+|y_2|+16\le 36.$$

Portanto,

$$\frac{5}{4}\big|(x_1+6)^2-(x_2+6)^2\big|\le 40|x_1-x_2|,$$
$$\big|(y_1-8)^2-(y_2-8)^2\big|\le 36|y_1-y_2|\le 40|y_1-y_2|.$$

Assim,

$$\begin{aligned}|f(x_1,y_1)-f(x_2,y_2)|&\le 40\big(|x_1-x_2|+|y_1-y_2|\big)\\&\le 40\sqrt{2}\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}\\&=40\sqrt{2}\,\|(x_1,y_1)-(x_2,y_2)\|.\end{aligned}$$

Logo, a função f *é* $40\sqrt{2}$*-Lipschitz em* D*. A regra de atualização do Gradiente Descendente é*

$$\omega^{t+1}=\omega^t-\eta\nabla f(\omega^t),$$

onde o passo pode ser escolhido como

$$\eta=\sqrt{\frac{B^2}{T\rho^2}}=\sqrt{\frac{B^2}{T(40\sqrt{2})^2}}=\frac{B}{40\sqrt{2T}}.$$

3 Subgradiente

O algoritmo GD requer que a função f seja diferenciável. Agora vamos generalizar a discussão para além das funções diferenciáveis. Mostraremos que o algoritmo GD pode ser aplicado a funções não diferenciáveis utilizando um recurso chamado subgradiente de $f(\mathbf{w})$ em $\mathbf{w}^{(t)}$, em vez do gradiente.

Para motivar a definição de subgradientes, lembre-se de que, para uma função convexa f, o gradiente em $\mathbf{w}$ define a inclinação de uma tangente que fica abaixo de f,

ou seja,

$$\forall \mathbf{u}, \quad f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \nabla f(\mathbf{w}) \rangle \tag{14.7}$$

A existência de uma tangente que fica abaixo de f é uma propriedade importante das funções convexas, que é, na verdade, uma caracterização alternativa da convexidade.

Lema 3.1. *Seja S um conjunto convexo aberto. Uma função $f : S \to \mathbb{R}$ é convexa se, e somente se, para cada $\mathbf{w} \in S$ existe $\mathbf{v}$ tal que*

$$\forall \mathbf{u} \in S, \quad f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \mathbf{v} \rangle \tag{3.0.1}$$

A prova deste lema pode ser encontrada em muitos livros de análise convexa (por exemplo, (BORWEIN; LEWIS, 2005)). A desigualdade precedente nos leva à definição de subgradientes.

Definição 4. *(Subgradientes). Um vetor $\mathbf{v}$ que satisfaz a Equação 3.0.1 é chamado de subgradiente de f em $\mathbf{w}$. O conjunto de subgradientes de f em $\mathbf{w}$ é chamado de conjunto de subdiferenciais e é denotado por $\partial f(\mathbf{w})$.*

A Figura 3 ilustra um exemplo de função convexa que não é diferenciável em todo o seu domínio. Nesse caso, destaca-se em vermelho um subgradiente no ponto de não diferenciabilidade, representando uma das retas de suporte à função nesse ponto.

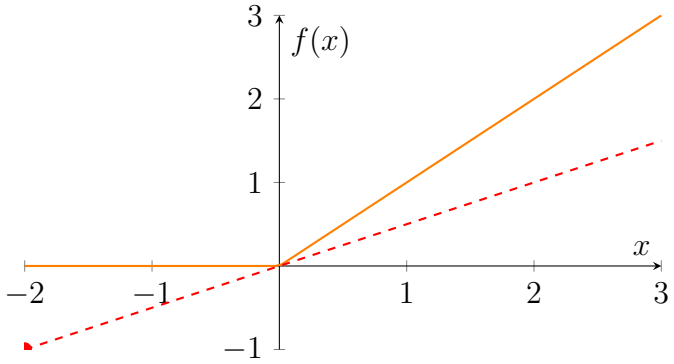


Figura 3: Função convexa $f(x) = \max(0, x)$ e um subgradiente no ponto não diferenciável $w = (0, 0)$.

3.1 Calculando subgradientes

Como construímos subgradientes de uma dada função convexa? Se uma função é diferenciável em um ponto $\mathbf{w}$, então o conjunto diferencial é trivial, como mostra a

seguinte afirmação.

Proposição 3.2. *Se f é diferenciável em $\mathbf{w}$, então $\partial f(\mathbf{w})$ contém um único elemento - o gradiente de f em $\mathbf{w}, \nabla f(\mathbf{w})$.*

Demonstração. A demonstração pode ser encontrada em (SHALEV-SHWARTZ; BEN-DAVID, 2014). □

Exemplo 5. *(O Conjunto Diferencial da Função Modular). Considere a função valor absoluto $f(x) = |x|$. Usando a proposição 3.2, podemos construir o conjunto diferencial para as partes diferenciáveis de f, e o único ponto que requer atenção especial é $x_0 = 0$. Nesse ponto, é possível verificar que o subdiferencial é o conjunto de todos os números entre -1 e 1. Portanto:*

$$\partial f(x) = \begin{cases} \{1\} & se\ x > 0 \\ \{-1\} & se\ x < 0 \\ (-1, 1) & se\ x = 0 \end{cases}$$

Para muitos usos práticos, não precisamos calcular todo o conjunto de subgradientes em um ponto dado, pois um membro desse conjunto é suficiente. A afirmação a seguir mostra como construir um subgradiente para funções de máximo pontual.

Proposição 3.3. *Seja $g(\mathbf{w}) = \max_{i \in [r]} g_i(\mathbf{w})$ para r funções convexas diferenciáveis $g_1, \ldots, g_r$. Dado algum $\mathbf{w}$, seja $j \in \operatorname{argmax}_i g_i(\mathbf{w})$. Então $\nabla g_j(\mathbf{w}) \in \partial g(\mathbf{w})$.*

Demonstração. Como g_j é convexa, temos que para todo $\mathbf{u}$,

$$g_j(\mathbf{u}) \geq g_j(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \nabla g_j(\mathbf{w}) \rangle .$$

Como $g(\mathbf{w}) = g_j(\mathbf{w})$ e $g(\mathbf{u}) \geq g_j(\mathbf{u})$, obtemos que

$$g(\mathbf{u}) \geq g(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \nabla g_j(\mathbf{w}) \rangle .$$

O que conclui nossa prova. □

3.2 Subgradientes de funções Lipschitz

Relembre que uma função $f : A \to \mathbb{R}$ é ρ-Lipschitz se para todos $\mathbf{u}, \mathbf{v} \in A$

$$|f(\mathbf{u}) - f(\mathbf{v})| \leq \rho\|\mathbf{u} - \mathbf{v}\|.$$

O seguinte lema dá uma definição equivalente usando normas de subgradientes.

