Accelerated Gradient Descent através do espelho ou como acelerar o Mirror Descent

José Arthur de Souza

20 de março de 2026

Dedico esta dissertação à minha família, que sempre me ofereceu apoio quando procurei.

Agradecimentos

Agradeço à minha família, em especial aos meus pais, Ivanor e Neli, por terem me criado sempre observando o potencial transformador da educação e por insistirem que ela fosse central em minha trajetória.

Agradeço ao meu orientador, Bernardo, pela disposição em me ensinar e pelas paciência e compreensão quando tive dificuldades em aprender. Agradeço-o por ter acreditado em mim quando eu mesmo duvidei e por ter me ajudado a persistir. Agradeço à professora Asla por ter possibilitado nossa conexão.

Aproveito para agradecer aos professores da EMAp, que tanto me ensinaram ao longo desses anos e, mais uma vez, à Asla, por ter me ensinado, além de matemática, sobre a vida. Agradeço também a todos os funcionários da FGV, pelos serviços prestados sempre com excelência.

Agradeço aos meus amigos e namorado, por sempre estarem lá por mim nos momentos difíceis e por fazerem meus dias mais completos e felizes. Agradeço-os por terem ampliado minha visão de mundo e proporcionado novas perspectivas sobre tantas coisas.

Por fim, agradeço à FGV e à CAPES, pelas bolsas que viabilizaram a conclusão do meu mestrado.

Resumo

Nesta dissertação, consideramos três métodos de primeira ordem com passo fixo para a resolução aproximada de problemas de otimização convexa. São eles: o método do gradiente descendente, o método de Nesterov (gradiente descendente acelerado) e o método de Bregman, ou *Mirror Descent*. Propomos ainda um quarto método, o *Accelerated Mirror Descent* (método de Bregman acelerado), concebido como uma versão acelerada (inspirada no método de Nesterov) do *Mirror Descent*, analisando duas possibilidades de aceleração: uma no espaço primal e outra no espaço dual. A formulação apresentada difere da versão do *Accelerated Mirror Descent* introduzida por [KBB15]. Além disso, apresentamos as equações diferenciais ordinárias (EDOs) associadas a cada método, cujas discretizaçõesdas das soluções correspondem às trajetórias dos algoritmos, incluindo a EDO do *Accelerated Mirror Descent*. Para esta última, demonstramos a existência e unicidade de solução dada uma condição inicial, bem como a convergência dos pontos gerados pelo método em direção à trajetória da solução, o que fornece uma prova mais detalhada também para resultados análogos ao método de Nesterov acelerado. Por fim, realizamos um experimento numérico para avaliar os efeitos das acelerações e da implementação do paradigma *mirror* na resolução de um problema de mínimos quadrados restrito ao simplex, em dimensões baixa (100) e alta (10.000). Os resultados indicam que ambas as estratégias melhoram substancialmente a convergência em alta dimensão.

Palavras-chave: Otimização Convexa, Métodos de primeira ordem, Aceleração de Nesterov, Divergência de Bregman, *Mirror Descent*.

Abstract

In this thesis, we consider three first-order methods with a fixed step size for the approximate resolution of convex optimization problems: the gradient descent method, Nesterov's method (or accelerated gradient descent), and Mirror Descent. We also propose a fourth method, the Accelerated Mirror Descent, conceived as an accelerated version (inspired by Nesterov's acceleration) of the Mirror Descent; this was done by analyzing two possibilities for acceleration, one in the primal space and another in the dual space. The proposed method differs from the Accelerated Mirror Descent version presented by [KBB15]. We also present the ordinary differential equations (ODEs) associated with each method, i.e., the ODEs whose solution discretizations correspond to the trajectories of the methods, including the one associated with Accelerated Mirror Descent. For the latter, we also prove the existence and uniqueness of the solution given an initial condition and the convergence of the points generated by the method towards the solution trajectory; this provides a more detailed demonstration of these same results relative to Nesterov's accelerated method. Finally, we conducted a numerical experiment designed to evaluate the effects of the accelerations, as well as the implementation of the mirror paradigm, in solving the least squares problem constrained to the simplex in low (100) and high (10,000) dimensions, and concluded that both substantially improve convergence in high dimension.

Keywords: Convex Optimization, First Order Methods, Nesterov's Acceleration, Bregman Divergence, Mirror Descent.

Sumário

1 Introdução **9**

2 Definições e resultados gerais **11**
2.1 Norma, Espaço Dual e Norma Dual 11
2.2 Continuidade e Diferenciabilidade 13
2.3 Convexidade 14
2.4 Conjuntos de Nível 16
2.5 Funções Lipschitz 17

3 Gradient Descent **20**
3.1 Definição e interpretações 20
3.2 Gradient Descent com Projeção para Conjuntos Convexos 21
3.3 EDO 23

4 Accelerated Gradient Descent **25**
4.1 Definição e interpretações 25
4.2 Accelerated Gradient Descent com Projeção 26
4.3 EDO e derivação informal 27

5 Mirror Descent **31**
5.1 Preliminares para Mirror Descent 32
5.1.1 Divergência de Bregman 32
5.1.2 Projeção de Bregman 34
5.2 Definição e interpretações 35
5.2.1 Mirror Map 36
5.3 Convergência do *Mirror Descent* 39
5.4 EDO 39
5.5 Aplicação do MD em problemas de otimização no Simplexo 40

6 Accelerated Mirror Descent **43**
6.1 Duas propostas de aceleração 43
6.2 Derivação informal da EDO 46
6.3 Aplicação do AMD em problemas de otimização no Simplexo 47

7 Uma EDO para o AMD **50**
7.1 Existência e Unicidade 50
7.1.1 Existência 50
7.1.2 Unicidade 59
7.2 Convergência do AMD 61
7.2.1 Caso 1: onde há suavização 68
7.2.2 Caso 2: onde não há suavização 69
7.2.3 Parte final 70

8 Experimento numérico **74**

9 Conclusão **78**

A Mirror Descent **83**

B EDO **85**
B.1 Teoria das EDOs 85
B.2 Convergência do método 85
B.2.1 Sobre a limitação da função avaliada nos pontos auxiliares primais 85
B.2.2 Um pouco mais sobre funções Lipschitz 88
B.2.3 Dois fatos auxiliares 90
B.2.4 Funções não descrescentes 91
B.2.5 Recorrência 91
B.2.6 Limites 93
B.3 Sobre a discretização da EDO associada ao AMD 94

C Experimento Numérico **95**

D Taxas de convergência empíricas **98**

Capítulo 1

Introdução

Neste documento consideramos problemas de otimização da forma

$$(P) = \begin{cases} \min f(x) \\ x \in \mathcal{X} \subseteq \mathbb{R}^n \end{cases}, \tag{P}$$

sendo $f : \mathbb{R}^n \to \mathbb{R}$ uma função diferenciável, convexa com gradiente ∇f L-Lipschitz e $\mathcal{X}$ um conjunto convexo.

Em otimização convexa, especialmente em contextos de larga escala como em Aprendizado de Máquina, métodos de primeira ordem, isto é, métodos que usam apenas informações sobre a primeira derivada da função objetivo, são fundamentais para a resolução aproximada de tais problemas. Essa relevância deve-se ao baixo custo associado a eles (sendo o cálculo de uma única derivada tipicamente a operação mais custosa computacionalmente) e à possibilidade de paralelização.

Neste trabalho estudamos três métodos de primeira ordem. O primeiro é possivelmente o método de primeira ordem mais simples conhecido: o método do gradiente descendente ou *Gradient Descent* (GD). O GD executa a cada iteração um passo na direção oposta a do gradiente, que representa a direção de maior descida no valor de f; esse passo é chamado de passo do gradiente. O segundo é o método do gradiente acelerado de Nesterov ou *Accelerated Gradient Descent* (AGD) ([Nes83]). O AGD opera da mesma forma que o GD, mas com a introdução de um passo a mais a cada iteração: um passo de momento, usado para acelerar a convergência e atingir a taxa ótima de convergência $\mathcal{O}(1/k^2)$, onde k é o número de iterações realizadas; este método tem aplicações em diversos problemas, como pode ser visto em [BT09], [Nes05] e [Nes13]. O terceiro é o método de descida do espelho ou *Mirror Descent* (MD), de Nemirovski e Yudin ([NY83b]), que representa uma importante generalização do GD para geometrias não euclidianas, como é explicado em [NY83b] e [BT03]. Ele executa o mesmo passo do gradiente do GD, mas no espaço dual, o espaço dos funcionais lineares contínuos sobre elementos de $\mathbb{R}^n$ (ao qual o gradiente de fato pertence), e com variáveis duais obtidas através de um difeomorfismo que chamamos de mapeamento espelho ou *mirror map*. Este método também tem diversas aplicações em otimização convexa ([BTN01], [BTMN01], [JNT11]) e otimização **online** ([BCB$^+$12], [CBL06]).

Propomos ainda um quarto e novo método, o Accelerated Mirror Descent (AMD), que implementa uma aceleração para o MD inspirada na aceleração de Nesterov para o GD, unindo vantagens das duas abordagens. Vale a pena ressaltar que existem outros métodos como o AMD, que visam acelerar o MD ([KBB15], [WWJ16]).

Vários métodos de otimização podem ser vistos como discretizações de soluções de equações diferenciais ordinárias (EDOs); vamos formalizar isso mais adiante, mas a ideia é que um método é a discretização de uma EDO quando a trajetória definida pelos pontos gerados por ele tende a uma solução da EDO conforme refinamos seus passos. Essa ligação nos permite aproveitar teoria das EDOs para analisar esses métodos (quanto às suas convergência, estabilidade ou interpretação) ou mesmo desenvolver outros algoritmos de otimização (a partir de discretizações diferentes, por exemplo) e já foi usada para resolver aproximadamente problemas de otimização com restrições ([Blo94], [BBB89], [HM12]) e sem restrições ([SS00]). Neste trabalho, explicitamos as EDOs associadas aos métodos GD, AGD e MD. Também derivamos a EDO da qual a discretização da solução deve corresponder ao AMD proposto e provamos que 1) esta EDO tem uma única

solução e 2) o AMD de fato converge para esta solução. Este é, na verdade, o objetivo central deste trabalho: desenvolver um método que complete o conjunto de algoritmos obtidos pela combinação ou não de dinâmicas de aceleração e de mapeamento espelho, alcançando assim uma aceleração do MD ou, de forma equivalente, uma generalização do AGD. Toda a construção é guiada por uma perspectiva unificadora, que integra de maneira sistemática a teoria das EDOs à formulação e análise de algoritmos de otimização.

O Capítulo 2 apresenta definições e resultados usados ao longo de todo o documento. Os métodos GD, AGD, MD e AMD são detalhados nos capítulos 3, 4, 5 e 6 respectivamente. Para cada método, apresentamos primeiro a sua versão para problemas sem restrição (isto é, para $\mathcal{X} = \mathbb{R}^n$) e depois sua versão projetada, usada para resolver o problema com restrição. As EDOs associadas são incluídas em seções específicas em cada um destes capítulos. As demonstrações relativas à EDO associada ao AMD se encontram no Capítulo 7. O Capítulo 8 trata de um experimento numérico conduzido para comparar as convergências dos quatro métodos ao resolver o problema de minimização de mínimos quadrados no simplexo em dois espaços de dimensões diferentes.

Todas as imagens foram produzidas pelo autor usando a biblioteca Matplotlib [Pyt23] para a linguagem de programação Python [Hun07] (os código está disponível em `https://github.com/josearthursouza/Dissertacao`) e o site Mathcha [Mat25] (para os diagramas em Tikz/LaTeX).

Capítulo 2

Definições e resultados gerais

Neste capítulo apresentamos definições e resultados chave para o desenvolvimento deste trabalho. Trazemos também um vislumbre de como cada objeto e resultado aparecerá nos próximos capítulos e a importância deles nesta dissertação.

2.1 Norma, Espaço Dual e Norma Dual

Uma das maiores motivações para a definição e uso do *Mirror Descent* está relacionada ao uso de diferentes normas em problemas específicos e o espaço e norma duais associados a elas. A seguir, exibimos as definições de norma, espaço dual e norma dual, bem como definições auxiliares, resultados relacionados e os exemplos mais comuns destes.

Definição (Norma)**.** *Seja X um espaço vetorial sobre $\mathbb{R}$. Uma* ***norma*** *em X é uma função $\|\cdot\| : X \to \mathbb{R}$ satisfazendo*

1. $\|x\| \geq 0$ *para todo* $x \in X$ *e* $\|x\| = 0 \Leftrightarrow x = 0$*;*
2. $\|\alpha x\| = |\alpha| \, \|x\|$ *para todo escalar* α*; e*
3. $\|x + y\| \leq \|x\| + \|y\|$ *para todos* $x, y \in X$*.*

Um espaço vetorial equipado com uma norma é chamado ***espaço normado****.*

A norma captura a noção de "tamanho" de um vetor e diferentes normas podem ser mais ou menos adequadas em diferentes contextos.

Exemplo. *Para $1 \leq p < \infty$, as normas ℓ_p são definidas por*

$$\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

Alguns casos particulares notáveis são

- *ℓ_2, a norma euclidiana, definida por $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$; e*
- *ℓ_1, a norma da soma, definda por $\|x\|_1 = \sum_{i=1}^{n} |x_i|$.*

Exemplo. *A norma ℓ_∞ é definida por*

$$\|x\|_\infty := \max_{1 \leq i \leq n} |x_i|.$$

Ela representa o limite da norma ℓ_p quando $p \to \infty$ e é também chamada de norma do máximo.

No caso de otimização sobre o simplexo, em problemas cujas variáveis são distribuições de probabilidade por exemplo, a norma ℓ_1 pode ser mais natural do que a norma euclidiana, a qual estamos mais habituados.

O fato a seguir garante que para qualquer norma, uma simples multiplicação serve de normalização.

Fato 2.1. *Seja $\|\cdot\|$ uma norma definida sobre $\mathcal{X}$, um espaço vetorial qualquer. Então para todo $x \in \mathcal{X}$, temos*

$$\left\| \frac{x}{\|x\|} \right\| = 1.$$

Demonstração. Pela propriedade *2.* da definição de norma, temos

$$\left\| \frac{x}{\|x\|} \right\| = \left| \frac{1}{\|x\|} \right| \|x\| = \frac{\|x\|}{\|x\|} = 1.$$

■

Além da norma, o conceito de espaço dual — e consequentemente o de norma dual também — será de suma importância para a própria definição do MD. As próximas duas definições introduzem formalmente esse conceito.

Definição (funcional linear)**.** *Seja X um espaço vetorial. Uma aplicação $T : X \to \mathbb{R}$ é chamada de* ***funcional linear*** *se para todos $x, y \in X$ e $\alpha \in \mathbb{R}$ satisfaz*

1. *$T(x + y) = T(x) + T(y)$; e*
2. *$T(\alpha x) = \alpha T(x)$.*

Definição (Espaço dual)**.** *O* ***espaço dual*** *X^* de um espaço normado X é o conjunto de todos os funcionais lineares contínuos $\varphi : X \to \mathbb{R}$.*

Ao introduzirmos o *Mirror Descent*, veremos que o gradiente de uma função f em um ponto $x \in X$ não é (pelo menos, não em geral) um elemento de X, mas sim do espaço dual X^*.

Exemplo. *Em um espaço de Hilbert H, isto é, um espaço vetorial dotado de produto interno $\langle \cdot, \cdot \rangle$, o teorema de Riesz estabelece que H^* é isometricamente isomorfo a H através do mapeamento $y \mapsto \langle y, \cdot \rangle$.*

O teorema de Riesz nos garante que ao realizar um passo do gradiente no *Gradient Descent* num espaço de Hilbert, podemos somar tranquilamente um vetor do espaço dual a um vetor do espaço primal. Num espaço qualquer, no entanto, isso não é possível, o que nos leva ao *Mirror Descent*. Quando estabelecermos o MD, vamos nos certificar de que elementos de espaços diferentes, primal e dual, não sejam somados entre si.

Definido o espaço dual, podemos agora definir a norma dual, a norma natural a ser usada no espaço dual, induzida pela norma usada no espaço primal.

Definição (Norma Dual)**.** *Para cada $\varphi \in X^*$, sua* ***norma dual*** *é definida por*

$$\|\varphi\|_* = \sup_{\|x\| \leq 1} |\varphi(x)|.$$

Da mesma forma que a norma primal, a dual mede o tamanho dos elementos do espaço dual. Geometricamente a norma dual representa o valor máximo que um funcional φ pode atingir (em módulo) ao ser aplicado a vetores da bola unitária do espaço primal X. Em outras palavras, ela mede o máximo que o funcional "deforma" ou "estica" a bola unitária de X ao mapeá-la em $\mathbb{R}$.

Exemplo. *Para $1 < p < \infty$, as normas duais relativas às normas ℓ_p são as normas ℓ_q, onde $\frac{1}{p} + \frac{1}{q} = 1$. Um caso particular importante é a norma euclidiana ℓ_2, que é auto dual, isto é, dual de si mesma.*

Exemplo. *A norma ℓ_∞ é dual da norma ℓ_1 e vice-versa.*

O fato a seguir será fundamental ao longo desta dissertação e é uma generalização da desigualdade de Cauchy-Schawrz (quando a norma é a euclidiana).

Fato 2.2. *Seja $\varphi \in X^*$ um funcional linear sobre X, um espaço dotado de norma $\|\cdot\|$. Então para todo $x \in X$, vale*

$$\varphi(x) \leq \|\varphi\|_* \|x\|.$$

Demonstração. pela definição de norma dual, temos que $\|\varphi\|_* \geq |\varphi(x/\|x\|)|$, pois $\|x/\|x\|\| = 1$ (fato 2.1). Pela linearidade de φ, temos então que $\varphi(x/\|x\|) = \varphi(x)/\|x\|$, o que nos leva a

$$\varphi(x) = \varphi(x/\|x\|)\|x\| \leq |\varphi(x/\|x\|)|\|x\| \leq \|\varphi\|_* \|x\|.$$

■

2.2 Continuidade e Diferenciabilidade

Os conceitos de continuidade e diferenciabilidade são fundamentais em otimização e análise e são usados diversas vezes neste trabalho.

Uma função contínua, por definição, é aquela em que pequenas variações no domínio resultam em pequenas variações na imagem. Na verdade, um pouco mais do que isso, é aquela em que, com uma variação tendendo a zero no domínio, alcançamos uma variação também tendendo a zero na imagem. Essa ideia é formalizada a seguir.

Definição (Continuidade)**.** *Sejam $(X, \|\cdot\|_X)$ e $(Y, \|\cdot\|_Y)$ espaços normados. Uma função $f : X \to Y$ é **contínua** em $x_0 \in X$ se para todo $\epsilon > 0$, existe $\delta > 0$ tal que*

$$\|x - x_0\|_X < \delta \implies \|f(x) - f(x_0)\|_Y < \epsilon.$$

Dizemos que f é contínua em X se for contínua em todos os pontos de X.