Lema 3.4. *Seja A um conjunto convexo aberto e seja $f : A \to \mathbb{R}$ uma função convexa. Então, f é ρ-Lipschitz sobre A se, e somente se, para todos $\mathbf{w} \in A$ e $\mathbf{v} \in \partial f(\mathbf{w})$ temos que $\|\mathbf{v}\| \leq \rho$.*

Demonstração. Suponha que para todos $\mathbf{v} \in \partial f(\mathbf{w})$ temos que $\|\mathbf{v}\| \leq \rho$. Como $\mathbf{v} \in \partial f(\mathbf{w})$, temos

$$f(\mathbf{w}) - f(\mathbf{u}) \leq \langle \mathbf{v}, \mathbf{w} - \mathbf{u} \rangle.$$

Limitando o lado direito usando a desigualdade de Cauchy-Schwartz, obtemos

$$f(\mathbf{w}) - f(\mathbf{u}) \leq \langle \mathbf{v}, \mathbf{w} - \mathbf{u} \rangle \leq \|\mathbf{v}\|\|\mathbf{w} - \mathbf{u}\| \leq \rho\|\mathbf{w} - \mathbf{u}\|.$$

Um argumento análogo pode mostrar que $f(\mathbf{u}) - f(\mathbf{w}) \leq \rho\|\mathbf{w} - \mathbf{u}\|$. Portanto, f é ρ-Lipschitz. Agora, suponha que f é ρ-Lipschitz. Escolha algum $\mathbf{w} \in A$, $\mathbf{v} \in \partial f(\mathbf{w})$. Como A é aberto, existe $\epsilon > 0$ tal que $\mathbf{u} = \mathbf{w} + \epsilon\mathbf{v}/\|\mathbf{v}\|$ pertence a A. Portanto, $\langle \mathbf{u} - \mathbf{w}, \mathbf{v} \rangle = \epsilon\|\mathbf{v}\|$ e $\|\mathbf{u} - \mathbf{w}\| = \epsilon$. Pela definição de subgradiente,

$$f(\mathbf{u}) - f(\mathbf{w}) \geq \langle \mathbf{v}, \mathbf{u} - \mathbf{w} \rangle = \epsilon\|\mathbf{v}\|.$$

Por outro lado, pela propriedade Lipschitz de f, temos

$$\rho_\epsilon = \rho\|\mathbf{u} - \mathbf{w}\| \geq f(\mathbf{u}) - f(\mathbf{w}).$$

Combinando as duas desigualdades, concluímos que $\|\mathbf{v}\| \leq \rho$. □

3.2.1 O Algoritmo do subgradiente

O método do subgradiente requer a definição de uma função convexa Lipschitz $f : \mathbb{R}^n \to \mathbb{R}$ para assegurar a existência de subgradientes $g_n \in \partial f(p_n)$ e a estabilidade das iterações. O conjunto viável $K \subset \mathbb{R}^n$, convexo, fechado e limitado, é essencial para representar restrições e garantir que a projeção $\Pi_K(y_n)$ seja bem definida e única. O ponto inicial $p_0 \in K$ assegura que o algoritmo comece em uma posição viável, enquanto a sequência de passos $\{\gamma_n\}_{n\geq 0}$, com $\gamma_n > 0$, controla o deslocamento na direção de descida $y_n = p_n - \gamma_n g_n$.

A tolerância $\varepsilon > 0$ define um critério de parada baseado na norma do subgradiente, indicando proximidade de um minimizador, e o número máximo de iterações $N_{\max}$ garante a finalização do algoritmo. Em conjunto, esses termos permitem gerar a sequência $\{p_n\} \subset K$ de forma bem definida, assegurando coerência teórica e viabilidade computacional do método, (SHALEV-SHWARTZ; BEN-DAVID, 2014).

Algoritmo 1 - Método do Subgradiente

Ensure: Aproximação $x^* \in K$

$n \leftarrow 0$
$p_n \leftarrow p_0$
while $n < N_{\max}$ **do**
    Escolha um subgradiente $g_n \in \partial f(p_n)$
    if $\|g_n\| \leq \varepsilon$ **then**
        return p_n
    end if
    $y_n \leftarrow p_n - \gamma_n g_n$
    $p_{n+1} \leftarrow \Pi_K(y_n)$
    $n \leftarrow n + 1$
end while
return p_n

Propriedades do Algoritmo:

O algoritmo na verdade não é um algoritmo descendente, ou seja, não podemos garantir que $f(x_{i+1}) \leq f(x_i)$. Uma implicação é que o ponto na iteração final pode não ser o melhor ponto. As soluções para isso são:

1. Acompanhar o valor da função e fornecer a melhor iteração até o momento;
2. Fornecer o ponto de busca ponderado.

Exemplo 6. *Considere a função*

$$f(x) = |x|,$$

que é convexa, porém não diferenciável no ponto $x = 0$. Nesse caso o conjunto de subgradientes de f é dado por:

$$\partial f(x) = \begin{cases} \{1\}, & se\ x > 0, \\ \{-1\}, & se\ x < 0, \\ [-1, 1], & se\ x = 0. \end{cases}$$

O mínimo global da função ocorre em $x^\star = 0$.

Método do Subgradiente

O método do subgradiente gera uma sequência $\{x_k\}$ *pela atualização*

$$x_{k+1} = x_k - \eta_k g_k,$$

onde $g_k \in \partial f(x_k)$ *é um subgradiente escolhido no ponto atual e* $\eta_k > 0$ *é o passo (taxa de aprendizado).*

Parâmetros

- *Passo constante:* $\eta = 0.5$*;*
- *Número de iterações:* $K = 10$*;*
- *Ponto inicial:* $x_0 = 5$.