Dada essa definição, temos o seguinte resultado que estabelece uma caracterização importante para funções contínuas.

Teorema 2.3. *Sejam $(X, \|\cdot\|_X)$ e $(Y, \|\cdot\|_Y)$ espaços normados. Uma função $f : X \to Y$ é **contínua** em x se*

$$\lim_{y \to x} f(y) = f(x).$$

Esse teorema garante que o valor de uma função contínua em x avaliada em x coincide com o limite da função em x. De maneira mais intuitiva, podemos pensar que uma função é contínua em um ponto se ela não tem um "salto" de valor nesse ponto. Por extensão, uma função é contínua se ela não tem nenhum "salto" em nenhum ponto do seu domínio.

Todos os algoritmos estudados neste trabalho levam em consideração o gradiente da função a ser otimizada, que nada mais é do que a sua derivada. A seguir formalizamos o conceito de função diferenciável e derivada. Vamos usar a definição de Fréchet aqui, pois ela evidencia que a derivada deve pertencer ao espaço dual.

Definição (Fréchet Diferenciabilidade)**.** *Uma função $f : X \to Y$ é **diferenciável** em $x \in X$ no sentido de Fréchet (ou Fréchet-diferenciável) se existe um operador linear limitado $A : X \to Y$ tal que*

$$\lim_{h \to \vec{0}} \frac{\|f(x+h) - f(x) - Ah\|_Y}{\|h\|_X} = 0.$$

*Neste caso, escrevemos $A = \nabla f(x)$ e dizemos que $\nabla f(x)$ é a **derivada de Fréchet** de f em x. Dizemos que f é diferenciável em X se for diferenciável em todos os pontos de X.*

Observação. *Nesta definição, $h \to \vec{0}$ significa que tomamos uma sequência de pontos cujo limite é $\vec{0}$. Não confundir com vetores de mesma direção e sentido com tamanho tendendo a zero (como é no caso de derivada direcional).*

Essa derivada é diferente da derivada mais usual: a derivada de Gâteaux. A derivada de gâteaux mede, para cada v, a variação da função ao longo da direção v, sendo, portanto, uma generalização da derivada direcional. No $\mathbb{R}^n$, quando dispomos de produto interno e a derivada de Gâteaux satisfaz condições de linearidade, costumamos pensar nessa derivada tomada em um ponto x como o operador $\langle \nabla^G f(x), \cdot \rangle$, onde $\nabla^G f(x) = (D_{e_1} f(x), D_{e_2} f(x), \ldots, D_{e_n} f(x)), \cdot \rangle$, sendo e_i o i-ésimo vetor canônico. Nesse caso, o gradiente $\nabla^G f(x)$ é visto também como um vetor do $\mathbb{R}^n$, o espaço primal, o que é possível graças ao teorema da representação de Riesz. É dessa forma que o GD usa o gradiente.

A derivada de Fréchet, por sua vez, é entendida como uma derivada no sentido de operador linear. Ela exige que a aproximação da função por esse operador seja válida uniformemente em todas as direções ao redor do ponto considerado. Em espaços normados, toda função Fréchet-diferenciável é também Gâteaux-diferenciável, mas o contrário nem sempre é verdadeiro. Além disso, a derivada de Fréchet nem sempre corresponde a um vetor no espaço primal: ela é um elemento do espaço dual. Isso exige que as operações

feitas com gradientes no GD sejam ajustadas ao espaço dual, o que nos leva ao *Mirror Descent*, onde o gradiente é tratado como um funcional dual e o mapeamento de volta ao espaço primal é feito com o auxílio de uma função espelho.

Uma função ser Fréchet-diferenciável em um ponto ou em um conjunto implica sua continuidade nesse mesmo ponto ou conjunto.

Quando a derivada de Gâteaux é representável como vetor — como no caso euclidiano — esse vetor aponta na direção de maior crescimento da função. Além disso, com essa representação, podemos usar o gradiente para classificar vetores como direções de subida ou de descida simplesmente verificando se sua correlação com o gradiente é positiva ou negativa. O teorema e o corolário a seguir expressam essa mesma ideia para a derivada de Fréchet, usando, em vez do produto interno com o gradiente, a aplicação do gradiente (como operador, dada a definição) sobre um ponto a se avaliar.

Teorema 2.4. *Seja $f : X \to \mathbb{R}$ Fréchet-diferenciável em x e $y \in X$ tal que $\nabla f(x)(y) < 0$. Então, existe $\delta > 0$ tal que para todo $\alpha \in (0, \delta)$*

$$f(x + \alpha y) < f(x).$$

Demonstração. Pela diferenciabilidade de Fréchet, tomando $h = \alpha y$ com $\alpha \to 0^+$ temos

$$\lim_{\alpha \to 0^+} \frac{|f(x + \alpha y) - f(x) - \alpha \nabla f(x) y|}{\alpha \|y\|_X} = 0.$$

Isso implica que para qualquer $\epsilon > 0$, existe $\delta_\epsilon > 0$ tal que para todo $\alpha \in (0, \delta_\epsilon)$, vale

$$\begin{aligned} |f(x + \alpha y) - f(x) - \alpha \nabla f(x) y| < \epsilon \alpha \|y\|_X \\ \implies \quad |f(x + \alpha y)| - |f(x)| - |\alpha \nabla f(x) y| < \epsilon \alpha \|y\|_X \\ \implies \quad |f(x + \alpha y)| < |f(x)| + |\alpha \nabla f(x) y| + \epsilon \alpha \|y\|_X . \end{aligned}$$

Fazendo $\epsilon = -\frac{\nabla f(x) y}{2 \|y\|_{\mathcal{X}}} > 0$, obtemos que para todo $\alpha \in (0, \delta_\epsilon)$,

$$\begin{aligned} f(x + \alpha y) < f(x) + \alpha \nabla f(x) y + \epsilon \alpha \|y\|_{\mathcal{X}} &= f(x) + \alpha \left(\nabla f(x) y + \frac{-\nabla f(x) y}{2} \right) \\ &= f(x) + \alpha \left(\frac{1}{2} \nabla f(x) y \right) < f(x). \end{aligned}$$

■

Ou seja, se um vetor y é negativamente correlacionado (mas não n sentido de produto escalar) a derivada $\nabla f(x)$ de uma função f num ponto x ou, equivalentemente, positivamente correlacionado ao oposto da derivada $-\nabla f(x)$, então esse vetor representa uma direção de descida de valor para a função f. Por ser linear, o gradiente define assim um plano separador que divide os vetores entre direções de descida e de subida para a função no ponto onde ele é tomado. O próximo corolário apresenta uma pequena variação do teorema anterior e sua demonstração se dá praticamente da mesma maneira, apenas invertendo alguns sinais.

Corolário 2.5. *Seja $f : X \to \mathbb{R}$ Fréchet-diferenciável em x e $y \in \mathcal{X}$ tal que $\nabla f(x) y > 0$. Então, existe $\delta > 0$ tal que para todo $\alpha \in (0, \delta)$*

$$f(x + \alpha y) > f(x).$$

Além desses resultados, vale destacar que valem também resultados clássicos de cálculo para a derivada de Fréchet que usaremos neste trabalho. São eles: Teorema Fundamental do Cálculo, Teorema do Valor Médio, Teorema do Valor Intermediário ([Die11],[AB06], [Zei12]), Teorema da Função Inversa ([Lan12], [AP95]), e regras operacionais como linearidade da derivada, regra do produto e do quociente e regra da cadeia ([Die11], [AB06], [Car12]).

2.3 Convexidade

A convexidade desempenha um papel fundamental em otimização, pois assegura que todo mínimo local seja também um mínimo global. Além disso, ela permite alcançar aproximações do mínimo global por meio de

métodos iterativos baseados em decisões locais, como os analisados aqui. Nesta seção, exploramos conjuntos e funções convexas, bem como variantes importantes, como funções estritamente convexas e fortemente convexas. Iniciamos com a definição de conjuntos convexos.

Definição (Conjunto convexo)**.** *Um conjunto $\mathcal{X} \subseteq \mathbb{R}^n$ é dito* ***convexo*** *se para quaisquer $x, y \in \mathcal{X}$ e qualquer $t \in [0,1]$, vale que*

$$tx + (1-t)y \in \mathcal{X}.$$

Isto significa que o segmento de reta ligando x a y está inteiramente contido em $\mathcal{X}$.

Combinações de x e y da forma $tx + (1-t)y$, com $t \in [0,1]$ são chamadas de combinações convexas e definem pontos do segmento de reta que une os dois pontos (mas não da reta toda passando pelos dois pontos). Intuitivamente, um conjunto convexo é aquele que não tem buracos ou sobressalências ou ainda aquele cuja fronteira está sempre "curvada para dentro", como uma bola.

Exemplo. *Bolas fechadas, definidas por $B_r(x) = \{y \in \mathbb{R}^n | \, \|y - x\| \leq r\}$, onde x é o centro da bola e r seu raio, são convexas.*

Exemplo. *Hiperplanos, definidos por $H = \{x \in \mathbb{R}^n | Ax = b\}$, são convexos.*

Exemplo. *semiespaços, definidos por $H^+ = \{x \in \mathbb{R}^n | Ax \leq b\}$, são convexos.*

Exemplo. *O simplexo n-dimensional, definido por*

$$\Delta^{n-1} = \left\{ x \in \mathbb{R}^n \mid x_i \geq 0, \sum_{i=1}^{n} x_i = 1 \right\},$$

é convexo.

Prosseguimos agora com a definição de funções convexas e duas de suas versões mais fortes. Neste trabalho, a convexidade da função objetivo f garante a existência de soluções aproximadas por métodos iterativos. Já a convexidade estrita da função geradora usada no mapeamento espelho (em Mirror Descent) assegura a injetividade do gradiente, enquanto a convexidade forte dessa mesma função permitirá relacionar distâncias em coordenadas duais e primais.

Definição (Função convexa)**.** *Seja $\mathcal{X}$ um conjunto convexo. Uma função $f : \mathcal{X} \to \mathbb{R}$ é dita* ***convexa*** *se para quaisquer $x, y \in \mathcal{X}$ e $t \in [0,1]$ vale*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

Definição (Função estritamente convexa)**.** *Seja $\mathcal{X}$ um conjunto convexo. Uma função $f : \mathcal{X} \to \mathbb{R}$ é* ***estritamente convexa*** *se para quaisquer $x, y \in \mathcal{X}$ com $x \neq y$ e $t \in (0,1)$ vale*

$$f(tx + (1-t)y) < tf(x) + (1-t)f(y).$$

Definição (Função fortemente convexa)**.** *Seja $\mathcal{X}$ um conjunto convexo. Uma função $f : \mathcal{X} \to \mathbb{R}$ é* ***ν-fortemente convexa*** *se para quaisquer $x, y \in \mathcal{X}$ e $t \in [0,1]$ vale*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{\nu}{2} t(1-t) \|x - y\|^2.$$

Assim, uma função f definida sobre $\mathcal{X}$ é convexa se para todo par de pontos $x, y \in \mathcal{X}$, o gráfico de f no intervalo (x, y) fica abaixo ou coincide com o segmento de reta que liga $(x, f(x))$ e $(y, f(y))$. Para convexidade estrita, o gráfico deve ficar abaixo do segmento, sem coincidir com ele (exceto nas extremidades). Para convexidade forte, o gráfico da função f deve ficar abaixo ou coincidir com uma certa parábola (não só um segmento de reta) que une os pontos $(x, f(x))$ e $(y, f(y))$. Notemos que convexidade forte implica em convexidade estrita, que por sua vez implica em convexidade simples.

Para funções diferenciáveis, temos as caracterizações dadas pelos próximos teoremas.

Teorema 2.6. *Seja $\mathcal{X}$ um conjunto convexo. Uma função $f : \mathcal{X} \to \mathbb{R}$ diferenciável é convexa se, e somente se, para quaisquer $x, y \in \mathcal{X}$,*

$$f(y) \geq f(x) + \nabla f(x)(y - x).$$

Teorema 2.7. *Seja $\mathcal{X}$ um conjunto convexo. Uma função $f : \mathcal{X} \to \mathbb{R}$ diferenciável é estritamente convexa se, e somente se, para quaisquer $x, y \in \mathcal{X}$, com $x \neq y$,*

$$f(y) > f(x) + \nabla f(x)(y - x).$$

Teorema 2.8. *Seja $\mathcal{X}$ um conjunto convexo. Uma função $f : \mathcal{X} \to \mathbb{R}$ diferenciável é ν-fortemente convexa se, e somente se, para quaisquer $x, y \in \mathcal{X}$,*

$$f(y) \geq f(x) + \nabla f(x)(y - x) + \frac{\nu}{2} \|y - x\|^2 .$$

Ou seja, uma função f diferenciável é convexa se o gráfico de f fica acima ou coincide com o gráfico de qualquer hiperplano tangente definido pelo gradiente. Para convexidade estrita, o gráfico deve ficar acima de qualquer hiperplano, sem coincidir com ele exceto, é claro, nos pontos onde as derivadas que os definem são tomadas. Para convexidade forte, o gráfico de f deve ficar acima de uma certa quadrática com mesmo valor e mesmo gradiente no ponto x.

Para funções de classe C^2, isto é, duas vezes diferenciáveis, com derivadas de segunda ordem contínuas, temos ainda as caracterizações dadas pelos próximos teoremas.

Teorema 2.9. *Seja $\mathcal{X}$ um conjunto convexo. Uma função $f : \mathcal{X} \to \mathbb{R}$ duas vezes diferenciável é convexa se, e somente se, para qualquer $x \in \mathcal{X}$, $\nabla^2 f(x)$ é semidefinida positiva.*

Teorema 2.10. *Seja $\mathcal{X}$ um conjunto convexo. Uma função $f : \mathcal{X} \to \mathbb{R}$ duas vezes diferenciável é estritamente convexa se, e somente se, para qualquer $x \in \mathcal{X}$, $\nabla^2 f(x)$ é definida positiva.*

Teorema 2.11. *Seja $\mathcal{X}$ um conjunto convexo. Uma função $f : \mathcal{X} \to \mathbb{R}$ duas vezes diferenciável é ν-fortemente convexa se, e somente se, para qualquer $x \in \mathcal{X}$, $\nabla^2 f(x) - \nu I$ é semidefinida positiva.*

A positividade definida da Hessiana $\nabla^2 f(x)$ implica que o gradiente $\nabla f(x)$ é injetivo em torno de x, o que significa que variações em x resultam em variações em todas as direções possíveis no gradiente. Assim, se f é estritamente ou fortemente convexa, seu gradiente define um difeomorfismo entre o domínio $\mathcal{X}$ e a imagem $f(\mathcal{X})$.

As definições e caracterizações desta seção estão resumidas na Tabela 2.1.

Convexidade	**Definição**	$f \in C^1$	$f \in C^2$
Simples	$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$	$f(y) \geq f(x) + \nabla f(x)(y - x)$	$\nabla^2 f(x) \succeq 0$
Estrita	$f(tx + (1-t)y) < tf(x) + (1-t)f(y)$	$f(y) > f(x) + \nabla f(x)(y - x)$	$\nabla^2 f(x) \succ 0$
Forte	$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$ $-\frac{\nu}{2} t(1-t) \|y - x\|^2$	$f(y) \geq f(x) + \nabla f(x)(y - x) + \frac{\nu}{2} \|y - x\|^2$	$\nabla^2 f(x) \succeq \nu I$

Tabela 2.1: Tabela recapitulativa das noções de convexidade.

2.4 Conjuntos de Nível

A ideia de conjuntos de nível, apesar de não ser central aqui, é bastante útil para entendermos mais facilmente alguns resultados. Nesta seção exploramos este objeto e algumas de suas propriedades, em especial no contexto de otimização convexa.

Definição (Conjunto de nível)**.** *Dada uma função $f : \mathbb{R}^n \to \mathbb{R}$, definimos os* ***conjuntos de nível*** *de f para um valor $c \in \mathbb{R}$ como*

$$L_c(f) = \{x \in \mathbb{R}^n \mid f(x) \leq c\}.$$

No caso de funções convexas, os conjuntos de nível possuem uma propriedade especial, estabelecida pelo teorema a seguir.

Teorema 2.12. *Seja $\mathcal{X}$ um conjunto convexo e $f : \mathcal{X} \to \mathbb{R}$ é uma função convexa, então para todo $c \in \mathbb{R}$, o conjunto de nível $L_c(f)$ é um conjunto convexo.*

Demonstração. Sejam $x, y \in L_c(f)$, isto é, tais que $f(x) < c$ e $f(y) < c$. Então, pela definição de função convexa, para todo $t \in [0, 1]$,$f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y) \leq tc + (1 - t)c = c$ de forma que $tx + (1 - t)y \in L_c(f)$. Como x e y são pontos arbitrários, isso vale para qualquer par de pontos pertencente a $L_c(f)$, o que caracteriza a convexidade de $L_c(f)$. ■

Os fatos a seguir tratam da relação entre os gradientes e os conjuntos de nível de uma função convexa.

Fato 2.13. *Sejam $\mathcal{X}$ um conjunto convexo e $f : \mathcal{X} \to \mathbb{R}$ uma função convexa. Sejam $c \in \mathbb{R}$ qualquer e $x \in \mathcal{X}$ um ponto tal que $f(x) = c$. Então para todo $y \in L_c(f)$,*

$$\nabla f(x)(x - y) \geq 0.$$

Demonstração. Seja x tal que $f(x) = c$. Suponhamos por absurdo que exista $y \in L_c(f)$ tal que $\nabla f(x)(x - y) < 0$. Então $\nabla f(x)(y - x) > 0$ e aí, pelo corolário 2.5, com $\alpha > 0$ pequeno o suficiente, o ponto $y' = x + \alpha(y - x) = (1 - \alpha)x + \alpha y \in L_c(f)$ é tal que $f(y') > f(x) = c$. Mas isso é um absurdo, pois $y' \in L_c(f)$ e deveria satisfazer $f(y') \leq c$. ■

Geometricamente, quando vemos o gradiente $\nabla f(x)$ como um vetor, isso significa que ele aponta sempre "para fora" do conjunto de nível para o valor $f(x)$. Além disso, ele define um hiperplano que divide o espaço em dois semiespaços, sendo que um deles contém o conjunto de nível inteiro.

Fato 2.14. *Sejam $f : \mathbb{R}^n \to \mathbb{R}$ contínua e diferenciável e $\mathcal{X} \subseteq \mathbb{R}^n$ convexo e fechado. Então*

$$y^* \in \arg\min_{x \in \mathcal{X}} f(x) \quad \Longrightarrow \quad \forall y \in \mathcal{X}, \nabla f(y^*)(y^* - y) \leq 0.$$

Demonstração. Seja $y^* \in \arg\min_{x \in \mathcal{X}} f(x)$. Suponhamos por absurdo que exista $y \in \mathcal{X}$ tal que $\nabla f(y^*)(y^* - y) > 0$. Então, pelo corolário 2.5, com um $\alpha > 0$ pequeno o suficiente, o ponto $y' = y^* + \alpha(y^* - y)$, que pertence a $\mathcal{X}$ pela convexidade do conjunto, é tal que $f(y') < f(y^*)$, o que é absurdo, pois y^* deveria ser minimizador de f. ■

Isso implica que o ponto minimizador do Problema P, sujeito à restrição $x \in \mathcal{X}$, deve satisfazer uma das seguintes condições de otimalidade:

1. $\nabla f(y^*) = 0$, ou
2. $\nabla f(y^*)(y - y^*) \geq 0$ para todo $y \in \mathcal{X}$, com y^* na fronteira de $\mathcal{X}$.

Essa caracterização é análoga à descrita no Fato 2.13 para conjuntos de nível.

Geometricamente, quando visto como vetor, o oposto gradiente $\nabla f(x^*)$ indica a direção de maior decrescimento da função. No segundo caso, como o ponto ótimo x^* situa-se na fronteira de $\mathcal{X}$ e esse vetor aponta para fora do conjunto, ele não pode ser melhorado por movimentos viáveis dentro de $\mathcal{X}$. É como se o gradiente tentasse "empurrar" o ponto ótimo para fora de $\mathcal{X}$, mas não pudesse por conta da fronteira.

Embora a recíproca também seja válida, neste documento interessamo-nos apenas pela implicação direta. Essa propriedade será fundamental para demonstrar a existência e unicidade de uma certa projeção.

2.5 Funções Lipschitz

Por fim, nesta seção definimos e exploramos o conceito de função Lipschitz, que será importante em vários momentos para controlar distâncias ou tamanhos de gradientes, assegurando convergências, além de garantir a aplicação de importantes teoremas de teoria das EDOs (Picard-Lindeöf e teorema da dependência contínua nas condições iniciais).

Definição (Condição de Lipschitz)**.** *Sejam $(X, \|\cdot\|_X)$ e $(Y, \|\cdot\|_Y)$ espaços normados. Uma função $f : X \to Y$ é L-**Lipschitz** em torno de $x_0 \in X$ se para todo x em uma vizinhança de x_0, vale que*

$$\|f(x) - f(x_0)\|_Y \leq L \|x - x_0\|_X .$$

*Dizemos que f é L-**Lipschitz** em X se o for em todos os pontos de X.*

Uma função f ser Lipschitz significa que ela varia de maneira controlada, sem que seu valor mude drasticamente com pequenas mudanças nas variáveis. Se f for (Fréchet) diferenciável, o seu gradiente em um ponto x pode ser usado como uma aproximação para f em torno de x, nos permitindo medir variações em uma vizinhança de x. Lembrando que a norma dual do gradiente mede a maior transformação (em módulo) sobre vetores unitários, podemos usá-la para elaborar uma caracterização da condição de Lipschitz, como no fato a seguir

Fato 2.15 (Caracterização via derivada)**.** *Sejam $(X, \|\cdot\|_X)$ e $(Y, \|\cdot\|_Y)$ espaços normados e $f : X \to Y$ uma função diferenciável. Então f é L-Lipschitz em torno de x_0 se, e somente se,*

$$\|\nabla f(x_0)\|_{\mathcal{L}(X,Y)} \leq L,$$

onde $\|\nabla f(x_0)\|_{\mathcal{L}(X,Y)}$ denota a norma de operador da derivada de f em x_0.

Além disso, f é L-Lipschitz em X se, e somente se,

$$\|\nabla f(x)\|_{\mathcal{L}(X,Y)} \leq L \quad \textit{para todo } x \in X.$$

Demonstração. ($\Rightarrow$) Suponha que f é L-Lipschitz em X. Queremos mostrar que $\|\nabla f(x)\|_{\mathcal{L}(X,Y)} \leq L$ para todo $x \in X$.

Seja $x \in X$ arbitrário. Como f é Fréchet diferenciável em x, existe um operador linear limitado $\nabla f(x) : X \to Y$ tal que

$$\lim_{h \to \vec{0}} \frac{\|f(x+h) - f(x) - \nabla f(x)(h)\|_Y}{\|h\|_X} = 0.$$

Mais precisamente, para qualquer $\varepsilon > 0$, existe $\delta > 0$ tal que se $\|h\|_X < \delta$, então

$$\frac{\|f(x+h) - f(x) - \nabla f(x)(h)\|_Y}{\|h\|_X} < \varepsilon.$$

Agora, para qualquer $h \neq 0$ com $\|h\|_X < \delta$, temos pela desigualdade triangular

$$\|\nabla f(x)(h)\|_Y \leq \|f(x+h) - f(x)\|_Y + \|f(x+h) - f(x) - \nabla f(x)(h)\|_Y.$$

Dividindo ambos os lados por $\|h\|_X$

$$\frac{\|\nabla f(x)(h)\|_Y}{\|h\|_X} \leq \frac{\|f(x+h) - f(x)\|_Y}{\|h\|_X} + \frac{\|f(x+h) - f(x) - \nabla f(x)(h)\|_Y}{\|h\|_X}.$$

Como f é L-Lipschitz, temos

$$\|f(x+h) - f(x)\|_Y \leq L \|x + h - x\|_X = L \|h\|_X \quad \Longrightarrow \quad \frac{\|f(x+h) - f(x)\|_Y}{\|h\|_X} \leq L.$$

Além disso, pela diferenciabilidade de Fréchet, o segundo termo é menor que ε para $\|h\|_X$ suficientemente pequeno. Portanto

$$\frac{\|\nabla f(x)(h)\|_Y}{\|h\|_X} \leq L + \varepsilon.$$

Agora, para qualquer vetor unitário $u \in X$ (isto é, $\|u\|_X = 1$), tome $h = tu$ com $0 < t < \delta$. Então

$$\frac{\|\nabla f(x)(tu)\|_Y}{\|tu\|_X} = \frac{\|t\nabla f(x)(u)\|_Y}{t} = \|\nabla f(x)(u)\|_Y \leq L + \varepsilon.$$

Como $\varepsilon > 0$ é arbitrário, concluímos que $\|\nabla f(x)(u)\|_Y \leq L$ para todo vetor unitário $u \in X$. Portanto

$$\|\nabla f(x)\|_{\mathcal{L}(X,Y)} = \sup_{\|u\|_X=1} \|\nabla f(x)(u)\|_Y \leq L.$$

($\Leftarrow$) Suponha que $\|\nabla f(x)\|_{\mathcal{L}(X,Y)} \leq L$ para todo $x \in X$. Queremos mostrar que f é L-Lipschitz.

Sejam $x, y \in X$ distintos. Considere o segmento $[x, y] = \{x + t(y - x) : t \in [0, 1]\}$ e defina a função $\varphi : [0, 1] \to Y$ por

$$\varphi(t) = f(x + t(y - x)).$$

Pela regra da cadeia para derivada de Fréchet, φ é diferenciável em $(0, 1)$ e

$$\varphi'(t) = \nabla f(x + t(y - x))(y - x).$$

Agora, pela desigualdade triangular e pela hipótese sobre a norma da derivada

$$\|\varphi'(t)\|_Y = \|\nabla f(x + t(y - x))(y - x)\|_Y \leq \|\nabla f(x + t(y - x))\|_{\mathcal{L}(X,Y)} \cdot \|y - x\|_X \leq L\|y - x\|_X.$$

Para concluir, usamos o teorema fundamental do cálculo para funções vetoriais com valores em espaços de Banach. No nosso caso, como f é Fréchet diferenciável, φ é diferenciável e portanto contínua, e sua derivada φ' é limitada por $L\|y - x\|_X$. Assim

$$f(y) - f(x) = \varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)\, dt.$$

Tomando normas e usando a propriedade da integral

$$\|f(y) - f(x)\|_Y = \left\| \int_0^1 \varphi'(t)\, dt \right\|_Y \leq \int_0^1 \|\varphi'(t)\|_Y\, dt \leq \int_0^1 L\|y - x\|_X\, dt = L\|y - x\|_X.$$

Portanto, f é L-Lipschitz em X. ■

Também podemos provar que toda função Lipschitz é contínua. Além disso, o teorema de Rademacher ([Eva18]) garante que se $\mathcal{X} = \mathbb{R}^n$ e $\mathcal{Y} = R^m$ para $n, m \in \mathbb{N}$ quaisquer, então $f : \mathcal{X} \to \mathcal{Y}$ ser Lipschitz implica que f é diferenciável em quase toda parte. Para espaços normados gerais, contudo, essa conclusão não vale sem hipóteses adicionais: é necessário que o espaço de chegada satisfaça a propriedade de Radon–Nikodým ou que seja separável, por exemplo ([Bat93]).

Capítulo 3

Gradient Descent

Neste capítulo, apresentamos o método do gradiente descendente, ou *Gradient Descent*. Neste capítulo e no próximo, vamos assumir que a norma usada é a euclidiana, de forma que faça sentido pensar no gradiente de uma função como um vetor primal e somá-lo ou tomar seu produto interno com outro vetor primal.

3.1 Definição e interpretações

O *Gradient Descent* (GD) é um algoritmo iterativo bastante conhecido, usado para resolver o problema de otimização convexa P. O algoritmo funciona escolhendo-se um ponto viável $x_0 \in \mathbb{R}^n$ e, para cada iteração, isto é, para cada k, resolvendo o problema

$$x_{k+1} = \arg\min_x \left\{ \eta \langle \nabla f(x_k), x \rangle + \frac{1}{2} \|x - x_k\|_2^2 \right\}, \tag{3.1}$$

onde $\eta > 0$ é chamado de tamanho de passo ou taxa de aprendizado.

Fazer $x_{k+1} = x_k - \alpha \nabla f(x_k)$, isto é, mover-se na direção de $-\nabla f(x_k)$ (a direção de maior descida de f em x_k), diminui o termo $\eta \langle \nabla f(x_k), x \rangle$ de maneira proporcional tanto ao tamanho α do passo dado quanto ao tamanho de η, o que contribui para a minimização da expressão. No entanto, afastar-se de x_k em qualquer direção aumenta o termo $\frac{1}{2} \|x - x_k\|_2^2$ na proporção do quadrado do passo dado, o que prejudica a minimização. O equilíbrio entre os termos é determinado pela taxa η; quanto maior, mais vale a pena se mover na direção $-\nabla f(x_k)$. O problema então consiste em encontrar o tamanho ótimo do passo a ser dado na direção de $-\nabla f(x_k)$ e a ideia por trás do algoritmo é, na iteração $k+1$, atualizar o ponto anterior x_k movendo-se nesta direção, mas sem afastar-se muito de x_k.

Aqui estamos definindo o método com um passo de tamanho η fixo. O método pode ser adaptado para que o tamanho do passo varie em cada iteração. Uma alternativa popular é fazer o passo variar conforme a relação $\eta_k \propto 1/k$, que permite que passos maiores sejam dados no começo e tem um refinamento com o passar do tempo, mas sem limitar o espaço de busca (já que a série harmônica é divergente). Outras alternativas podem ser desenhadas para cada método de maneira mais ótima. Além disso, alguns algoritmos otimizam o tamanho do passo a cada iteração. Essas otimizações secundárias são as chamadas buscas lineares ou *line searches*, como a de Armijo ([Arm66]), Goldstein-Price ([Gol65]) e Wolfe ([Wol69] e [Wol71]).

Seja $g_k(x) = \eta \langle \nabla f(x_k), x \rangle + \frac{1}{2} \|x - x_k\|_2^2$, a função a ser minimizada no tempo $k+1$. Então o método pode ser escrito como

$$x_{k+1} = \arg\min_x \ g_k(x).$$

Notemos que, como $\eta > 0$, minimizar g_k é o mesmo que minimizar

$$g_k/\eta = \langle \nabla f(x_k), x \rangle + \frac{1}{2\eta} \|x - x_k\|_2^2 = (\langle \nabla f(x_k), x \rangle - f(x_k)) + \left(\frac{1}{2\eta} \|x - x_k\|_2^2 + f(x_k) \right).$$

Escrito como um problema de minimização de g_k/η, o GD é equivalente ao método do ponto proximal aplicado à aproximação linear de f introduzido Rockafellar [Roc76] e Martinet [Mar70].

Para cada $k \in \mathbb{N}$, o primeiro termo é a aproximação linear de f no ponto x_k e o segundo termo é uma função quadrática cujo mínimo é x_k, que funciona como uma função de penalização segundo a distância do ponto atual. Por esta ótica, o método busca minimizar a aproximação linear da função f, mas com o cuidado para não se afastar muito, o que faz bastante sentido na verdade, já que a aproximação é local, isto é, só é boa perto de onde ela é tomada.

Apesar de g_k ser uma simples função quadrática, a formulação mais usual do GD não é a escrita como um problema de minimização de g_k em cada tempo k, mas sim a que contém a sua solução explícita. O ponto x_{k+1} que minimiza g_k deve ser tal que

$$\begin{aligned} 0 &= \frac{\partial g_k}{\partial x_{k+1}}(x_{k+1}) \\ &= \frac{\partial}{\partial x_{k+1}} \left(\eta \langle \nabla f(x_k), x_{k+1} \rangle - \frac{1}{2} \left\| x_{k+1} - x_k \right\|^2 \right) \\ &= \eta \nabla f(x_k) - x_{k+1} + x_k. \end{aligned}$$

Ou seja, para todo k, x_{k+1} deve satisfazer

$$x_{k+1} = x_k - \eta \nabla f(x_k). \tag{3.2}$$

Esta formulação mais comum do GD evidencia que o tamanho do passo ótimo a ser dado é controlado por η. Quando maior o tamanho de η maiores são os passos dados na direção de descida, o que justifica enfim o nome "taxa de aprendizado". A Figura 3.1 ilustra o passo dado pelo GD em uma iteração.

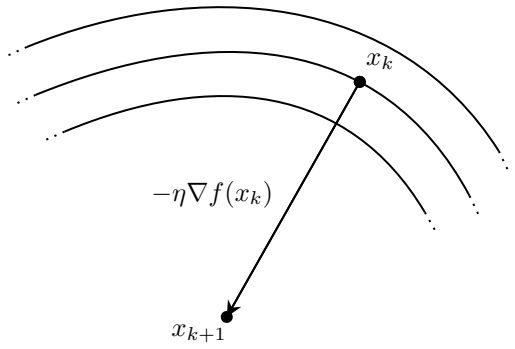

Figura 3.1: Uma iteração do método do gradiente

A Figura 3.2 ilustra a trajetória do *Gradient Descent* sobre as curvas de nível da função objetivo. Para a confecção dessa e outras figuras que ilustram trajetórias de métodos ou de soluções de EDOs associadas, a função objetivo utilizada foi a quadrática $f(x, y) = (x - 128)^2 + (y + 56)^2 + (x - 128)(y + 56)$. Notemos que o percurso é perpendicular às curvas de nível, uma vez que os passos são dados na direção do gradiente.

3.2 Gradient Descent com Projeção para Conjuntos Convexos

O método de *Gradient Descent* pode ser facilmente adaptado para resolver aproximadamente problemas de otimização restritos a conjuntos convexos fechados $\mathcal{X} \subset \mathbb{R}^n$, isto é, problemas escritos como

$$(P') = \begin{cases} \min f(x) \\ x \in \mathcal{X} \end{cases}, \tag{P'}$$

com f convexa e diferenciável em um aberto contendo $\mathcal{X}$ com gradiente ∇f L-Lipschitz.

Uma forma direta de lidar com a restrição é, após cada passo de descida, projetar o ponto obtido de volta em $\mathcal{X}$. A atualização assume a forma

$$\begin{cases} x'_{k+1} = x_k - \eta \nabla f(x_k) \\ x_{k+1} = \Pi_{\mathcal{X}}\left(x'_{k+1}\right), \end{cases}$$

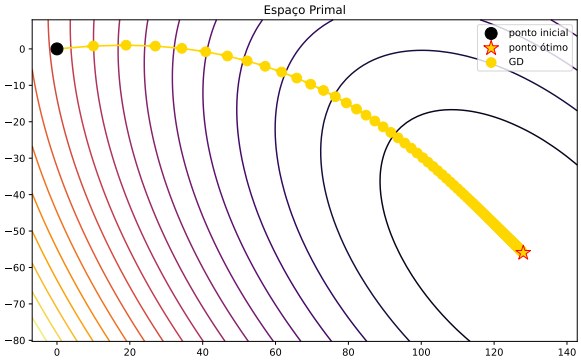

Figura 3.2: Trajetória do GD sobre curvas de nível da função objetivo.

onde $\Pi_{\mathcal{X}}$ denota a projeção euclidiana sobre $\mathcal{X}$, definida por

$$\Pi_{\mathcal{X}}(y) = \arg\min_{z \in \mathcal{X}} \|z - y\|_2 .$$

Se além de convexo e fechado, $\mathcal{X}$ for não vazio, então a projeção euclidiana existe e é única.

Essa adaptação é conhecida como *Projected Gradient Descent* (PGD) ou método do gradiente projetado e mantém a simplicidade conceitual do método original.A Figura 3.3 ilustra uma iteração do método do gradiente projetado.

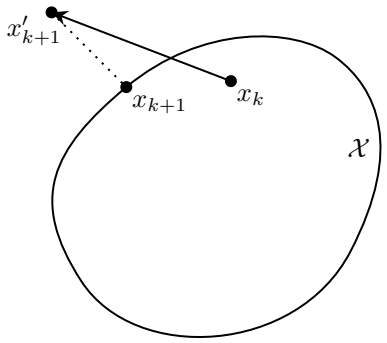

Figura 3.3: Uma iteração do método do gradiente projetado

Em particular, quando $\mathcal{X}$ possui estrutura especial (como em caixas do tipo $[a, b]^n$, ou conjuntos definidos por desigualdades lineares), a solução do problema da projeção é dada em uma fórmula fechada e pode ser implementada de forma eficiente.

Convergência do Método do Gradiente

Sob as hipóteses que fizemos sobre a função objetivo f é possível estabelecer uma taxa sublinear de convergência para a sequência $(x_k)_{k \in \mathbb{N}}$ gerada pelo método do gradiente com passo constante $\eta = \frac{1}{L}$. O teorema a seguir faz isso.

Teorema 3.1. *Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função convexa, diferenciável e com gradiente L-Lipschitz. Sejam x^* um ponto ótimo de f e $\{x_k\}_{k \in \mathbb{N}}$ a sequência de pontos gerada pelo GD com tamanho de passo $1/L$. Então, para todo $k \in \mathbb{N}$,*

$$f(x_k) - f(x^*) \leq \frac{L}{2k} \|x_0 - x^*\|^2 .$$

A demonstração deste resultado com argumentos diretos pode ser encontrada em [Cal20] ou [Bub15a]. Uma demonstração alternativa, usando uma função energia, pode ser encontrada em [GG23] (que também contém a demonstração com argumentos diretos). Esse resultado também vale para o método do gradiente projetado.