Seguindo o Algortimo 1 tem-se

Tabela 1: Iterações do método do subgradiente com $x_0 = 5$ e $\eta = 0.5$

k	x_k	$g_k = \text{sign}(x_k)$	$x_{k+1} = x_k - 0.5\, g_k$
0	5.0	1	4.5
1	4.5	1	4.0
2	4.0	1	3.5
3	3.5	1	3.0
4	3.0	1	2.5
5	2.5	1	2.0
6	2.0	1	1.5
7	1.5	1	1.0
8	1.0	1	0.5
9	0.5	1	0.0

Onde a expressão $\text{sign}(x_k)$ representa a função sinal, que, neste caso, indica o sinal do número x_k, isto é:

$$\text{sign}(x_k) = \begin{cases} 1 & \text{se } x_k > 0 \\ 0 & \text{se } x_k = 0 \\ -1 & \text{se } x_k < 0 \end{cases}$$

4 Gradiente descendente estocástico

Em vez de impor que a direção de atualização coincida exatamente com o gradiente, admite-se que ela seja um vetor aleatório cujo valor esperado, a cada iteração, corresponda ao gradiente. De forma mais geral, requer-se que o valor esperado desse vetor aleatório seja um subgradiente da função no ponto corrente (SHALEV-SHWARTZ; BEN-DAVID, 2014). Essa reformulação constitui um elemento central no aprimoramento do método do gradiente descendente, que passa então a ser denominado gradiente descendente estocástico[2]. A adoção dessa abordagem permite que, em cenários não convexos, o método apresente maior capacidade de escapar de mínimos locais — uma limitação típica do gradiente descendente determinístico. Nesse sentido, o gradiente descendente estocástico destaca-se como uma ferramenta particularmente relevante para problemas de otimização não convexa.

4.1 Otimização de funções limitadas convexas Lipschitzianas

Recorde o limite que obtivemos para o algoritmo de GD no Corolário 2.2. Para o caso estocástico, no qual apenas a esperança de $\mathbf{v}_t$ pertence a $\partial f\left(\mathbf{w}^{(t)}\right)$, não podemos aplicar diretamente a equação

$$f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*) \leq \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla f(\mathbf{w}^{(t)}) \rangle.$$

No entanto, como o valor esperado de $\mathbf{v}_t$ é um subgradiente de f em $\mathbf{w}^{(t)}$, ainda podemos derivar um limite semelhante para a saída esperada do gradiente descendente estocástico. Isso é formalizado no seguinte teorema.

TEOREMA 14.8 Sejam $B, \rho > 0$. Seja f uma função convexa e seja

$$\mathbf{w}^\star \in \operatorname{argmin}_{\mathbf{w}:\|\mathbf{w}\|\leq B} f(\mathbf{w}).$$

Suponha que o GDE seja executado por T iterações com

$$\eta = \sqrt{\frac{B^2}{\rho^2 T}}.$$

[2]Em inglês a terminologia empregada é Stochastic Gradient Descent (SGD).

Suponha também que, para todo t, vale $\|\mathbf{v}_t\| \leq \rho$ com probabilidade 1. Então,

$$\mathbb{E}[f(\overline{\mathbf{w}})] - f\left(\mathbf{w}^\star\right) \leq \frac{B\rho}{\sqrt{T}}.$$

Portanto, para qualquer $\epsilon > 0$, para obter

$$\mathbb{E}[f(\overline{\mathbf{w}})] - f\left(\mathbf{w}^\star\right) \leq \epsilon,$$

é suficiente executar o algoritmo GDE por um número de iterações que satisfaça

$$T \geq \frac{B^2\rho^2}{\epsilon^2}.$$

Prova. Vamos introduzir a notação $\mathbf{v}_{1:t}$ para denotar a sequência $\mathbf{v}_1, \ldots, \mathbf{v}_t$. Tomando a esperança da Equação $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta\nabla f(\mathbf{w}^{(t)})$, obtemos

$$\mathbb{E}_{\mathbf{v}_{1:T}}\left[f(\overline{\mathbf{w}}) - f\left(\mathbf{w}^\star\right)\right] \leq \mathbb{E}_{\mathbf{v}_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}\left(f\left(\mathbf{w}^{(t)}\right) - f\left(\mathbf{w}^\star\right)\right)\right].$$

Como o Lema 2.1 vale para qualquer sequência $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_T$, ele também se aplica ao GDE. Tomando a esperança do limite do lema, temos

$$\mathbb{E}_{\mathbf{v}_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}\left\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t\right\rangle\right] \leq \frac{B\rho}{\sqrt{T}}. \tag{4.1.1}$$

Resta mostrar que

$$\mathbb{E}_{\mathbf{v}_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}\left(f\left(\mathbf{w}^{(t)}\right) - f\left(\mathbf{w}^\star\right)\right)\right] \leq \mathbb{E}_{\mathbf{v}_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}\left\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t\right\rangle\right], \tag{4.1.2}$$

o que provaremos a seguir.

Usando a linearidade da esperança, temos

$$\mathbb{E}_{\mathbf{v}_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}\left\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t\right\rangle\right] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathbf{v}_{1:T}}\left[\left\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t\right\rangle\right].$$

Em seguida, recordamos a lei da esperança total: para quaisquer duas variáveis aleatórias α, β e uma função g, vale

$$\mathbb{E}_\alpha[g(\alpha)] = \mathbb{E}_\beta\,\mathbb{E}_\alpha[g(\alpha) \mid \beta].$$

Tomando $\alpha = \mathbf{v}_{1:t}$ e $\beta = \mathbf{v}_{1:t-1}$, obtemos

$$\begin{aligned}\mathbb{E}_{\mathbf{v}_{1:T}}\left[\left\langle \mathbf{w}^{(t)} - \mathbf{w}^{\star}, \mathbf{v}_t \right\rangle\right] &= \mathbb{E}_{\mathbf{v}_{1:t}}\left[\left\langle \mathbf{w}^{(t)} - \mathbf{w}^{\star}, \mathbf{v}_t \right\rangle\right] \\ &= \mathbb{E}_{\mathbf{v}_{1:t-1}}\mathbb{E}_{\mathbf{v}_{1:t}}\left[\left\langle \mathbf{w}^{(t)} - \mathbf{w}^{\star}, \mathbf{v}_t \right\rangle \mid \mathbf{v}_{1:t-1}\right].\end{aligned}$$

Uma vez que conhecemos $\mathbf{v}_{1:t-1}$, o valor de $\mathbf{w}^{(t)}$ deixa de ser aleatório e, portanto,

$$\mathbb{E}_{\mathbf{v}_{1:t-1}}\mathbb{E}_{\mathbf{v}_{1:t}}\left[\left\langle \mathbf{w}^{(t)} - \mathbf{w}^{\star}, \mathbf{v}_t \right\rangle \mid \mathbf{v}_{1:t-1}\right] = \mathbb{E}_{\mathbf{v}_{1:t-1}}\left\langle \mathbf{w}^{(t)} - \mathbf{w}^{\star}, \mathbb{E}_{\mathbf{v}_t}\left[\mathbf{v}_t \mid \mathbf{v}_{1:t-1}\right]\right\rangle.$$

Como $\mathbf{w}^{(t)}$ depende apenas de $\mathbf{v}_{1:t-1}$ e o GDE exige que

$$\mathbb{E}_{\mathbf{v}_t}\left[\mathbf{v}_t \mid \mathbf{w}^{(t)}\right] \in \partial f\left(\mathbf{w}^{(t)}\right),$$

obtemos que

$$\mathbb{E}_{\mathbf{v}_t}\left[\mathbf{v}_t \mid \mathbf{v}_{1:t-1}\right] \in \partial f\left(\mathbf{w}^{(t)}\right).$$

Assim,

$$\mathbb{E}_{\mathbf{v}_{1:t-1}}\left\langle \mathbf{w}^{(t)} - \mathbf{w}^{\star}, \mathbb{E}_{\mathbf{v}_t}\left[\mathbf{v}_t \mid \mathbf{v}_{1:t-1}\right]\right\rangle \geq \mathbb{E}_{\mathbf{v}_{1:t-1}}\left[f\left(\mathbf{w}^{(t)}\right) - f\left(\mathbf{w}^{\star}\right)\right].$$

No total, mostramos que

$$\begin{aligned}\mathbb{E}_{\mathbf{v}_{1:T}}\left[\left\langle \mathbf{w}^{(t)} - \mathbf{w}^{\star}, \mathbf{v}_t \right\rangle\right] &\geq \mathbb{E}_{\mathbf{v}_{1:t-1}}\left[f\left(\mathbf{w}^{(t)}\right) - f\left(\mathbf{w}^{\star}\right)\right] \\ &= \mathbb{E}_{\mathbf{v}_{1:T}}\left[f\left(\mathbf{w}^{(t)}\right) - f\left(\mathbf{w}^{\star}\right)\right].\end{aligned}$$

Somando em t, dividindo por T e usando a linearidade da esperança, obtemos que a Equação (4.1.2) é válida, o que conclui a prova.

Exemplo 7 (Minimização do Risco Empírico da regressão logística via GDE)**.** *Considere um problema de classificação binária modelado por regressão logística.*

- *Domain set:* $\mathcal{X} \subset \mathbb{R}^2$
- *Label set:* $\mathcal{Y} = \{0, 1\}$

Para $\boldsymbol{x} \in \mathcal{X}$*, definimos*

$$\mathbb{P}(Y = 1 \mid \boldsymbol{x}) = h_{\boldsymbol{\beta}}(\boldsymbol{x}) = \frac{1}{1 + \exp\left(-\langle \boldsymbol{\beta}, \boldsymbol{x}' \rangle\right)},$$

onde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ *e* $\boldsymbol{x}' = (1, x_1, x_2)$*. Dada a amostra de treinamento*

$$\mathbf{o}_n = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\},$$

a função de perda [3] *é*

$$\ell(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\Big[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\Big],$$

onde $\hat{y}_i = h_{\boldsymbol{\beta}}(\boldsymbol{x}_i)$*. O gradiente da perda é dado por*

$$\frac{\partial}{\partial \beta_k}\ell(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)x'_{i,k}, \qquad k = 0, 1, 2.$$

Definindo $\epsilon_i = \hat{y}_i - y_i$*, obtemos*

$$\nabla\ell(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i \boldsymbol{x}'_i.$$

No algoritmo GDE, a atualização dos parâmetros é

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta_t(\hat{y}_i - y_i)\boldsymbol{x}'_i.$$

Em termos de algortimos tem-se:

Algoritmo 2 - Mini-batch GDE para Regressão Logística

Ensure: Parâmetros estimados $\hat{\boldsymbol{\beta}}$

$\boldsymbol{\beta} \leftarrow \boldsymbol{\beta}_0 \in \mathbb{R}^3$
$t \leftarrow 0$
while $t < T$ **do**
    Embaralhe os dados
    for cada mini-batch B de tamanho m **do**
        Construa $X'_B \in \mathbb{R}^{m\times 3}$ e $\boldsymbol{y}_B \in \mathbb{R}^m$
        $\boldsymbol{z}_B \leftarrow X'_B\boldsymbol{\beta}$
        $\hat{\boldsymbol{y}}_B \leftarrow \sigma(\boldsymbol{z}_B)$
        $\boldsymbol{\epsilon}_B \leftarrow \hat{\boldsymbol{y}}_B - \boldsymbol{y}_B$
        $\boldsymbol{g} \leftarrow \frac{1}{m}X'^{\top}_B\boldsymbol{\epsilon}_B$
        $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} - \eta\boldsymbol{g}$
    end for
    $t \leftarrow t + 1$
end while
return $\boldsymbol{\beta}$

[3] Nesse caso a Minimização do Risco Empírico (ERM) e Máxima Verossimilhança (MLE) são equivalentes.

5 Aplicações

Nesta fase do trabalho, será aplicado o algoritmo de gradiente descendente estocástico no contexto estatístico. Inicialmente, será abordada a aplicação do algoritmo na estimação de parâmetros de modelos de misturas de normais. Em seguida, seu grau de eficiência será avaliado com base em métricas e estatísticas apropriadas, como o viés e a raíz do erro quadrático médio[4].

5.1 Análise descritiva dos dados

O conjunto de dados Old Faithful descreve o comportamento eruptivo do gêiser Old Faithful, localizado no Yellowstone National Park, no estado de Wyoming, nos Estados Unidos. O objetivo do conjunto é registrar e analisar a relação entre a duração das erupções e o tempo de espera até a ocorrência da próxima erupção.

Os dados estão organizados em um data frame contendo 272 observações e duas variáveis numéricas. A primeira variável, denominada eruptions, corresponde ao tempo de duração de cada erupção, medido em minutos. A segunda variável, waiting, representa o tempo de espera, também em minutos, até a próxima erupção do gêiser.

Existem diversas versões desse conjunto de dados, sendo que algumas apresentam informações mais completas. Estudos posteriores utilizaram variações aprimoradas dos registros originais, contribuindo para análises estatísticas mais refinadas sobre o comportamento do gêiser Old Faithful.