3.3 EDO

Métodos iterativos, especialmente de primeira ordem, podem ser pensados também como discretizações de EDOs, o que às vezes resulta no aproveitamento de argumentos de Lyapunov sobre a convergência de soluções na construção de argumentos sobre a convergência dos próprios métodos. O *Gradient Descent* pode ser visto também como a dicretização explícita, com passo η, dada pelo *método de Euler Forward* ou *método de Euler progressivo* relativo à EDO

$$\frac{\mathrm{d}}{\mathrm{d}t}X(t) = -\nabla f(X(t)). \tag{3.3}$$

A Figura 3.4 mostra a solução dessa EDO com a mesma condição inicial usada para o GD, também sobre as curvas de nível da função objetivo. A curva gerada pela solução é conhecida como *Gradient Flow* ou fluxo do gradiente. Vale notar que aqui, a curva também é perpendicular às curvas de nível da função objetivo.

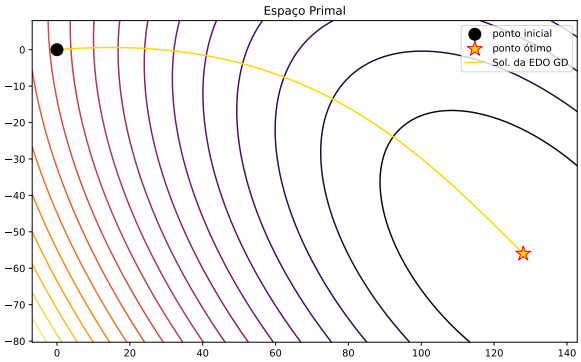

Figura 3.4: Solução da EDO associada ao GD sobre curvas de nível da função objetivo.

Podemos imaginar que a superfície da função representa uma superfície física e aí a solução da EDO pode ser pensada como a trajetória de uma partícula (ou uma bola) descendo essa superfície sob o efeito da gravidade e parando no seu ponto mais baixo.

Quanto à discretização, intuitivamente, sendo $X(t)$ a solução da EDO, podemos aproximar via método das *diferenças finitas* sua derivada por $\frac{X(t+\Delta t)-X(t)}{\Delta t}$ e a própria EDO por $\frac{X(t+\Delta t)-X(t)}{\Delta t} = -\nabla f(X(t))$ ou

$$X(t+\Delta t) = X(t) - \Delta t \nabla f(X(t)).$$

Discretizando também o tempo de forma apropriada, com $t = k\Delta t$ e consequentemente $t+\Delta t = k\Delta t + \Delta t = (k+1)\Delta t$, podemos reescrever a aproximação da EDO como

$$X((k+1)\Delta t) = X(k\Delta t) - \Delta t \nabla f(X(t)).$$

Por fim, se chamarmos $X(k\Delta t)$ de x_k ($X((k+1)\Delta t)$ será x_{k+1}) e Δt de η, obtemos

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

que é o método GD como descrito em (3.2).

De maneira mais rigorosa, usando expansão de Taylor, podemos escrever

$$X(t+\eta) = X(t) + \eta \frac{\mathrm{d}}{\mathrm{d}t}X(t) + o(\eta^2) = X(t) - \eta \nabla f(X(t)) + o(\eta^2).$$

Chamando de novo $X(t)$ de x_k e $X(t+\eta)$ de x_{k+1}, obtemos

$$x_{k+1} = x_k - \eta \nabla f(x_k) + o(\eta^2). \tag{3.4}$$

Quando η tende 0, o termo $o(\eta^2)$ tende a 0 também e aí a EDO se aproxima de

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

que, mais uma vez, é o método GD.

Aqui, η surge naturalmente como o parâmetro que controla linearmente o tamanho dos passos tanto na dimensão temporal quanto na espacial. Essa relação ocorre porque a EDO estabelece uma proporcionalidade simples entre as variações no espaço e suas correspondentes no tempo. Entretanto, no Accelerated Gradient Descent — como exploraremos no próximo capítulo — η mantém seu papel de controlador linear do passo temporal, mas assume uma relação não-linear com o passo espacial, que passa a depender de sua raiz quadrada.

Em [BBF16] e [SH98], prova-se o teorema a seguir que corrobora com a convergência do método para a solução da EDO em um intervalo finito.

Teorema 3.2. *Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função convexa, diferenciável e com gradiente L-Lipschitz. Seja $T > 0$ um tempo finito fixo. Sejam $X : \mathbb{R}_+ \to \mathbb{R}^n$ uma solução de (3.3) com condição inicial $X(0) = x_0$ para algum $x_0 \in \mathbb{R}^n$ e $\{x_k\}_{k\in\mathbb{N}}$ a sequência de pontos gerada pelo GD com tamanho de passo η e $\lfloor T/\eta \rfloor$ passos. Então*

$$\|X(\eta k) - x_k\| \leq \frac{\eta M}{2L}(\exp\{\eta k L\} - 1),$$

onde M é uma constante tal que $\left\| \frac{\mathrm{d}^2}{\mathrm{d}t^2} X(t) \right\| < M$ para todo $t \in R_+$.

Isso significa que com $\eta \to 0$, os pontos produzidos pelo GD se aproximam de seus correspondentes na curva solução da EDO. Vale frisar que isso vale para todo tempo, isto é, para todo T fixo. Assim a diferença de $X(T)$ para algum ponto gerado pelo GD (que varia com $\eta \to 0$) sempre tende a zero conforme refinamos o método.

Além disso, em [Cal20] demonstra-se o teorema a seguir, que estabelece uma taxa de convergência para a solução da EDO, unindo a taxa de convergência do GD à sua convergência para a solução da EDO.

Teorema 3.3. *Sejam $f : \mathbb{R}^n \to \mathbb{R}$ uma função convexa e diferenciável e x^* seu ponto ótimo. Seja ainda $X : \mathbb{R}^+ \to \mathbb{R}^n$ uma solução de (3.3) com condição inicial $X(0) = x_0$ para algum $x_0 \in \mathbb{R}^n$. Então*

$$f(X(t)) - f(x^*) \leq \frac{1}{2t} \|x_0 - x^*\|^2.$$

Observação. *Aqui vimos como o GD representa uma discretização via um método já bem conhecido (de Euler progressivo). O mesmo vai ocorrer no Capítulo 5, mas não nos capítulos Capítulo 4 e Capítulo 6. Nestes dois últimos casos, vamos apresentar uma noção diferente de discretização.*

Capítulo 4

Accelerated Gradient Descent

O Accelerated Gradient Descent (AGD), introduzido por Yuri Nesterov em [Nes83], é um algoritmo que busca superar de maneira eficiente limitações do *Gradient Descent* (como o fato deste não convergir em problemas não convexos, mesmo que relativamente bem comportados) incorporando junto ao gradiente um termo de momento ou inércia, que acaba funcionando como um "impulso".

4.1 Definição e interpretações

O algoritmo funciona tomando-se um ponto viável qualquer $x_0 \in \mathbb{R}^n$, fazendo-se $y_0 = x_0$ e, a cada iteração, atualizando-os seguindo o esquema

$$\begin{cases} x_{k+1} &= y_k - \eta \nabla f(y_k) \\ y_k &= x_k + \mu_k (x_k - x_{k-1}) \end{cases},$$

onde $\mu_k = \frac{k-1}{k+2}$ e $\eta > 0$ é a taxa de aprendizado. As atualizações feitas pelo método é seguem a ordem inversa a apresentada pelo sistema de equações, que prioriza a "importância" de cada uma delas. A Figura 4.1 ilustra o passo dado pelo Método de Nesterov em uma iteração.

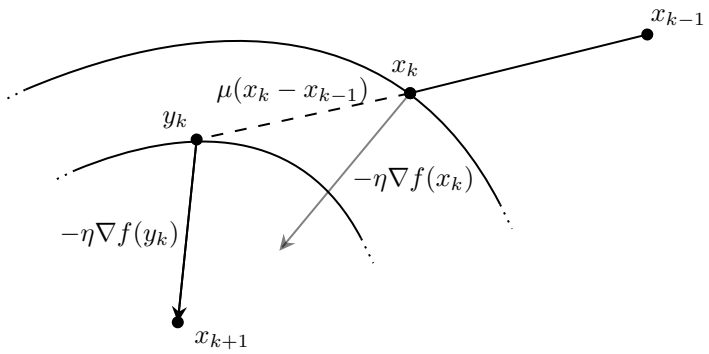

Figura 4.1: Uma iteração do método de Nesterov

A Figura 4.2 mostra a sequência de pontos gerada pelo AGD sobre as curvas de nível da função objetivo de interesse.

O termo $\mu_k(x_k - x_{k-1})$ representa o momento no tempo k e ponto x_k. O ponto y_k que é igual ao ponto x_k mais o termo do momento funciona como uma antecipação de qual deverá ser a próxima atualização baseada em qual foi a atualização anterior. A atualização real se dá então de maneira igual à feita no GD, mas em y_k, a previsão do ponto atualizado, em vez de x_k, o ponto corrente. Esse processo é chamado de *look-ahead step*, ou passo de predição, e é o que promove a aceleração do método. Na Figura 4.2 um efeito do momento pode ser visualizado observando que mesmo já estando próximo do ponto ótimo, os passos continuam sendo

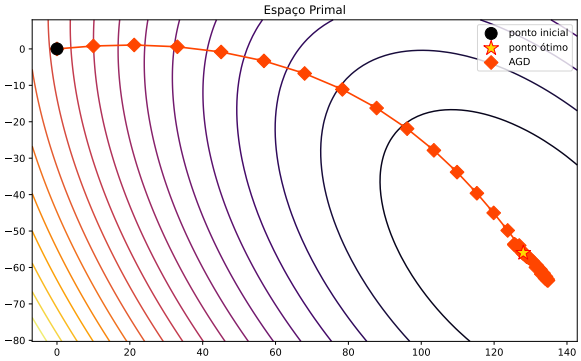

Figura 4.2: Trajetória do AGD sobre curvas de nível da função objetivo.

dados na direção e sentido com que se vinha, como uma partícula sob efeito da inércia; após passar pelo ponto ótimo, os gradientes acumulados enfim anulam o efeito do momento e a trajetória volta a se aproximar do ótimo.

No entanto, como fazemos $y_0 = x_0$, o primeiro passo a ser dado é exatamente igual ao passo dado pelo GD. Além disso, notemos que μ_k começa sendo igual a 0 (na iteração $k = 1$) e cresce sempre se aproximando de 1 conforme aumenta o número de iterações. Dessa maneira, o AGD começa de maneira muito similar ao GD e só depois, quando momento passa a ter um peso maior, se diferencia dele. A Figura 4.3 apresenta as trajetórias do *Gradient Descent* e do Accelerated Gradient Descent, de forma que seja visível sua semelhança, pelo menos nas primeiras iterações.

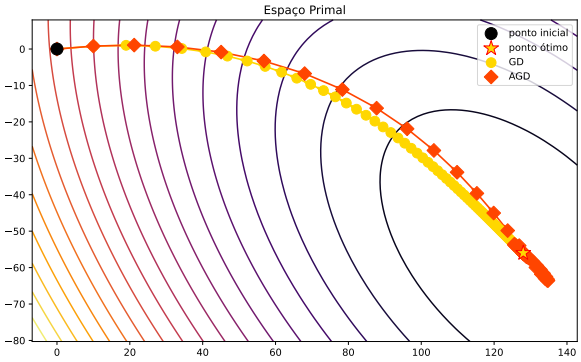

Figura 4.3: Trajetórias do GD e do AGD sobre curvas de nível da função objetivo

O AGD, assim como o GD, pode ser formulado como um algoritmo que resolve um problema de minimização a cada passo seguindo a fórmula

$$x_{k+1} = \arg\min_x \left\{ \eta \langle \nabla f(y_k), x \rangle + \frac{1}{2} \|x - y_k\|^2 \right\},$$

onde y_k é definido de maneira igual à em 4.1. Com isso, o algoritmo também minimiza uma certa aproximação linear de f sem se afastar muito do ponto onde ela é tomada, mas com este ponto sendo o y_k desta vez.

4.2 Accelerated Gradient Descent com Projeção

De forma análoga ao caso do *Gradient Descent*, o *Accelerated Gradient Descent* (AGD) pode ser adaptado para problemas com restrições em conjuntos convexos da forma P' simplesmente inserindo uma etapa de projeção. No esquema de Nesterov, essa projeção é aplicada no ponto de atualização principal. Um exemplo

de iteração projetada é

$$\begin{cases} y_k = x_k + \mu_k(x_k - x_{k-1}) \\ x'_{k+1} = y_k - \eta \nabla f(y_k) \\ x_{k+1} = \Pi_{\mathcal{X}}(x'_{k+1}). \end{cases}$$

Como $\mathcal{X}$ é convexo e y_k é uma combinação convexa de x_k e x_{k-1}, pontos pertencentes a $\mathcal{X}$, temos $y_k \in \mathcal{X}$, não sendo necessária a projeção deste ponto. Ao aplicar o AGD em problemas gerais, ela pode se fazer necessária para o cálculo do gradiente que pode estar definido apenas num conjunto.

Esse método é conhecido como Projected Accelerated Gradient Descent (PAGD) e combina as vantagens da aceleração de Nesterov com a simplicidade do mecanismo de projeção. A Figura 4.4 ilustra os passos dados pelo PAGD em uma iteração.

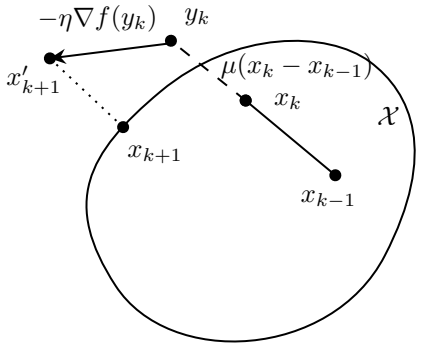

Figura 4.4: Uma iteração do método do gradiente projetado acelerado

Convergência do Método de Nesterov

Sob as mesmas hipóteses feitas para o GD, conseguimos demonstrar a convergência para a sequência $(x_k)_{k\in\mathbb{N}}$ gerada pelo método de Nesterov com passo constante $\eta = \frac{1}{L}$.

Teorema 4.1. *Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função convexa, diferenciável e com gradiente L-Lipschitz. Sejam x^* um ponto ótimo de f e $\{x_k\}_{k\in\mathbb{N}}$ a sequência de pontos gerada pelo AGD com tamanho de passo $1/L$. Então, para todo $k \in \mathbb{N}$,*

$$f(x_k) - f(x^*) \leq \frac{L}{2(k+1)^2} \|x_0 - x^*\|^2 .$$

Esse resultado estabelece uma taxa de convergência em valor funcional da ordem $\mathcal{O}(1/k^2)$, o que representa uma melhora significativa em relação à taxa $\mathcal{O}(1/k)$ do método do gradiente. Essa aceleração foi obtida originalmente por Nesterov [Nes83] e permanece como uma das contribuições mais fundamentais em otimização convexa de primeira ordem. A demonstração também pode ser encontrada em [Cal20] ou [GG23].

Um resultado parecido também vale para a adaptação com projeção.

4.3 EDO e derivação informal

Assim como o GD, o AGD também pode ser pensado como a discretização de uma EDO. Em [SBC16, Seção 2], mostra-se que o AGD representa a discretização da EDO

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} X(t) + \gamma(t) \frac{\mathrm{d}}{\mathrm{d}t} X(t) - \nabla f(X(t)) = 0, \tag{4.1}$$

onde $\gamma(t) = \frac{3}{t}$. A Figura 4.5 mostra a solução dessa EDO com a mesma condição inicial usada para o AGD, também sobre as curvas de nível da função objetivo.

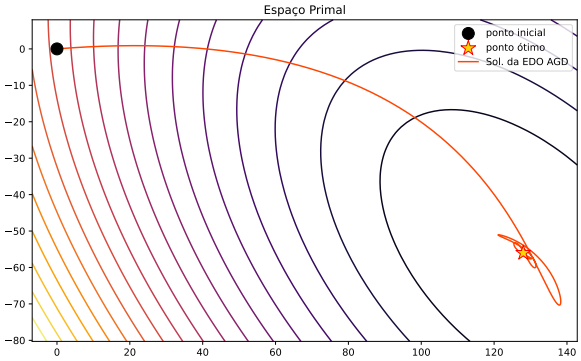

Figura 4.5: Solução da EDO associada ao AGD sobre curvas de nível da função objetivo.

Pensando no gráfico da função como uma superfície física, a curva da solução representa a trajetória de uma partícula com massa sob o efeito da gravidade e de inércia. Com uma pequena adaptação do método de Nesterov, tomando o gradiente no ponto atual em vez do ponto de predição, recuperamos o que é chamado de *Heavy Ball Method*, método do momento ou Polyak momentum ([Pol64]), cuja interpretação é exatamente essa e tem mesma EDO associada (onde s pontos x_k e y_k coincidem).

Nos próximos parágrafos, reproduziremos a derivação informal apresentada em [SBC16]. A ideia geral é supor a existência de uma função do tempo $X(t)$ que produza uma sequência de pontos satisfazendo o esquema de Nesterov quando amostrada em intervalos de tempo iguais, derivar daí uma equação e analisar o comportamento dessa equação no limite em que o intervalo de tempo tende a zero, chegando a uma equação diferencial ordinária (EDO) que representa o caso contínuo do método acelerado.

Seja $\Delta t \in \mathbb{R}_+$ um tamanho de intervalo de tempo qualquer. Suponhamos que exista uma função $X : \mathbb{R}_+ \to \mathbb{R}^n$ de classe C^2, com $\frac{\mathrm{d}}{\mathrm{d}t}X(0) = \vec{0}$ e tal que os pontos iterados da forma $x_k = X(k\Delta t) = X(t)$ satisfaçam o esquema de Nesterov com tamanho de passo $\eta = (\Delta t)^2$. Nosso objetivo é mostrar que, no limite em que $\Delta t \to 0$, a função X satisfaz a EDO apresentada.