A seguir é apresentada a Tabela 2 com estatísticas resumo do conjunto de dados em estudo.

Tabela 2: Estatísticas descritivas do conjunto de dados

Variável	**Media**	**Mediana**	**DP**	**CV**	**Mínimo**	**Maximo**	**Assimetria**	**Curtose**
Eruptions	3,49	4,00	1,14	0,33	1,60	5,10	-0,42	1,50
Waiting	70,90	76,00	13,59	0,19	43,00	96,00	-0,42	1,86

A Tabela 2 apresenta as estatísticas descritivas das variáveis Eruptions e Waiting, permitindo uma visão geral sobre o comportamento dos dados. Para a variável Eruptions, observa-se que a média 3,49 é ligeiramente inferior à mediana 4,00, o que indica uma distribuição levemente assimétrica à esquerda, confirmada pelo valor negativo da assimetria -0,42. O desvio padrão de 1,14 e o coeficiente de variação de 0,33 sugerem uma dispersão

[4]O código pode ser acessado por meio da página: Código Fonte.

moderada em relação à média. Os valores mínimo 1,60 e máximo 5,10 evidenciam a amplitude dos dados, enquanto a curtose de 1,50 indica uma distribuição mais achatada quando comparada à normal.

De forma semelhante, a variável Waiting apresenta média de 70,90 e mediana de 76,00, também apontando para uma assimetria negativa. O desvio padrão é de 13,59, com coeficiente de variação de 0,19, indicando menor variabilidade relativa em comparação à variável Eruptions. Os valores mínimo de 43,00 e máximo de 96,00 mostram uma amplitude considerável, e a curtose de 1,86 sugere uma distribuição relativamente achatada. Em conjunto, essas estatísticas permitem compreender a tendência central, a dispersão e o formato das distribuições das duas variáveis analisadas. Por fim, é conveniente analisar a distribuição dos dados, apresentada na Figura 4.

Observa-se de imediato a presença de multimodalidade, característica que evidencia que os dados não são gerados por um único processo homogêneo, mas sim por dois comportamentos ou regimes distintos subjacentes. Como consequência, medidas descritivas globais, como a média e o desvio-padrão, tornam-se menos representativas do conjunto como um todo, uma vez que tendem a mascarar a estrutura interna da distribuição. Dessa forma, a multimodalidade observada reforça a necessidade de uma análise segmentada dos dados, reconhecendo a coexistência de múltiplos padrões no fenômeno estudado.

Ademais, a ausência de valores extremos isolados sugere consistência interna dos dados dentro de cada agrupamento identificado. Entretanto, a presença de dois padrões distintos inviabiliza a suposição de uma distribuição normal simples, tornando mais apropriada a interpretação dos dados por meio de uma distribuição bimodal ou de uma abordagem baseada em modelos de mistura de distribuições normais, (DEISENROTH; FAISAL; ONG, 2020).

Neste caso, buscar-se-á modelar os dados por meio de uma mistura de distribuições normais, utilizando o algoritmo de gradiente descendente. Para isso, faz-se inicialmente necessário verificar a validade da hipótese de que os dados — em particular, os tempos de espera, que constituir-se-ão como o foco deste estudo — apresentam uma distribuição bimodal. Nesse sentido pode-se empregar o teste Teste de Dip de Hartigan (HARTIGAN; HARTIGAN, 1985). Esse teste mede o chamado *Dip*, que é a maior diferença entre a função de distribuição empírica dos dados e a função de distribuição unimodal que mais se aproxima dela. Quanto mais elevado o *Dip*, maior a evidência contra a unimodalidade. Dessa forma, as hipóteses do teste são

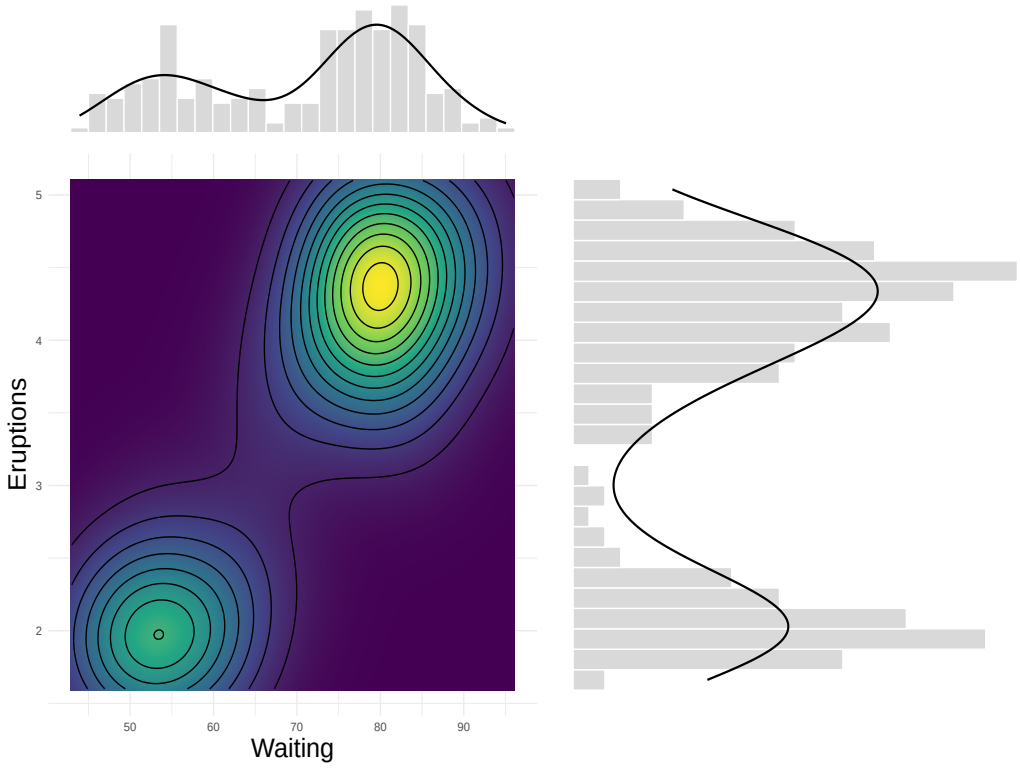


Figura 4: Distribuições conjunta e marginais

$$\begin{cases} H_0: & \text{os dados são unimodais,} \\ H_1: & \text{os dados não são unimodais.} \end{cases}$$

Nesse teste um p-valor pequeno implica na evidência de multimodalidade. Para o caso os resultados foram:

Tabela 3: Resultado do teste de Hartigan's Dip

Medida	**Valor**
Dip statistic	0.0414
p-value	0.00181

Assumindo um nível de significância de 5% tem-se que é possível rejeitar a hipótese nula de unimodalidade. Dessa forma, é possível dar seguimento ao processo de modelagem por meio de misturas de distriuições normais. Essa abordagem é interesante visto que modelos baseados em misturas de distribuições são uma classe de modelos probabilísticos usados para representar populações heterogêneas, nas quais os dados observados podem ser gerados por diferentes subpopulações latentes, (SUGIYAMA, 2015). Em vez de assumir que todas as observações provêm de uma única distribuição, esses modelos postulam que cada observação é originada de um entre vários componentes probabilísticos, cada

um descrito por sua própria distribuição (TITTERINGTON; SMITH; MAKOV, 1985).

5.2 Modelagem dos dados via misturas de normais

A metodologia aqui empregada configura-se como uma estimação direta de um modelo de mistura Gaussiana unidimensional por otimização estocástica da log-verossimilhança, com parametrizações que garantem restrições naturais, inicialização por k-means, tratamento do problema de label switching por permutações e diagnóstico por resíduos padronizados e comparação gráfica com uma referência, (SUGIYAMA, 2015). O vetor de parâmetros é organizado em três blocos: os logits das proporções $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$[5], as médias $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ e os log-desvios $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_K)$, sendo a transformação inversa $\sigma_j = \exp(\ell_j)$ usada para garantir $\sigma_j > 0$. As proporções misturadas são obtidas pela função softmax, numericamente estabilizada, isto é,

$$\pi_j = \frac{\exp(\alpha_j)}{\sum_{l=1}^{K} \exp(\alpha_l)} \qquad (j = 1, \ldots, K),$$

onde, subtrai-se o máximo de α antes de exponenciar para evitar overflow.

A função objetivo maximizada é a log-verossimilhança do modelo de mistura sobre a amostra $x_1, \ldots, x_n$:

$$\ell(\theta) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{K} \pi_j, \varphi\big(x_i \mid \mu_j, \sigma_j^2\big) \right),$$

onde $\varphi(\cdot \mid \mu, \sigma^2)$ é a densidade normal. Para facilitar a otimização por gradiente estocástico, calcula-se a contribuição de um ponto x à derivada da log-verossimilhança. Define-se primeiro as densidades ponderadas por componente e a probabilidade marginal do ponto:

$$d_j(x) = \pi_j, \varphi(x \mid \mu_j, \sigma_j^2), \qquad p(x) = \sum_{j=1}^{K} d_j(x).$$

Os pesos da posterior probabilities de componente dado x são então

$$r_j(x) = \frac{d_j(x)}{p(x)}.$$

A partir disso, obtêm-se expressões fechadas para o gradiente em relação aos parâmetros

[5]Um vetor livre de parâmetros reais que depois é mapeado para proporções pela softmax.

transformados. Em particular, a derivada em relação aos logits α resulta em

$$\frac{\partial \log p(x)}{\partial \alpha_j} = r_j(x) - \pi_j,$$

resultado que decorre da regra da cadeia aplicada à softmax usando $\partial \pi_k / \partial \alpha_j = \pi_k(\delta_{jk} - \pi_j)$. As derivadas em relação às médias e aos log-desvios seguem das derivadas usuais da densidade normal combinadas com o peso r_j:

$$\frac{\partial \log p(x)}{\partial \mu_j} = r_j(x), \frac{x - \mu_j}{\sigma_j^2}, \qquad \frac{\partial \log p(x)}{\partial \ell_j} = r_j(x) \left(\frac{(x - \mu_j)^2}{\sigma_j^2} - 1 \right),$$

onde a derivada em relação a ℓ_j incorpora a cadeia $\partial \sigma_j / \partial \ell_j = \sigma_j$.

A otimização é feita pelo GDE em mini-batchs[6], (GOODFELLOW; BENGIO; COURVILLE, 2016). A cada iteração é extraída uma amostra mini-batch da amostra total, calculam-se as contribuições de gradiente por ponto e toma-se a média para obter o gradiente estimado do mini-lote. O passo de atualização é da forma

$$\theta_{t+1} = \theta_t + \eta_t, \overline{\nabla \ell},$$

onde $\eta_t = \frac{\eta_0}{(1+\gamma t)}$ é uma taxa de aprendizado com decaimento e $\overline{\nabla \ell}$ é o gradiente médio no mini-batch. Observa-se que a rotina implementa a soma do gradiente, pois-se se busca maximizar a log-verossimilhança. Para monitoramento, a log-verossimilhança total sobre os n pontos é avaliada e registrada a cada atualização, garantindo traços temporais e trajetória dos parâmetros para diagnóstico e visualização de convergência.

A inicialização dos parâmetros é feita tomando as médias como fixadas nos centros obtidos por k-means sobre os dados, os parâmetros α são inicializados iguais (vetor nulo, implicando π_j iniciais iguais) e os log-desvios são todos iguais a $\log(\mathrm{sd}(x)/2)$, uma escolha razoável de escala. Esta escolha é conveniente em muitos problemas empíricos, pois k-means oferece uma partição inicial que costuma estar próxima das componentes reais quando há heterogeneidade multimodal nos dados (MCLACHLAN; PEEL, 2000).

Calculam-se métricas de desempenho como viés (Bias) e RMSE elementares para cada tipo de parâmetro:

$$\mathrm{Bias} = \hat{\theta} - \theta^{\mathrm{ref}}, \qquad \mathrm{RMSE} = \sqrt{\frac{1}{K} \sum j = 1^K (\hat{\theta}_j - \theta_j^{\mathrm{ref}})^2}.$$

[6]O GDE nesse caso, tem os parâmetros do modelo atualizados usando o gradiente calculado a partir de pequenos subconjuntos aleatórios dos dados, chamados mini-batchs, em vez de toda a amostra ou de apenas uma observação.