Como X é de classe C^2, podemos expandir $X(t + \Delta t)$ em série de Taylor, obtendo

$$X(t + \Delta t) = X(t) + \frac{\mathrm{d}}{\mathrm{d}t}X(t)(\Delta t) + \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}X(t)(\Delta t)^2 + o((\Delta t)^2),$$

que, após um rearranjo nos termos, resulta em

$$\frac{X(t + \Delta t) - X(t)}{\Delta t} = \frac{\mathrm{d}}{\mathrm{d}t}X(t) + \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}X(t)(\Delta t) + o((\Delta t)^2).$$

De maneira análoga, expandindo agora $X(t - \Delta t)$ em série de Taylor, temos

$$\frac{X(t) - X(t - \Delta t)}{\Delta t} = \frac{\mathrm{d}}{\mathrm{d}t}X(t) - \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}X(t)(\Delta t) + o((\Delta t)^2).$$

Agora, de Nesterov, temos $x_{k+1} = y_k - \eta\nabla f(y_k)$ e $y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1})$. Substituindo a segunda equação na primeira, obtemos

$$x_{k+1} = x_k + \frac{k-1}{k+2}(x_k - x_{k-1}) - \eta\nabla f(y_k) \tag{4.2}$$

ou

$$(k+2)(x_{k+1} - x_k) = (k-1)(x_k - x_{k-1}) - (k+2)\eta\nabla f(y_k).$$

Dividindo tudo por $\sqrt{\eta}$ obtemos

$$(k+2)\frac{x_{k+1} - x_k}{\sqrt{\eta}} = (k-1)\frac{x_k - x_{k-1}}{\sqrt{\eta}} - (k+2)\sqrt{\eta}\ \nabla f(y_k).$$

Como $\eta = (\Delta t)^2$, podemos substituir Δt por $\sqrt{\eta}$. Observemos aqui que o esquema anda numa escala de tempo diferente da escala da solução da EDO: em uma iteração com passo η o método avança de $X(k\Delta t)$ para $X((k+1)\Delta t)$, o mesmo que a solução da EDO em um intervalo de tempo Δt. Isso sugere que o método compensa a diferença de escalas, alcançando o mesmo avanço temporal (Δt) com um passo interno menor (η) ou pelo menos quando o tamanho do passo é menor que 1. Além disso, como supomos que X segue o esquema de Nesterov, podemos usar a notação introduzida por X e escrever

$$(k+2)\frac{X(t+\Delta t)-X(t)}{\Delta t} = (k-1)\frac{X(t)-X(t-\Delta t)}{\Delta t} - (k+2)\Delta t\ \nabla f(y_k).$$

Agora, a fim de substituir $\nabla f(y_k)$ por um termo em função de $X(t)$, vamos provar que $\Delta t\nabla f(y_k) = \Delta t\nabla f(X(t)) + o(\Delta t)$. Via expansão de Taylor em $X(t-\Delta t)$, temos

$$X(t-\Delta t) = X(t) - \frac{\mathrm{d}}{\mathrm{d}t}X(t)(\Delta t) + o(\Delta t) \quad \Longrightarrow \quad X(t) - X(t-\Delta t) = \frac{\mathrm{d}}{\mathrm{d}t}X(t)(\Delta t) + o(\Delta t).$$

Dessa maneira, temos

$$y_k = X(t) + \mu_k(X(t) - X(t-\Delta t)) = X(t) + \mu_k\left(\frac{\mathrm{d}}{\mathrm{d}t}X(t)(\Delta t) + o(\Delta t)\right).$$

Usando expansão de Taylor mais uma vez, agora em $\nabla f\left(X(t) + \mu_k\left(\frac{\mathrm{d}}{\mathrm{d}t}X(t)(\Delta t) + o(\Delta t)\right)\right)$, obtemos

$$\begin{aligned}\nabla f(y_k) &= \nabla f\left(X(t) + \mu_k\left(\frac{\mathrm{d}}{\mathrm{d}t}X(t)(\Delta t) + o(\Delta t)\right)\right)\\ &= \nabla f(X(t)) + \nabla^2 f(X(t))\left(\mu_k\left(\frac{\mathrm{d}}{\mathrm{d}t}X(t)(\Delta t) + o(\Delta t)\right)\right).\end{aligned}$$

Notemos que $\frac{\mathrm{d}}{\mathrm{d}t}X(t)$ é constante em relação a Δt, de forma que $\mu_k\frac{\mathrm{d}}{\mathrm{d}t}X(t)(\Delta t) = O(\Delta t)$. Além disso, como ∇f é L-Lipschitz, temos que $\nabla^2 f$ é limitado. Assim

$$\nabla f(y_k) = \nabla f(X(t)) + O(\Delta t).$$

Multiplicando os dois lados da equação por Δt,

$$\Delta t\nabla f(y_k) = \Delta t\nabla f(X(t)) + \Delta t\cdot O(\Delta t) = \Delta t\nabla f(X(t)) + o(\Delta t),$$

como gostaríamos.

Com isso, podemos substituir o termo $\Delta t\nabla f(y_k)$, ficando com

$$\begin{aligned}(k+2)\left(\frac{\mathrm{d}}{\mathrm{d}t}X(t) + \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}X(t)(\Delta t) + o((\Delta t)^2)\right) =& \\ (k-1)\left(\frac{\mathrm{d}}{\mathrm{d}t}X(t) - \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}X(t)(\Delta t) + o((\Delta t)^2)\right)& \\ - (k+2)\left(\Delta t\nabla f(X(t)) + o(\Delta t)\right).&\end{aligned}$$

Após algum desenvolvimento, temos, enfim

$$3\frac{\mathrm{d}}{\mathrm{d}t}X(t) + k\Delta t\frac{\mathrm{d}^2}{\mathrm{d}t^2}X(t) + k\Delta t\nabla f(X(t)) = \Delta t\left(\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}X(t) + 2\nabla f(X(t)) + o(\Delta t)\right).$$

Ou ainda,

$$3\frac{\mathrm{d}}{\mathrm{d}t}X(t) + t\frac{\mathrm{d}^2}{\mathrm{d}t^2}X(t) + t\nabla f(X(t)) = \Delta t\left(\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}X(t) + 2\nabla f(X(t)) + o(\Delta t)\right).$$

Quando o passo η tende a 0, Δt também tende a 0, e o lado direito da equação todo tende a 0. Assim, no limite, temos

$$3\frac{\mathrm{d}}{\mathrm{d}t}X(t) + t\frac{\mathrm{d}^2}{\mathrm{d}t^2}X(t) + t\nabla f(X(t)) = 0.$$

Por fim, dividindo tudo por t, chegamos à equação que define a EDO,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}X(t) + \frac{3}{t}\frac{\mathrm{d}}{\mathrm{d}t}X(t) + \nabla f(X(t)) = 0.$$

Esta derivação evidencia a relação quadrática (ou com respeito à raiz quadrada) entre os incrementos temporais (Δt) da discretização da EDO e incrementos espaciais (η) do esquema de Nesterov.

No que diz respeito à EDO e sua relação com o método de Nesterov, temos dois teoremas centrais, enunciados a seguir.

Teorema 4.2 (Existência e Unicidade)**.** *Sejam $f \in \mathcal{F}_L$ uma função diferenciável L-Lipschitz e $x_0 \in \mathbb{R}^n$ um ponto qualquer. Então a EDO 4.1 com condições iniciais $X(0) = x_0$ e $\frac{\mathrm{d}}{\mathrm{d}t}X(0) = \vec{0}$ tem uma única solução $X \in C^2((0,\infty);\mathbb{R}^n)) \cap C^1([0,\infty);\mathbb{R}^n))$.*

Teorema 4.3 (Convergência do método de Nesterov para a solução da EDO associada)**.** *Sejam $f \in \mathcal{F}_L$ uma função diferenciável L-Lipschitz e $x_0 \in \mathbb{R}^n$ um ponto qualquer. Sejam ainda $X : \mathbb{R}^n_+ \to \mathbb{R}^n$ a solução da EDO* (6.1) *com condições iniciais $X(0) = x_0$ e $\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(0)) = \vec{0}$ e $\{x_k\}_{\mathbb{N}}$ a sequência de pontos gerada pelo método de Nesterov com ponto inicial x_0 e passo s. Então para qualquer $T \in R_+$, vale*

$$\lim_{s\to 0} \sup_{k\in[T/\sqrt{s}]} \left\| x_k - X(k\sqrt{s}) \right\| = 0 \ .$$

As demonstrações destes dois teoremas encontram-se em [SBC16]. No Capítulo 7 apresentaremos demonstrações similares às apresentadas no artigo, aplicadas à EDO e ao método descritos no Capítulo 6, que representam generalizações da EDO e do método descritos neste capítulo.

Aqui, diferentemente de como fizemos para o GD, a noção de discretização se dá através da "distância" da trajetória do método para a solução da EDO. Na verdade, da distância dos pontos gerados pelo método para alguns pontos bem escolhidos da solução da EDO. A ideia fica mais clara quando pensamos no limite, fazendo o passo do método tender a zero e a trajetória tender a uma curva contínua, que corresponde justamente à solução. No Seção B.3, mostramos um método de discretização da EDO correspondente ao AMD que gera os pontos da trajetória do método; como veremos, o AMD vai representar uma generalização do AGD, de forma que essa discretização também vale para o AGD.

Capítulo 5

Mirror Descent

O *Mirror Descent* (MD) foi introduzido por Nemirosky e Yudin em 1983 em [NY83b]. O método apresenta uma generalização do *Gradient Descent* entre outros algoritmos, como o *Multiplicative Weights* ([AHK12]) ou o algoritmo de Hedge ([KKB15]). Os próximos parágrafos justificam o MD seguindo as estruturas e argumentos usados em [Anu20] e [Bub15a].

Como visto no Capítulo 3, o *Gradient Descent* é um algoritmo poderoso e bastante geral, podendo ser aplicado para resolver aproximadamente problemas de minimização de qualquer função convexa sobre qualquer conjunto convexo. No entanto, essa generalidade significa que o GD pode não ser o melhor algoritmo para certas classes de funções ou conjuntos. Para minimizações sobre o simplexo, por exemplo, sabe-se que algoritmos especializados como o de Hedge são significativamente melhores que o GD.

O *Mirror Descent* é um algoritmo que generaliza o *Gradient Descent* quanto à medida de distância usada, fazendo com que seja possível adequar o método a uma possível geometria específica que seja mais adequada para o problema, dependendo, por exemplo, da função objetivo ou do conjunto viável (principalmente). Essa generalização dá-se trocando-se a função usada para medir a distância do ponto atual a qualquer outro - a princípio usamos a norma euclidiana do vetor que leva o ponto atual a este qualquer outro - pela chamada Divergência de Bregman. A Divergência de Bregman, como veremos a seguir, generaliza o conceito de distância, usando uma função convexa chamada função geradora de distâncias.

Ainda, o *Mirror Descent* generaliza o Gradient Descent, fazendo com que seja possível resolver aproximadamente problemas de otimização convexa em espaços de Banach. O GD é definido em espaços de Hilbert. Isso quer dizer que podemos usar o MD em contextos em que a norma utilizada não depende de um produto interno, como acontece com a norma ℓ_1, por exemplo. Nesses casos, o GD simplesmente não faz sentido, uma vez que não se pode somar um gradiente, que é um elemento do espaço dual, a um ponto do espaço primal, como acontece quando se faz $x_k - \eta \nabla f(x_k)$. Isso não era um problema anteriormente, pois em um espaço de Hilbert $\mathcal{H}$, graças ao Teorema da Representação de Riesz, o espaço dual $\mathcal{H}^*$ é isomorfo a $\mathcal{H}$.

Nemirovski e Yudin perceberam que ainda é possível realizar o passo do gradiente, mas aplicando-o a um ponto do espaço dual. Dado um ponto primal, podemos mapeá-lo para um ponto dual, atualizar esse ponto dual com o gradiente e mapear o ponto atualizado de volta para o espaço primal. Caso o ponto resultante estiver fora do conjunto viável, o projetamos de volta. Tanto esse mapeamento quanto a projeção são baseados nas ideias de divergência de Bregman e mapeamento espelho, que definiremos na próxima seção. O termo *Mirror* foi empregado por conta da visão do espaço dual como uma imagem espelhada do espaço primal.

Este capítulo apresenta as definições da divergência de Bregman e do mapeamento espelho, bem como alguns resultados associados a eles. Em seguida, apresenta a definição do *Mirror Descent* por uma abordagem análoga à usada para o Gradient Descent. Por fim, apresenta a EDO cuja discretização corresponde ao Mirror Descent.

A partir daqui vamos deixar de escrever $\langle \nabla f(x), y \rangle$ onde x e y são pontos primais e passaremos a escrever $\nabla f(x)(y)$, uma vez que estaremos pensando na derivada como operador linear. A exceção será quando estivermos falando de casos específicos, como o euclidiano, mas deixaremos isso claro.

5.1 Preliminares para Mirror Descent

Nesta seção apresentamos as definições da divergência de Bregman e Projeção de Bregman, necessárias para a definição do *Mirror Descent*, bem como alguns resultados associados a elas.

5.1.1 Divergência de Bregman

Consideremos a definição da Divergência de Bregman.

Definição (Divergência de Bregman)**.** *Dada uma função $\Phi : \mathcal{D} \to \mathbb{R}$ diferenciável e estritamente convexa, a* ***divergência de Bregman*** *de x para y com respeito a Φ é definida por*

$$D_\Phi(x, y) = \Phi(x) - [\Phi(y) + \nabla\Phi(y)(x - y)].$$

Em alguns contextos, a função Φ é chamada de função geradora de distâncias. Nesse sentido, a divergência de Bregman é uma espécie de generalização de uma função de distância. Quando $\Phi(x) = \|x\|_2^2$, por exemplo, temos $\nabla\Phi(x) = 2x$ e

$$\begin{aligned} D_\Phi(x, y) &= \|x\|_2^2 - \|y\|_2^2 - \langle 2y, x\rangle + \langle 2y, y\rangle \\ &= \|x\|_2^2 - \|y\|_2^2 - \langle 2y, x\rangle + 2\|y\|_2^2 \\ &= \|x\|_2^2 - 2\langle y, x\rangle + \|y\|_2^2 \\ &= \langle y - x, y - x\rangle = \|y - x\|_2^2, \end{aligned}$$

a distância euclidiana. A função Φ também recebe o nome de função de Bregman e pode ser pensada como um potencial.

Por Φ ser estritamente convexa, segue que para quaisquer $x, y \in \mathbb{R}^n$, $D_\Phi(x, y) \geq 0$ ou ainda que $D_\Phi(x, y) = 0$ se, e somente se, $x = y$, entre outras propriedades associadas ao conceito de distância. Dessas duas propriedades citadas, segue também que $D_\Phi(x, y)$ atinge seu mínimo global em y, onde tem valor 0. A divergência de Bregman (como outras divergências) se parece com uma métrica, mas não satisfaz a desigualdade triangular nem é simétrica de maneira geral.

Geometricamente, pensando na divergência de Bregman como uma função de x para cada y, ela é a diferença da função Φ avaliada em x para a aproximação linear de Φ tomada em y avaliada em x; a Figura 5.1 ilustra isso para uma função bidimensional. Vale a pena notar que, sendo uma diferença de uma função estritamente convexa e uma função linear, a divergência de Bregman também é uma função estritamente convexa. Mais do que isso, a divergência de Bregman se assemelha à sua própria função geradora de distância no sentido de que seus gradientes diferem apenas por um vetor constante, $\nabla\Phi(y)$; em termos de segunda derivada, elas são iguais. Essa ideia é expressa pelo fato a seguir.

Fato 5.1. *Seja $\Phi : \mathcal{D} \to \mathbb{R}$ uma função diferenciável e estritamente convexa e sejam $x, y \in \mathcal{D}$ quaisquer. Então*

$$\frac{\partial}{\partial x} D_\Phi(x, y) = \nabla\Phi(x) - \nabla\Phi(y).$$

Demonstração. Basta acompanhar o desenvolvimento a seguir.

$$\begin{aligned} \frac{\partial}{\partial x} D_\Phi(x, y) &= \frac{\partial}{\partial x}[\Phi(x) - \Phi(y) - \nabla\Phi(y)(x - y)] \\ &= \frac{\partial}{\partial x}[\Phi(x) - \Phi(y) - \nabla\Phi(y)(x) + \nabla\Phi(y)(y)] \\ &= \nabla\Phi(x) - 0 - \nabla\Phi(y) + 0 = \nabla\Phi(x) - \nabla\Phi(y). \end{aligned}$$

■

Usaremos este fato para encontrar uma formulação explícita para o MD mais adiante.

A divergência de Bregman vai representar uma adaptação à geometria do problema, especialmente quando fizer sentido usar normas diferentes da euclidiana para as variáveis envolvidas. É o caso de otimização sobre o simplexo, como veremos ao final deste capítulo.

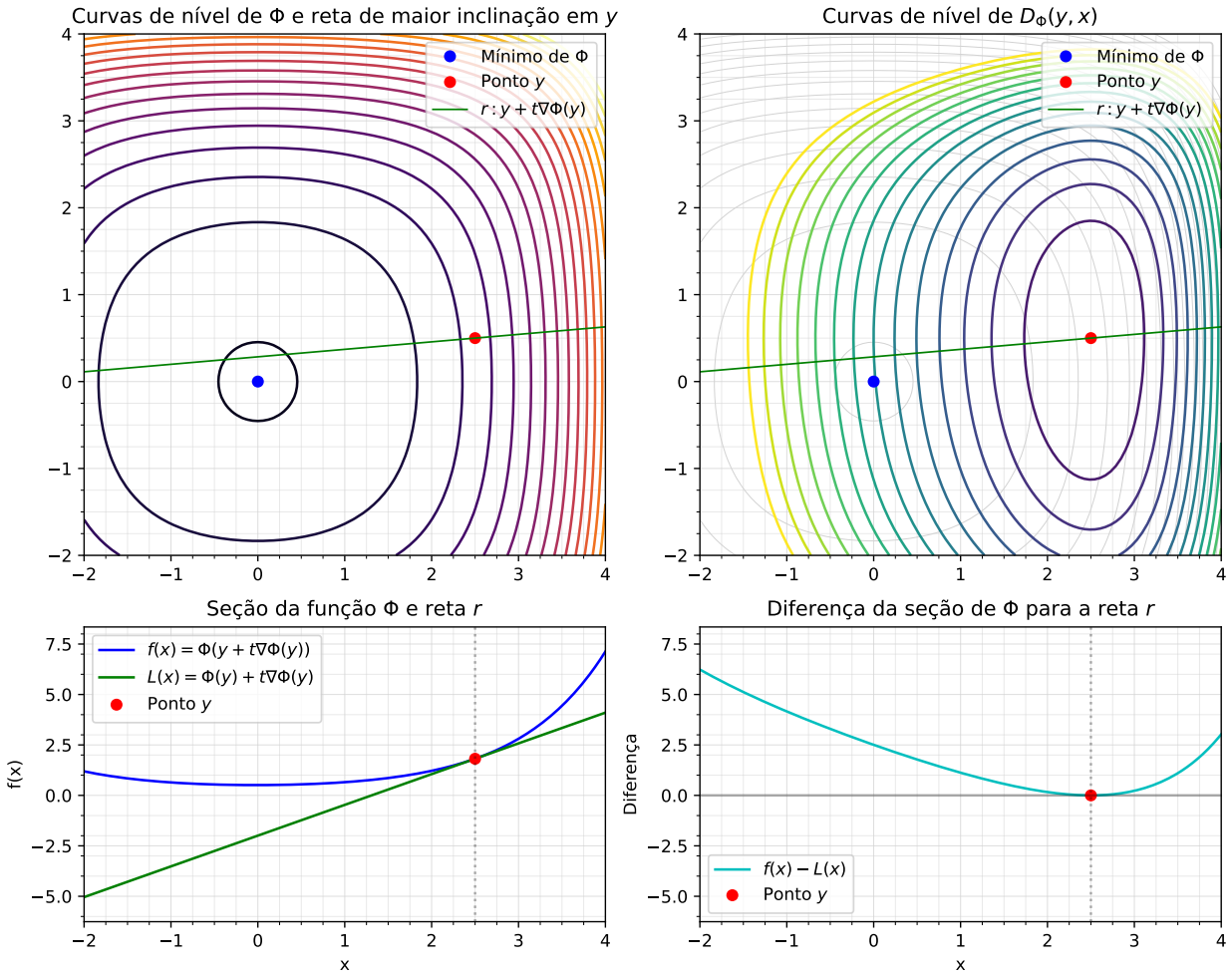

Figura 5.1: Curvas de nível da função Φ (canto superior esquerdo); curvas de nível da divergência de Bregman $D_\Phi(x, y)$ (canto superior direito); seções das função Φ e sua linearização em y (canto inferior esquerdo); seção da diferença entre a função Φ e sua linearização em y (canto inferior direito).

5.1.2 Projeção de Bregman

Assim como fizemos nos Capítulo 3 e Capítulo 4, podemos adaptar o *Mirror Descent* em problemas restritos a conjuntos convexos, fechados e não vazios(como o simplexo) através de projeções. No caso do MD, para aproveitar a adaptação à geometria do problema possibilitada pela divergência de Bregman, vamos definir também uma projeção induzida pela função de Bregman.

Definição (Projeção de Bregman)**.** *Seja $\Phi : \mathcal{D} \to \mathbb{R}$ uma função diferenciável e estritamente convexa e seja $\mathcal{X}$ convexo e compacto, tal que $\mathcal{X} \subseteq \overline{\mathcal{D}}$. A* ***projeção de Bregman em $\mathcal{X}$ com respeito a Φ*** *é dada por*

$$\Pi_{\mathcal{X}}^{\Phi}(y) = \arg\min_{x \in \mathcal{X} \cap \mathcal{D}} D_{\Phi}(x, y).$$

Como a divergência de Bregman é interpretada como uma distância, a projeção de Bregman de um ponto y em um conjunto $\mathcal{X}$ é então o ponto em $\mathcal{X}$ mais próximo (cuja distância é dada pela divergência de Bregman) do ponto y. A Figura 5.2 mostra a projeção de Bregman de um ponto y em um conjunto poligonal sobre as curvas de nível da divergência de Bregman. Lembremo-nos de que divergência de Bregman não é simétrica em geral. A escolha de minimizar $D_{\Phi}(x, y)$ e não $D_{\Phi}(y, x)$ se justifica pela simplicidade, tanto em interpretação quanto computação. O teorema a seguir evidencia isso.

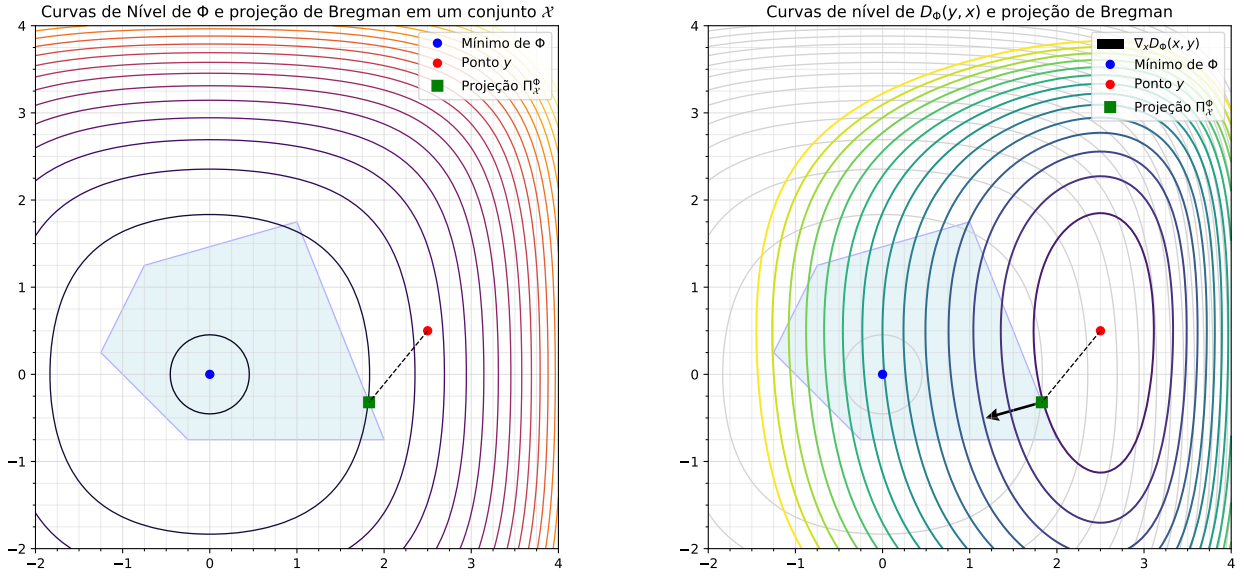

Figura 5.2: Curvas de nível de uma função Φ e conjunto convexo $\mathcal{X}$ (à esquerda) e curvas de nível da divergência de Bregman em relação a um ponto y, conjunto $\mathcal{X}$ e a projeção de Bregman em $\mathcal{X}$ de y (à direita)

Teorema 5.2 (existência e unicidade da projeção)**.** *Seja $\Phi : \mathcal{D} \to \mathbb{R}$ uma função diferenciável e estritamente convexa e seja $\mathcal{X}$ um conjunto fechado e convexo tal que $\mathcal{X} \subseteq \overline{\mathcal{D}}$. Se $\lim_{x \to \partial\mathcal{D}} \|\nabla\Phi(x)\|_* = +\infty$, então para qualquer $y \in \mathcal{D}$, a projeção de Bregman de y em $\mathcal{X}$ com respeito a Φ existe e é única.*

Demonstração. Como $\|\nabla\Phi(x)\|_*$ diverge conforme x se aproxima do bordo de $\mathcal{D}$, a função $D_{\Phi}(x, y)$ também deve divergir conforme x se aproximar do bordo de $\mathcal{D}$. Dessa forma, toda sequência de pontos convergindo para o mínimo de $D_{\Phi}(x, y)$ em $\mathcal{X} \cap \mathcal{D}$ não pode estar se aproximando do bordo de $\mathcal{D}$ e deve, portanto, convergir para um ponto no interior de $\mathcal{D}$. Toda sequência dessa forma também não pode estar se aproximando de um ponto que que não está em $\mathcal{X} \cap \mathcal{D}$ (pois como $\mathcal{X}$ é fechado, todo ponto dessa forma deve estar no bordo de $\mathcal{D}$ e já vimos que a convergência deve ser para o interior de $\mathcal{D}$). Assim, toda sequência em $\mathcal{X} \cap \mathcal{D}$ convergindo para o mínimo de $D_{\Phi}(x, y)$ converge para um ponto em $\mathcal{X} \cap \mathcal{D}$, o que garante a existência da projeção.

Como a $D_{\Phi}(x, y)$ é estritamente convexa, essa projeção é única. ■

De maneira mais intuitiva, levando em consideração conjuntos de nível da divergência de Bregman para um ponto y, a projeção de y em um conjunto $\mathcal{X}$ é o ponto em $\mathcal{X}$ que pertence ao conjunto de nível de menor valor dentre os que intersectam o $\mathcal{X}$; ver Figura 5.2. O fato de Φ divergir na fronteira de $\mathcal{D}$, implica que todo conjunto de nível de valor finito está no interior de $\mathcal{D}$ de forma que esse ponto está, afinal, bem definido. Além disso, o fato de Φ ser estritamente convexa implica que seus conjuntos de nível são fronteiras de conjuntos estritamente convexos, de forma que a menor interseção não vazia com $\mathcal{X}$, um conjunto convexo e fechado, é um único ponto.

A propriedade $\lim_{x\to\partial\mathcal{D}} \|\nabla\Phi(x)\|_* = +\infty$ é também chamada de coercividade na fronteira de $\mathcal{D}$. Em [Ius95], Iusem prova que o fato da projeção não estar na fronteira de $\mathcal{D}$ é devido ao subdiferencial de uma divergência de Bregman coerciva ser vazio na fronteira de $\mathcal{D}$.

O próximo lema estabelece uma espécie de teorema de separação para a projeção de Bregman.

Lema 5.3. *Seja $\Phi : \mathcal{D} \to \mathbb{R}$ uma função diferenciável e estritamente convexa. Sejam $x \in \mathcal{X} \cap \mathcal{D}$, $y \in \mathcal{D}$ e $y^* = \Pi^\Phi_\mathcal{X}(y)$. Então*

$$(\nabla\Phi(y^*) - \nabla\Phi(y))(y^* - x) \leq 0.$$

Demonstração. Como $y^* = \Pi^\Phi_\mathcal{X}(y) \in \arg\min_{x\in\mathcal{X}\cap\mathcal{D}} D_\Phi(x, y)$, pelo fato 2.14, temos que para todo $x \in \mathcal{X}\cap\mathcal{D}$, $\nabla D_\Phi(y^*, y)(y^* - x) \leq 0$. Agora, pelo fato 5.1, temos então que $(\nabla\Phi(y^*) - \nabla\Phi(y))(y^* - x) \leq 0$. ■

Aqui o vetor $\nabla\Phi(y^*) - \nabla\Phi(y) = \nabla_{y^*} D_\Phi(y^*, y)$ (via fato 5.1) é o vetor normal ao conjunto de nível que contém a projeção y^* (ver Figura 5.2). Este lema estabelece que o vetor que sai se qualquer ponto $x \in \mathcal{X}$ e vai até o ponto $y*$ está negativamente correlacionado com o vetor normal a esse conjunto. Isso quer dizer que esse vetor é normal a um hiperplano que define dois semiespaços, um deles contendo totalmente o conjunto $\mathcal{X}$. Nesse sentido, para $y \notin \mathcal{X}$, esse resultado é uma espécie de teorema do plano separador.

O fato a seguir é conhecido como identidade dos três pontos.

Fato 5.4. *Seja $\Phi : \mathbb{R}^n \to \mathbb{R}$ uma função diferenciável e estritamente convexa. Então para quaisquer $x, y, z \in \mathbb{R}^n$, vale que*

$$D_\Phi(x, y) + D_\Phi(z, x) - D_\Phi(z, y) = (\nabla\Phi(x) - \nabla\Phi(y))(x - z).$$

Demonstração.

$$\begin{aligned} D_\Phi(x, y) + D_\Phi(z, x) - D_\Phi(z, y) &= \Phi(x) - \Phi(y) + \Phi(z) - \Phi(x) - \Phi(z) + \Phi(y) \\ &\quad + \nabla\Phi(y)(x - y) + \nabla\Phi(x)(z - x) - \nabla\Phi(y)(z - y) \\ &= \nabla\Phi(y)(x - y - z + y) + \nabla\Phi(x)(z - x) \\ &= \nabla\Phi(y)(x - z) - \nabla\Phi(x)(x - z) \\ &= (\nabla\Phi(y) - \nabla\Phi(x))(x - z) \end{aligned}$$ ■

Este fato e o Lema anterior a ele resultam no seguinte corolário que estabelece um caso especial para o qual afinal vale a desigualdade triangular para a divergência de Bregman.

Corolário 5.5. *Seja $\Phi : \mathcal{D} \to \mathbb{R}$ uma função diferenciável e estritamente convexa. Sejam $x \in \mathcal{X} \cap \mathcal{D}$, $y \in \mathcal{D}$ e $y^* = \Pi^\Phi_\mathcal{X}(y)$. Então*

$$D_\Phi(x, y^*) + D_\Phi(y^*, y) \leq D_\Phi(x, y).$$

Demonstração. Pelo Lema 5.3, temos que $(\nabla\Phi(y^*) - \nabla\Phi(y))(y^* - x) \leq 0$.

Pelo fato 5.4, temos que $(\nabla\Phi(y^*) - \nabla\Phi(y))(y^* - x) = D_\Phi(x, y^*) + D_\Phi(y^*, y) - D_\Phi(x, y)$.

Assim, $D_\Phi(x, y^*) + D_\Phi(y^*, y) - D_\Phi(x, y) \leq 0$ e portanto $D_\Phi(x, y^*) + D_\Phi(y^*, y) \leq D_\Phi(x, y)$. ■

5.2 Definição e interpretações

Da mesma maneira como definimos o *Gradient Descent* pela primeira vez, o *Mirror Descent* pode ser definido via uma abordagem de ponto proximal resolvendo-se, a cada iteração, o problema

$$x_{k+1} = \arg\min_x \{\eta\nabla f(x_k)(x) + D_\Phi(x, x_k)\}, \tag{5.1}$$

onde, mais uma vez, $\eta > 0$ é a chamada de taxa de aprendizado. Por conta dessa formulação, o método é chamado também de método do ponto proximal para divergência de Bregman ou apenas método de Bregman. A função minimizada pelo MD difere da função otimizada pelo GD apenas na função usada para medir a distância entre o ponto atual e os candidatos a atualizações. Se $\Phi(x) = \|x\|_2^2$, de forma praticamente igual à feita na Subseção 5.1.1, temos que as funções distância do GD e do MD são iguais e os métodos são equivalentes.

Fazer $x_{k+1} = x_k + \alpha \nabla f(x_k)$, como visto, não faz sentido. Ainda assim, como $\nabla f(x_k)$ é um operador linear, minimizar o termo $\eta \nabla f(x_k)(x)$ com um tamanho de passo fixo α envolve andar na direção de um certo vetor v_k, isto é, fazer $x_{k+1} = x_k + \alpha v_k$. Entretanto, afastar-se de x_k também aumenta o termo $D_\Phi(x, x_k)$, mas dessa vez de forma que a direção desse afastamento tenha mais importância; em comparação, a distância usada no GD é indiferente à direção, o que é bastante particular, na verdade. A atualização a ser feita nesse caso não é tão óbvia: pode ser que, andando na direção de v_k, a divergência de Bregman aumente muito, fazendo com que o tamanho do passo que podemos dar acabe sendo pequeno, enquanto que, andando em uma outra direção, a divergência aumente pouco, permitindo-nos dar passos maiores, o que pode compensar o fato de não estarmos indo na direção do vetor v_k (de norma 1) minimizador de $\nabla f(x_k)v$. Felizmente o MD também tem uma formulação um pouco mais explícita sobre a solução do problema de minimização a ser resolvido, como veremos a seguir.

Seja $g_k(x) = \eta \nabla f(x_k)(x) + D_\Phi(x, x_k)$ a função a ser minimizada em cada passo do esquema. Seguindo a mesma lógica apresentada no Capítulo 3 Seção 3.1, minimizar $g_k(x)$ é o mesmo que minimizar $(\nabla f(x_k)(x) - f(x_k)) + (D_\Phi(x, x_k) + f(x_k))$. O primeiro termo dessa expressão é a aproximação linear de f no ponto x_k. O segundo termo é uma penalização do distanciamento de x_k. Sendo assim, o MD é também um método que busca minimizar a aproximação de f sem se afastar muito do ponto atual, mas dessa vez considerando uma geometria possivelmente diferente da usual.

Usando novamente o fato 5.1, o ponto x_{k+1} que minimiza g_k no passo k deve satisfazer a condição de otimalidade

$$\begin{aligned} 0 = \frac{\partial g_k}{\partial x_{k+1}}(x_{k+1}) &= \frac{\partial}{\partial x_{k+1}}(\eta \nabla f(x_k)(x_{k+1}) - D_\Phi(x_{k+1}, x_k)) \\ &= \eta \nabla f(x_k) + \nabla \Phi(x_{k+1}) - \nabla \Phi(x_k). \end{aligned}$$

Ou seja, em cada passo x_{k+1} deve ser tal que

$$\nabla \Phi(x_{k+1}) = \nabla \Phi(x_k) - \eta \nabla f(x_k).$$

Chamando $\nabla \Phi(x_k)$ de θ_k de forma que enfatizemos que estes pontos pertencem ao espaço dual de $\mathbb{R}^n$, escrevemos

$$\theta_{k+1} = \theta_k - \eta \nabla f(x_k).$$

Notemos que este é exatamente o passo que é dado no GD na iteração k, mas desta vez, feito no espaço dual, ao qual de fato $\nabla f(x_k)$ pertence. Neste sentido, o MD é essencialmente o GD com a garantia de que, ao somar dois elementos, estes pertençam ao mesmo espaço.

A transformação de elementos do espaço primal em elementos do espaço dual e vice-versa aparece naturalmente pelo gradiente da Φ, a função geradora de distâncias que usamos para a divergência de Bregman. Dada a condição de otimalidade e a importância de $\nabla \Phi$ para ela, introduzimos a seguir os conceitos de função potencial e mapeamento espelho ou *mirror map*.

5.2.1 Mirror Map

Definição (Mapeamento espelho)**.** *Uma função $\Phi : \mathcal{D} \to \mathbb{R}$ é chamada de **função potencial** se satisfaz as seguintes propriedades:*

- *Φ é diferenciável e estritamente convexa em $\mathcal{D}$;*
- *$\nabla \Phi : \mathcal{D} \to \mathbb{R}^{n*}$ é sobrejetiva, ou seja, $\nabla \Phi(\mathcal{D}) = \mathbb{R}^{n*}$;*
- *A norma do gradiente diverge quando x se aproxima da fronteira de $\mathcal{D}$, isto é,*

$$\lim_{x \to \partial \mathcal{D}} \|\nabla \Phi(x)\|_* = \infty.$$

Neste caso, a aplicação $\nabla\Phi$ é denominada ***mapeamento espelho*** *(ou* ***mirror map****).*

É comum assumir (às vezes na definição mesmo) que a função potencial seja fortemente convexa em vez de só estritamente convexa. A convexidade forte implica na convexidade estrita, como já vimos, e possibilita a obtenção de alguns resultados importantes. Nesta dissertação, faremos essa hipótese em alguns momentos.

O passo dado pelo MD é feito no espaço dual, mas o problema se dá no espaço primal. Para atualizar o ponto x_k, devemos portanto encontrar uma maneira de remapear o ponto atualizado θ_{k+1} para o espaço primal. O fato de Φ ser estritamente convexa faz com que ela seja injetiva, garantindo a existência da inversa. Assim, podemos escrever $x_{k+1} = (\nabla\Phi)^{-1}(\theta_{k+1})$ tranquilamente; a inversa de $\nabla\Phi$ é às vezes denotada por $\nabla\Phi^*$ e chamada de conjugada de $\nabla\Phi$. Isto nos fornece a formulação mais comum do MD,

$$\begin{cases} x_{k+1} &= (\nabla\Phi)^{-1}(\theta_{k+1}) \\ \theta_{k+1} &= \nabla\Phi(x_k) - \eta\nabla f(x_k). \end{cases} \tag{5.2}$$

Aqui mais uma vez as equações estão apresentadas em ordem de importância, mesmo que a ordem de aplicação delas em cada passo seja a ordem contrária a essa. A Figura 5.3 ilustra os passos dados pelo MD em uma iteração.