Para avaliação diagnóstica do ajuste, a metodologia inclui duas vertentes: comparação gráfica das densidades e análise de resíduos padronizados. A comparação gráfica sobrepõe ao histograma empírico a densidade total ajustada pela mistura estimada e a densidade de referência k-means, permitindo visualizar aproximações e desajustes. Os resíduos são construídos atribuindo a cada observação ao componente de maior responsabilidade $\arg\max_j r_j(x_i)$ e calculando o resíduo padronizado

$$e_i = \frac{x_i - \hat{\mu}_{a_i}}{\hat{\sigma}_{a_i}}, \qquad a_i = \arg\max_j r_j(x_i).$$

A série de resíduos é então examinada por um QQ-plot com envelope de simulação.

5.3 Resultados

A análise dos resultados indica que a estimação realizada por meio de otimização estocástica, utilizando como ponto inicial os centróides obtidos pelo algoritmo k-means, produziu soluções estáveis e satisfatórias para os parâmetros do modelo. Essa estratégia de inicialização mostrou-se adequada para orientar o processo de otimização em direção a regiões plausíveis do espaço paramétrico, reduzindo potenciais problemas associados à sensibilidade a valores iniciais.

A Tabela 4 apresenta as estimativas dos parâmetros obtidas por meio do procedimento de otimização via GDE. Observa-se que as proporções estimadas indicam predominância da primeira componente, com peso aproximado de 0.64, enquanto a segunda componente apresenta peso em torno de 0.36. As médias estimadas evidenciam uma clara separação entre as componentes, situando-se aproximadamente em 80.25 e 54.76, respectivamente. Já os desvios padrão estimados são bastante semelhantes entre as duas componentes, ambos próximos de 5.9, sugerindo variabilidade interna comparável entre os grupos.

Tabela 4: Parâmetros estimados do modelo via GDE

Parâmetro	Componente 1	Componente 2
$\hat{\boldsymbol{\pi}}$	0.6372	0.3628
$\hat{\boldsymbol{\mu}}$	80.2538	54.7576
$\hat{\boldsymbol{\sigma}}$	5.8418	5.9018

A Tabela 5 apresenta as métricas de desempenho associadas às estimativas, considerando o viés e a raiz do erro quadrático médio (RMSE). Observa-se que os vieses das estimativas são praticamente nulos, indicando ausência de distorção sistemática relevante no processo de estimação. Em particular, o viés associado à proporção π_2 é extremamente pequeno,

enquanto os vieses dos parâmetros de localização e escala também permanecem próximos de zero.

Em relação à precisão das estimativas, os valores de RMSE para os parâmetros de média encontram-se na ordem de 10^{-2}, o que evidencia elevada precisão na recuperação das localizações das componentes. Por outro lado, os parâmetros de escala apresentam RMSE na ordem de 10^{-1}, refletindo maior variabilidade relativa nessas estimativas.

De forma geral, os resultados indicam que a maior parcela da incerteza do modelo está concentrada nos parâmetros de escala, enquanto as estimativas das médias e das proporções são bastante estáveis e precisas.

Tabela 5: Métricas estimadas do modelo via GDE

Parâmetro	Viés	RMSE
$\hat{\boldsymbol{\mu}}$	0.0076	0.0226
$\hat{\boldsymbol{\sigma}}$	0.0065	0.1517
$\hat{\boldsymbol{\pi}}$	-0.0048	0.0048

A estabilidade dos estimadores entre a inicialização por k-means e a solução final por GDE sugere que o problema de otimização é bem comportado para esta amostra: a rotina de inicialização leva o algoritmo a uma vizinhança do ótimo e o refinamento por gradiente estocástico ajusta apenas pequenas discrepâncias, sobretudo nas variâncias.

A validação gráfica corrobora esse diagnóstico numérico. A sobreposição da densidade ajustada pela mistura estimada ao histograma empírico mostra excelente concordância, com a mistura estimada capturando corretamente os modos e a largura das cristas observadas no histograma; esta concordância visual aparece na Figura 5.

O QQ-plot, que pode ser observado na Figura 6 dos resíduos padronizados, construídos atribuindo cada observação ao componente de maior responsabilidade e padronizando pelo respectivo estimador, apresenta pontos próximos à reta teórica de uma normal padrão, com desvios modestos nas caudas que não chegam a comprometer a adequação do modelo — isto é, os resíduos não revelam discrepâncias sistemáticas graves nem estrutura de não-normalidade pronunciada.