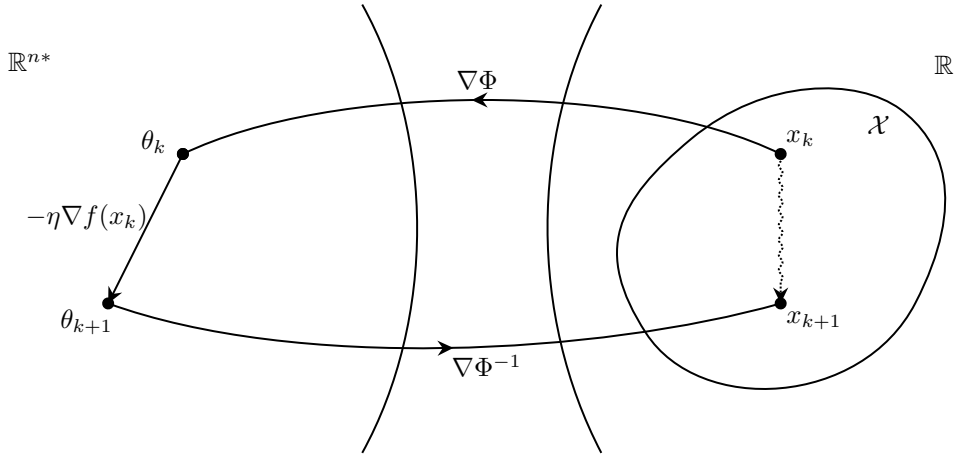

Figura 5.3: Uma iteração do *Mirror Descent*

Se por ventura o ponto gerado pelo MD estiver fora do conjunto viável, o projetamos de volta usando a projeção de Bregman. Nesse caso o método é descrito por

$$\begin{cases} x_{k+1} &= \Pi^{\Phi}_{\mathcal{X}}(x'_{k+1}) \\ x'_{k+1} &= (\nabla\Phi)^{-1}(\theta_{k+1}) \\ \theta_{k+1} &= \nabla\Phi(x_k) - \eta\nabla f(x_k). \end{cases}$$

Sob a hipótese de $\mathcal{X}$ compacto (fechado e limitado), esse método é equivalente ao método do ponto proximal definido por

$$x_{k+1} = \arg\min_{x\in\mathcal{X}}\{\eta\nabla f(x_k)(x) + D_\Phi(x, x_k)\}.$$

A Figura 5.4 ilustra os passos do método neste caso.

A Figura 5.5 mostra as trajetórias do GD e do MD sobre as curvas de nível da função objetivo $f(x)$ no espaço primal, e da função objetivo $f((\nabla\Phi)^{-1}(\theta))$ no espaço dual, bem como o campo gradiente de Φ no espaço primal. Para a confecção dessa imagem e da imagem que ilustra a trajetória do AMD no próximo capítulo, bem como as imagens que ilustram as soluções das EDOs associadas ao MD e AMD, a função que induziu o mapeamento espelho foi $\Phi(x, y) = x^2 + y^2 + \frac{3}{2}xy$. Aqui é um certo abuso ilustrar os gradientes de Φ no espaço primal, mas é como normalmente visualizamos um campo gradiente; podemos pensar nesses vetores como *proxys* para as direções de maior subida da *Phi* e, por extensão, de D_Φ. Notemos que a trajetória do MD no espaço primal parece dar preferência para direções onde os vetores gradientes

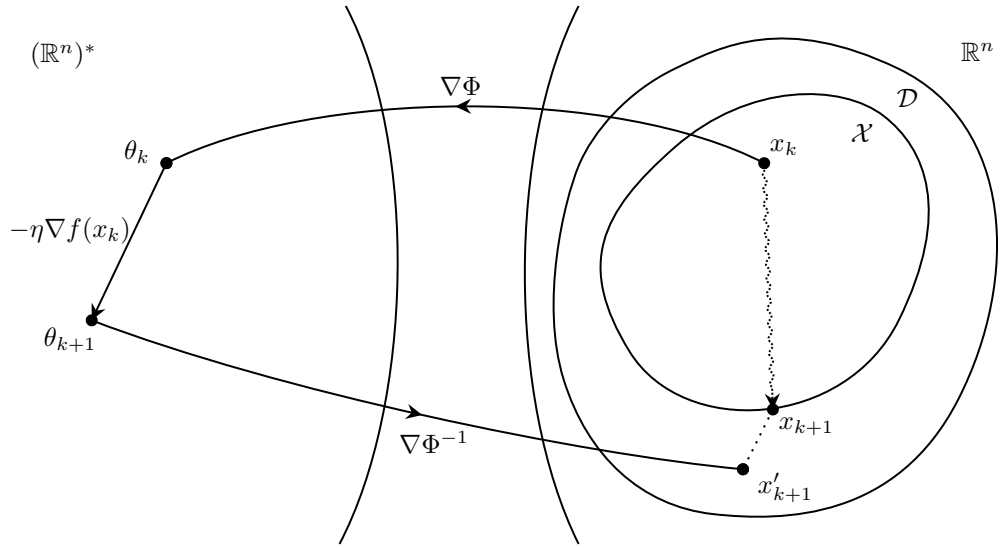

Figura 5.4: Uma iteração do método do *Projected Mirror Descent*

$\nabla\Phi(x)$ são menores. Isso ocorre porque são nessas direções que a divergência de Bregman é menor e passos maiores podem ser dados. Esse fenômeno se manifesta mais nas primeiras iterações, pois depois de um dado momento, os vetores que minimizam os termos relativos às minimizações da linearização da f e à distância ao ponto atual se tornam menos correlacionadas, fazendo com que ir na direção de v_k não compense mais o desvio do gradiente de f.

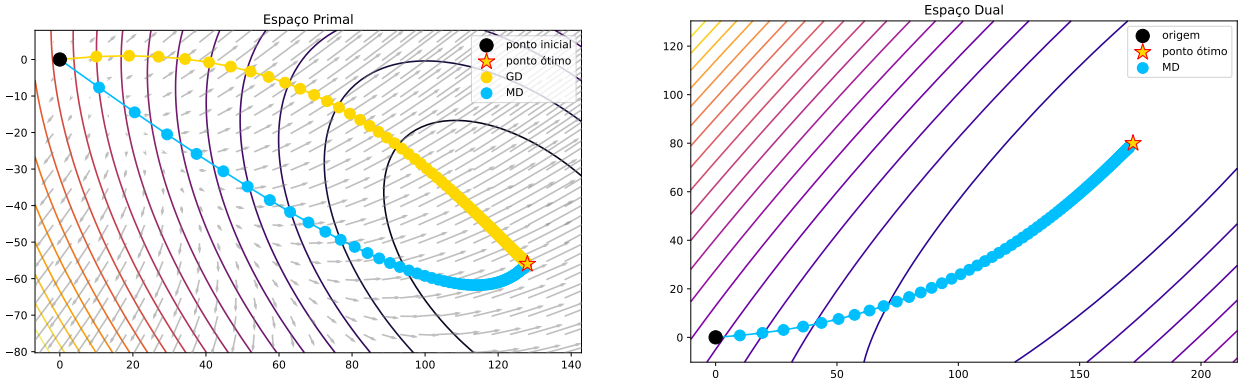

Figura 5.5: Trajetórias dos GD e MD sobre curvas de nível da função objetivo.

O passo dado pelo MD é sempre positivamente correlacionado com o oposto do gradiente de f de forma que os passos dados pelo GD e pelo MD partindo de um mesmo ponto são sempre positivamente correlacionados. Na Figura 5.5 isso pode ser observado no ponto inicial. Isso acontece porque o vetor de máximo decrescimento de $\nabla f(x_k)$, digamos u_k, é ortogonal a um hiperplano que divide o espaço em dois semiespaços, um para o qual todo passo para seu interior representa um decrescimento na f (H^-) e outro para o qual todo passo para seu interior representa um crescimento (H^+). Um passo dado pelo MD para H^+ não teria nada a ganhar: aumenta o termo relativo à linearização da f e aumenta a distância ao ponto atual. Portanto, resta dar um passo para H^-. Este fato é formalizado a seguir, lembrando que para todo $k \in \mathbb{N}$, $\nabla\Phi(x_{k+1}) - \nabla\Phi(x_k) = -\eta\nabla f(x_k)$, o passo dado pelo GD.

Teorema 5.6. *Seja $x_k \in \mathbb{R}^n$ um ponto qualquer qualquer. Então*

$$(\nabla\Phi(x_{k+1}) - \nabla\Phi(x_k))(x_{k+1} - x_k) \geq 0.$$

A demonstração deste teorema se encontra no Apêndice A, em A.1.

5.3 Convergência do *Mirror Descent*

Para o *Mirror Descent*, sob as hipóteses de convexidade de f, forte convexidade de Φ, e compacidade de $\mathcal{X}$, é possível estabelecer taxas de convergência de $O(1/\sqrt{k})$ para funções não suaves e $O(1/k)$ quando ∇f é Lipschitz contínua [Bub15b].

Resultados explícitos são encontrados principalmente em três contextos especializados. Em otimização online, [Haz16] demonstra *regret* $\mathcal{O}(\sqrt{\log k/k})$ com passos $\eta_k \propto 1/\sqrt{k}$ sob forte convexidade de Φ. Quando $\mathcal{X}$ é o simplexo unitário e Φ é a entropia negativa, [JN11] estabelece convergência $\mathcal{O}(\log k/k)$ explorando propriedades de auto-concordância. No caso euclidiano ($\Phi = \frac{1}{2}\|\cdot\|_2^2$), recupera-se a descida projetada clássica com convergência $\mathcal{O}(1/k)$ para funções suaves conforme demonstrado por [Nes18].

A aparente lacuna na literatura para a análise direta do MD decorre de fatores estruturais: a não linearidade da projeção $\Pi_{\mathcal{X}}^{\Phi}$ introduz complexidades analíticas adicionais comparativamente à formulação proximal, e aplicações práticas frequentemente requerem adaptações como passos variáveis ou regularizações. Consequentemente a análise de convergência é tipicamente conduzida através das formulações equivalentes ou em domínios estruturados, como o que será explorado no Capítulo 8.

5.4 EDO

Uma vez que o passo dado pelo MD no espaço dual é o mesmo que o passo dado pelo GD no espaço primal, podemos deduzir que a EDO cuja discretização corresponde ao MD é, de forma parecida com a do GD,

$$\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t)) = -\nabla f(X(t)), \tag{5.3}$$

onde $\Theta(X(t)) = \nabla\Phi(X(t))$. Na verdade, no trabalho seminal de Nemirovsky e Yudin ([NY83a]), o Mirror Descent é definido a partir desta EDO.

Apesar de ambos os lados da equação que define a EDO serem pontos no espaço dual, uma solução desta EDO é uma curva $X : \mathbb{R}_+ \to \mathbb{R}^n$ no espaço primal satisfazendo a equação. A Figura 5.6 mostra as trajetórias de $X(t)$ no espaço primal e de $\Theta(X(t))$ no espaço dual.

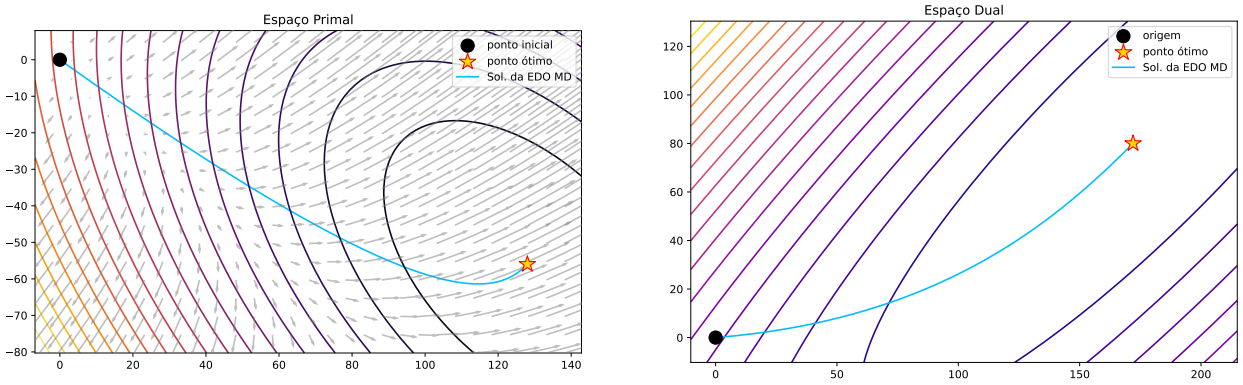

Figura 5.6: Trajetória do MD e solução da EDO associada sobre curvas de nível da função objetivo.

Da mesma maneira que fizemos no Capítulo 3, podemos aproximar a derivada de $\Theta(X(t))$ por $\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t)) = \frac{\Theta(X(t+\Delta t))-\Theta(X(t))}{\Delta t}$ e aí aproximar a EDO por $\frac{\Theta(X(t+\Delta t))-\Theta(X(t))}{\Delta t} = -\nabla f(X(t))$ ou mesmo

$$\Theta(X(t+\Delta t)) = \Theta(X(t)) - \Delta t\nabla f(X(t)).$$

Agora, pensando novamente no tempo de maneira discreta com $t = k\Delta t$ e $t+\Delta t = (k+1)\Delta t$, escrevemos

$$\Theta(X((k+1)\Delta t)) = \Theta(X(k\Delta t)) - \Delta t\nabla f(X(k\Delta t)).$$

Por fim, pensando que $\Theta(X(k\Delta t)) = \nabla\Phi(X(k\Delta t)) = \nabla\Phi(x_k) = \theta_k$, $\Theta(X((k+1)\Delta t)) = \theta_{k+1}$ e $\Delta t = \eta$, temos que

$$\theta_{k+1} = \theta_k - \eta\nabla f(x_k).$$

Isto, justamente com a definição de θ_k, isto é, $\theta_k = \nabla\Phi(x_k)$ é exatamente o MD como apresentamos.

De uma maneira muito parecida com a feita no Capítulo 3, também podemos derivar mais rigorosamente o MD a partir da EDO usando expansão de Taylor, mas não faremos isso aqui.

5.5 Aplicação do MD em problemas de otimização no Simplexo

Provavelmente o caso mais notável de utilização do MD é o de otimização no simplexo, isto é, no conjunto

$$\Delta^{n-1} = \left\{ x \in \mathbb{R}^n \mid x_i \geq 0, \sum_{i=1}^{n} x_i = 1 \right\}.$$

Esse tipo de problema aparece em contextos que exigem otimização de portfólios ou alocação de recursos, como investimentos em fontes de energia elétrica ou seleção de culturas para plantio. Também aparece em *machine learning* ou problemas cujas variáveis são distribuições de probabilidade. O fato de as coordenadas dos elementos de Δ^{n-1} serem todas não negativas e somarem um justifica sua utilização nesses contextos.

Nesse caso, o MD é aplicado com mapeamento espelho e divergência de Bregman induzidos pela função $\Phi(x) = \sum_{i=1}^{n}(x_i \log(x_i) - x_i) = \sum_{i=1}^{n} x_i \log(x_i) - \sum_{i=1}^{n} x_i$, chamada de entropia negativa. Ela é o oposto (em sinal) da entropia mais um termo linear (que é igual a 1 para todos os pontos do simplexo); a entropia recebe este nome por atingir seu máximo no ponto $(1/n, 1/n, \ldots, 1/n)$ e ser minimizada no simplexo por pontos que acumulam toda a probabilidade numa só coordenada, como $(1, 0, \ldots, 0)$. É possível encontrar na literatura a entropia negativa definida por $\Phi(x) = \sum_{i=1}^{n}(x_i \log(x_i) - x_i)$. Aqui, a não adição do termo $-\sum_{i=1}^{n} x_i$ se traduziria na constante 1 sendo adicionada e subtraída em dois lugares diferentes, não alterando o resultado final, mas poluindo o desenvolvimento. Por conta disso, vamos usar a primeira versão.

Dada essa Φ, o mapeamento espelho e seu inverso são dados por

- $\nabla\Phi(x) = (\log(x_1), \log(x_2), \ldots, \log(x_n))$; e
- $(\nabla\Phi)^{-1}(\theta) = (\exp(\theta_1), \exp(\theta_2), \ldots, \exp(\theta_n))$.

Como o passo do gradiente é feito no espaço dual e o mapeamento, definido sobre os reais positivos, é sobrejetivo sobre o $\mathbb{R}^n$ (por ser coercivo na fronteira), o mapeamento inverso do ponto atualizado deve ser um ponto pertencente aos reais positivos. Isso garante que todos os pontos gerados nunca cairão fora de $\mathbb{R}^n_+$ e podemos pensar no mapeamento como uma espécie de barreira.

A divergência de Bregman é dada por

$$\begin{aligned}
D_\Phi(x, y) &= \sum_{i=1}^{n} x_i \log(x_i) - \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i \log(y_i) - \sum_{i=1}^{n} y_i - (\log(y_1), \log(y_2), \ldots, \log(y_n)))(x - y) \\
&= \sum_{i=1}^{n} x_i \log(x_i) - \sum_{i=1}^{n} y_i \log(y_i) - \sum_{i=1}^{n} x_i \log(y_i) + \sum_{i=1}^{n} y_i \log(y_i) - \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i \\
&= \sum_{i=1}^{n} (x_i \log(x_i) - x_i \log(y_i)) - \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \\
&= \sum_{i=1}^{n} x_i \log\left(\frac{x_i}{y_i}\right) - \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.
\end{aligned}$$

Essa divergência é também conhecida como divergência de Kullback-Leibler ([KL51]). Como consequência, a projeção de Bregman de um ponto $y \in \mathbb{R}^n_+$ sobre o simplexo Δ^{n-1} é dada por

$$\Pi^{\Phi}_{\Delta^{n-1}}(y) = \arg\min_{x \in \Delta^{n-1}} D_\Phi(x, y) = \arg\min_{x \in \Delta^{n-1}} \sum_{i=1}^{n} x_i \log\left(\frac{x_i}{y_i}\right) - \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.$$

Notemos que o último termo, $\sum_{i=1}^n y_i$, depende apenas de y e é constante em relação a x, a variável de minimização, podendo ser ignorado. Além disso, como $x \in \Delta^{n-1}$, temos $\sum_{i=1}^n x_i = 1$, o que implica que o segundo termo também é constante. Assim, o problema da projeção se reduz a

$$\Pi^{\Phi}_{\Delta^{n-1}}(y) = \arg\min_{x \in \Delta^{n-1}} \sum_{i=1}^n x_i \log\left(\frac{x_i}{y_i}\right).$$

Para resolvê-lo, utilizaremos as condições de otimalidade dadas pelo Teorema de Karush-Kuhn-Tucker ([Kar39], [KT51]) ou teorema KKT. Para isso, escrevemos o Lagrangiano associado à restrição $\sum_{i=1}^n x_i = 1$ como

$$\mathcal{L}(x, \lambda) = \sum_{i=1}^n x_i \log\left(\frac{x_i}{y_i}\right) + \lambda\left(\sum_{i=1}^n x_i - 1\right).$$

Pelas condições de KKT, no ponto minimizador x^*, devemos ter, para todo $i \in [n]$,

$$\frac{\partial}{\partial x_i}\mathcal{L}(x^*, \lambda) = \log\left(\frac{x_i^*}{y_i}\right) + 1 + \lambda = 0 \quad \Longrightarrow \quad \log\left(\frac{x_i^*}{y_i}\right) = -1 - \lambda \quad \Longrightarrow \quad x_i^* = y_i \exp(-1-\lambda).$$

Denotando $E = \exp(-1-\lambda)$, que é constante em relação a x_i^* e y_i para todos os i, obtemos $x_i = Ey_i$. Impondo a restrição $\sum_{i=1}^n x_i^* = 1$, temos

$$\sum_{i=1}^n x_i = E\sum_{i=1}^n y_i = 1 \quad \Longrightarrow \quad E = \frac{1}{\sum_{i=1}^n y_i}.$$

Portanto,

$$x_i = \frac{y_i}{\sum_{j=1}^n y_j},$$

isto é, a projeção de Bregman de y sobre Δ^{n-1} com geradora a entropia negativa é simplesmente a normalização de y, tornando sua soma igual a 1, o que acaba sendo a projeção mais natural que se poderia imaginar. Frisamos que essa é a normalização relativa à norma ℓ_1.