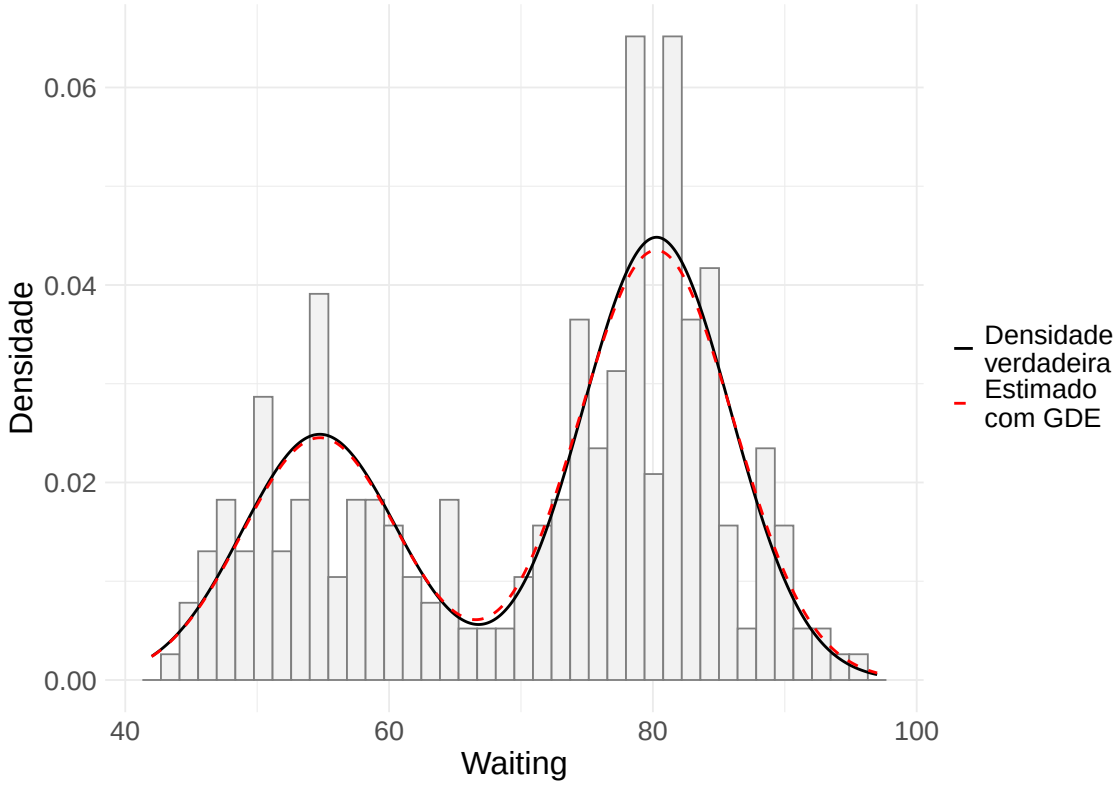


Figura 5: Gráfico de ajuste do modelo

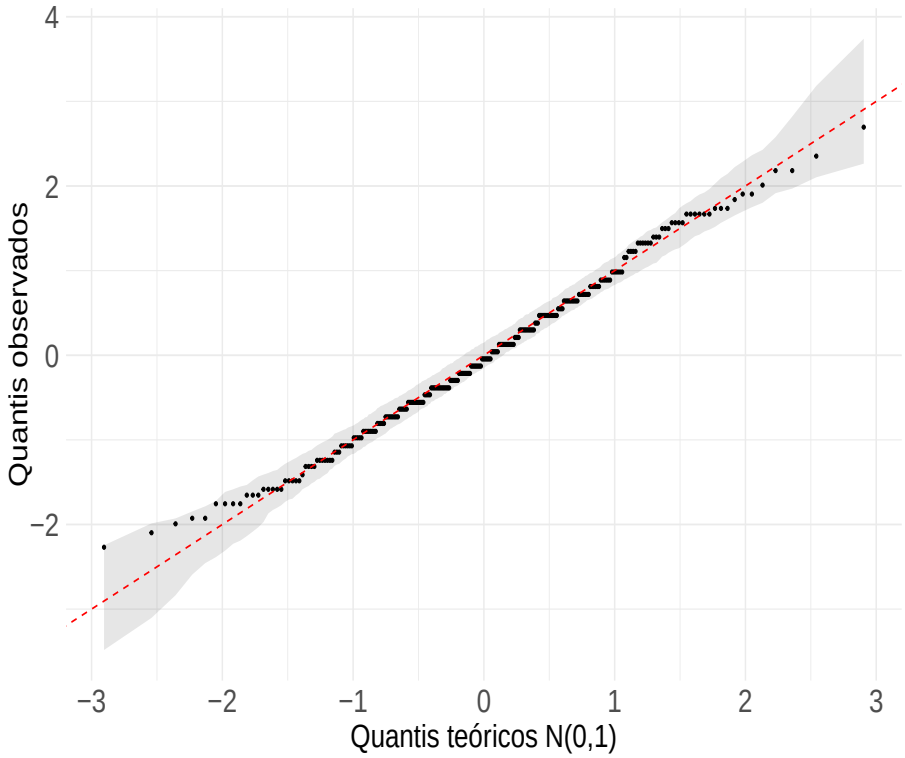


Figura 6: Ajuste dos resíduos dos modelos

6 Conclusão

Com base no estudo do gradiente descendente, foi possível observar que esse algoritmo possui uma série de atributos que o tornam bastante versátil. Ele pode ser utilizado como método de otimização tanto em contextos de minimização, amplamente empregados em machine learning, quanto em problemas de otimização estocástica, como analisado no presente trabalho.

Outro aspecto que reforça essa perspectiva é a possibilidade de aplicar o algoritmo até mesmo em funções não diferenciáveis, ainda assim obtendo resultados consistentes do ponto de vista analítico e matemático. Além disso, há flexibilidade na definição da taxa de aprendizagem, que pode ser ajustada de acordo com as propriedades da função considerada; por exemplo, quando a função é contínua e satisfaz condições do tipo Lipschitz.

Ademais, seu emprego em contextos estocásticos, nos quais a direção do vetor gradiente pode variar, constitui um recurso poderoso especialmente no caso de funções não convexas, nas quais também podem ser implementadas novas adaptações relacionadas à taxa de aprendizagem.

Diante dessas características, torna-se bastante conveniente o uso do Gradiente Descendente Estocástico (GDE) no contexto estatístico, no qual problemas de otimização são recorrentes, como ilustrado na aplicação relacionada à otimização da função de máxima verossimilhança. Nesse cenário, o GDE apresenta-se como uma alternativa eficiente aos métodos tradicionais de otimização numérica, tais como os métodos do tipo quasi-Newton, especialmente em situações que envolvem não continuidade ou estruturas de modelo mais complexas.

Referências

BORWEIN, J. J.; LEWIS, A. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. [S.l.: s.n.], 2005. ISBN 978-0-387-29570-1.

BRINK, H.; RICHARDS, J.; TETHEROLF, M. *Real-World Machine Learning*. [S.l.]: Manning Publications, 2016.

DEISENROTH, M. P.; FAISAL, A. A.; ONG, C. S. *Mathematics for Machine Learning*. Cambridge: Cambridge University Press, 2020. ISBN 9781108470049.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. Cambridge, Massachusetts: MIT Press, 2016. ISBN 9780262035613. Disponível em: ⟨https://www.deeplearningbook.org⟩.

HARTIGAN, J. A.; HARTIGAN, P. M. The dip test of unimodality. *The Annals of Statistics*, v. 13, n. 1, p. 70–84, 1985.

HOFFMAN, K.; KUNZE, R. *Linear Algebra*. 2. ed. Upper Saddle River: Prentice Hall, 1971.

LEI, Y. et al. Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. 2019.

LIMA, E. L. *Análise no Espaço* $\mathbb{R}^n$. 3. ed. Rio de Janeiro: IMPA, 2025. 120 p. Coleção Matemática Universitária. ISBN 978-85-244-0581-5.

MCLACHLAN, G. J.; PEEL, D. *Finite mixture models*. New York: Wiley, 2000.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. *Foundations of Machine Learning*. [S.l.]: MIT Press, 2018.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. *Understanding Machine Learning: From Theory to Algorithms*. [S.l.]: Cambridge University Press, 2014.

SUGIYAMA, M. *Introduction to Statistical Machine Learning*. San Francisco, CA, United States: Morgan Kaufmann, 2015. ISBN 978-0-12-802350-1.

TITTERINGTON, D. M.; SMITH, A. F. M.; MAKOV, U. E. *Statistical Analysis of Finite Mixture Distributions*. Chichester; New York: John Wiley & Sons, 1985.