Com tudo isso, para cada $i \in [n]$, os pontos gerados pelo MD satisfazem

$$\begin{aligned}
\theta_{k+1_i} &= \log(x_{k_i}) - \eta\nabla f(x_k)_i \\
x'_{k+1_i} &= \exp(\log(x_{k_i}) - \eta\nabla f(x_k)_i) = (x_{k_i}\exp(-\eta\nabla f(x_k)_i) \\
(x_{k+1})_i &= \Pi^{\Phi}_{\Delta^{n-1}}(x'_{k+1_i}) = \frac{x'_{k_i}}{\sum_{i=1}^n x'_{k_i}} = \frac{x_{k_i}\exp(-\eta\nabla f(x_k)_i)}{\sum_{i=1}^n x_{k_i}\exp(-\eta\nabla f(x_k)_i)}.
\end{aligned}$$

Agora, em vez de significar uma direção explícita de atualização, o gradiente funciona como uma espécie de peso para ajuste das coordenadas. Esse ajuste é feito multiplicando cada coordenada por um número que determina um aumento ou diminuição. Neste instante, fica claro como métodos como o *multiplicative weights* são também casos particulares do MD.

Digamos que para um certo $i \in [n]$, tenhamos que $-\eta\nabla f(x)_i$ é negativo. Nesse caso, esperaríamos que após uma atualização, a i-ésima coordenada de x diminuísse. No GD isso acontece, mas possivelmente com a coordenada se tornando negativa. No MD, como $-\eta\nabla f(x)_i$ é negativo, $\exp(-\eta\nabla f(x)_i)$ é menor do que 1 e aí $x_{k_i}\exp(-\eta\nabla f(x_k)_i)$ é menor do que x_i, significando – como no GD – uma diminuição nessa coordenada, mas dessa vez sem que ela se torne negativa. Da maneira análoga, se $-\eta\nabla f(x)_i$ é positivo, esperamos um aumento na i-ésima coordenada de x. No MD, isso se traduz na multiplicação dessa coordenada por $\exp(-\eta\nabla f(x_k)_i)$, um número maior que 1. O crescimento/decrescimento de cada coordenada depende ainda dos crescimentos e decrescimentos das outras coordenadas: após a normalização, uma coordenada que diminui de tamanho pode acabar aumentando se todas as outras diminuíram mais.

Aqui o fato do mapeamento e seu conjugado garantirem que o resultado após uma atualização ainda ser positivo acabou se traduzindo no fato de $\exp(-\eta\nabla f(x)_i)$ ser sempre positivo. Além disso, a normalização altera apenas a magnitude dos vetores, preservando a direção e o sentido e, assim, também a positividade.

O fato de $\nabla\Phi$ ser coerciva na fronteira de $\mathbb{R}^n_+$ e a dinâmica da normalização fazem com que todos os pontos primais gerados pelo método estejam no **interior** do simplexo (isto se deve também ao fato do bordo

do simplexo estar contido no bordo de $\mathbb{R}^n_+$). Sendo assim, nas configurações deste problema, o MD é um método de pontos interiores. Se a solução do problema for na fronteira do simplexo, o MD não vai atingir exatamente ela, mas vai sempre se aproximar dela.

Por fim, mas nem um pouco menos importante, por conta dos passos do gradiente serem dados no espaço dual sem restrições, o parâmetro η pode ser consideravelmente grande (pelo menos teoricamente). No espaço primal isso se traduz com o vetor $\exp\{-\eta\nabla f(x)\}$ poder ser consideravelmente pequeno, já que depois de implementá-lo ainda fazemos uma normalização. Na prática, no entanto, existem limitações para os tamanhos suportados pelos computadores.

Capítulo 6

Accelerated Mirror Descent

Com a introdução do *Mirror Descent* como uma generalização do *Gradient Descent* no capítulo anterior e o conhecimento do método acelerado de Nesterov, podemos pensar na possibilidade de acelerar também o Mirror Descent, conseguindo assim uma generalização do Accelerated Gradient Descent. Na próxima seção discutimos duas possibilidades de métodos de Bregman acelerados: uma cuja aceleração acontece no espaço primal e outra cuja aceleração acontece no dual. Em seguida escolhemos uma delas e enunciamos dois teoremas sobre ela e sua EDO associada, que serão provados no próximo capítulo, 7.

6.1 Duas propostas de aceleração

Quando apresentamos a aceleração de Nesterov para o Gradient Descent, não havia dúvidas sobre o que deveria ser acelerado uma vez que havia apenas uma variável espacial, x_k. Com a introdução do Mirror Descent, cada variável primal x_k fica associada a uma variável dual θ_k. Ao buscar um meio de acelerar o MD como se faz com o GD no método de Nesterov, podemos tentar fazê-lo no espaço primal ou no espaço dual. Uma aceleração no espaço primal resultaria no método descrito por

$$\begin{cases} x_{k+1} &= (\nabla\Phi)^{-1}(\theta_{k+1}) \\ \theta_{k+1} &= \xi_k - \eta\nabla f(y_k) \\ \xi_k &= \nabla\Phi(y_k) \\ y_k &= x_k + \mu_k(x_k - x_{k-1}), \end{cases}$$

enquanto uma aceleração no espaço dual resultaria no método descrito por

$$\begin{cases} x_{k+1} &= (\nabla\Phi)^{-1}(\theta_{k+1}) \\ \theta_{k+1} &= \xi_k - \eta\nabla f(y_k) \\ y_k &= (\nabla\Phi)^{-1}(\xi_k) \\ \xi_k &= \theta_k + \mu_k(\theta_k - \theta_{k-1}), \end{cases}$$

onde $\mu_k = \frac{k-1}{k+2}$ nos dois casos. Mais uma vez, as equações das duas recorrências se resolvem na ordem contrária a qual elas estão apresentadas nesses sistemas, ou seja, começando pela última e terminando pela primeira; esta ordem foi escolhida de acordo com a "importância" de cada equação e tendo em vista que a variável de interesse é, em última análise, x_k. Chamaremos o primeiro esquema de *Primal Accelerated Mirror Descent* (P-AMD) e o segundo de *Dual Accelerated Mirror Descent* (D-AMD). As Figuras 6.1 e 6.2 ilustram os passos dados pelo P-AMD e D-AMD, respectivamente, em uma iteração.

Assim como acontece no AGD, o P-AMD incorpora um termo de momento relativo à variável x_k. No esquema D-AMD isso acontece no correspondente dual da variável x_k, θ_k. Quando $\Phi(x) = \frac{1}{2}\|x\|_2^2$ ambos os métodos correspondem ao AGD, como esperaríamos.

A ideia de passo de predição é preservada em ambos os esquemas, mas é um pouco mais natural no P-AMD, já que a variável de interesse — e, portanto, a que faria mais sentido prever — é x_k. No D-AMD,

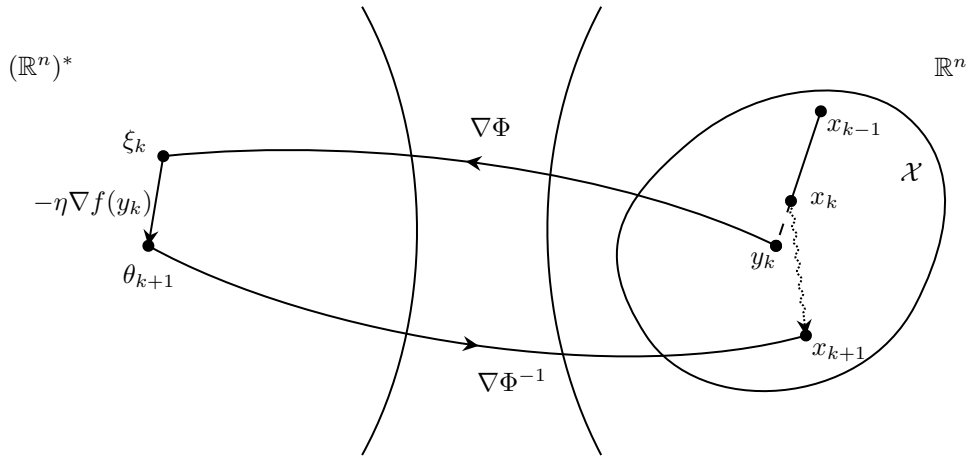

Figura 6.1: Uma iteração do método do P-AMD

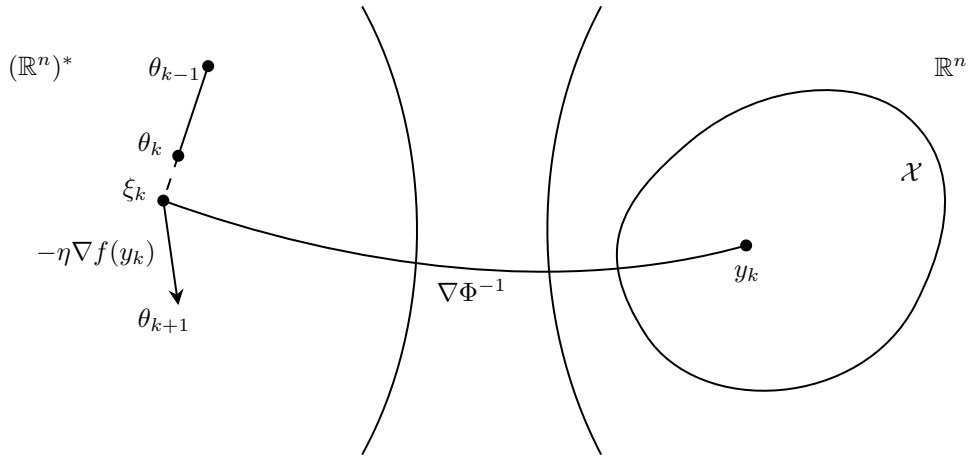

Figura 6.2: Uma iteração do método do D-AMD

de maneira um pouco menos natural mas ainda válida, o passo de predição se dá no espaço dual. Na verdade, todas as operações relevantes do D-AMD se dão no espaço dual: o passo de predição e o passo do gradiente. Se pensarmos de predição como um passo que incorpora um termo de inércia (que é a ideia original do método de Nesterov), sua função é utilizar informações (direção e magnitude) da última atualização do passo do gradiente para calcular o próximo, como um impulso. Dessa perspectiva, faz mais sentido que as duas operações se deem no mesmo espaço. A ideia de impulso se perde no P-AMD porque o mapeamento espelho distorce (pelo menos em algum nível) as informações da última atualização.

A função Φ é definida sobre um conjunto aberto $\mathcal{D}$ contendo o conjunto viável $\mathcal{X}$ para o problema de interesse. Como no P-AMD o passo de predição é dado no espaço primal, corremos o risco de sermos levados não só para fora de $\mathcal{X}$, mas para fora de $\mathcal{D}$, de forma que a projeção de Bregman não esteja definida para o ponto resultante. No caso do simplexo, por exemplo, como no GD, pode ser que o ponto resultante tenha coordenadas negativas. No caso do GD isso era resolvido pois a projeção usada era a euclidiana, definida para qualquer ponto no $\mathbb{R}^n$. Para o P-AMD esse problema é mais grave, pois exigiria uma outra projeção. No D-AMD, esse problema não existe: como os dois passos são dados no espaço dual e pela definição de mapeamento espelho, $\nabla\Phi(\mathcal{D}) = \mathbb{R}^{n*}$, o que faz com que a volta para $\mathcal{D}$ esteja garantida.

Ao tentar derivar informalmente, como feito na Seção 4.3, uma EDO para o P-AMD, não conseguimos seguir um raciocínio análogo ao de Su, Boyd e Candès [SBC16], porque o mapeamento impede que os objetos se comportem tão bem quanto no caso do AGD. Para obter a equação diferencial, partíamos da equação (4.2). No P-AMD, a equação equivalente seria

$$\theta_{k+1} = \nabla\Phi\left(x_k + \frac{k-1}{k+2}(x_k - x_{k-1})\right) - \eta\nabla f(y_k).$$

Nesse caso, o mapeamento impede que prossigamos de mesma forma.

Ao tentar derivar informalmente uma EDO para o D-AMD, no entanto, conseguimos aproveitar boa parte do raciocínio sem grandes problemas, como veremos na seção a seguir. Além disso, conseguimos provar rigorosamente, como feito em [SBC16, Apêndices A e B], que a EDO encontrada tem uma solução única e que o D-AMD converge para esta solução. Por essas razões, continuaremos com o D-AMD como a escolha mais natural de aceleração do MD e passaremos a chamá-lo simplesmente de AMD.

Analogamente aos outros métodos, o AMD também pode ser adaptado para resolver problemas com retrição usando projeções. Neste caso, assim como no caso do MD, a projeção usada é a de Bregman. A Figura 6.3 ilustra os passos dados pelo AMD no caso em que a projeção de Bregman é usada.

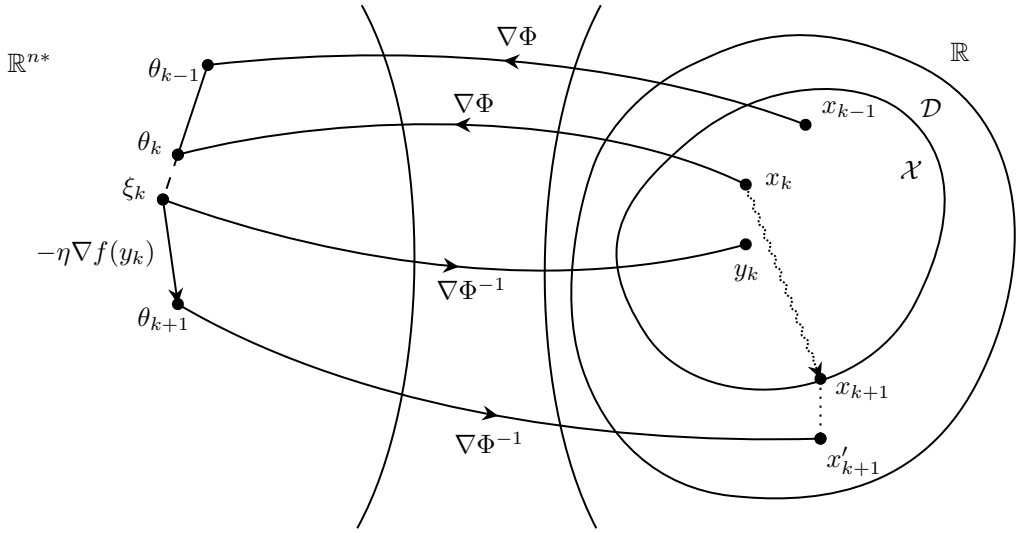

Figura 6.3: Uma iteração do método do AMD projetado

A Figura 6.4 mostra as trajetórias do MD e do D-AMD sobre as curvas de nível da função objetivo $f(x)$ no espaço primal, e da função objetivo $f((\nabla\Phi)^{-1}(\theta))$ no espaço dual, bem como o campo gradiente de Φ no espaço primal.

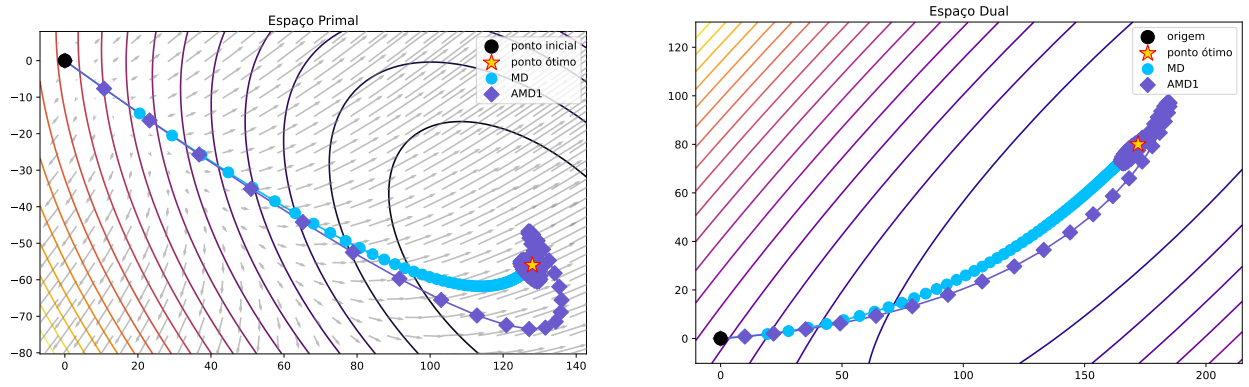

Figura 6.4: Trajetórias dos MD e AMD sobre curvas de nível da função objetivo.

Aqui vale notar que duas das equações que definem o AMD são equações de definição (a primeira e a terceira). Dessa forma, apesar de ser definido por quatro equações, a dinâmica do algoritmo é traduzida por duas delas. O mesmo acontece com o MD.

Observação. *Existe na literatura ainda um terceiro método que acelera o Mirror Descent. Ele pode ser derivado a partir de uma adaptação (troca de distância euclidiana por divergência de Bregman) em uma função de Lyapunov ([KBB15]) ou em um lagrangiano cuja integral é minimizada numa perspectiva variacional ([WWJ16].*

6.2 Derivação informal da EDO

Uma derivação muito análoga à feita na Seção 4.3 pode ser realizada para se obter uma EDO associada ao AMD. Aqui, a fim de evitar muitas repetições, destacamos apenas as principais mudanças decorrentes da incorporação do mapeamento espelho.

Para começar, anteriormente supusemos a existência de uma função $X : \mathbb{R}_+ \to \mathbb{R}^n$ de classe C^2, com $\frac{\mathrm{d}}{\mathrm{d}t}X(0) = \vec{0}$, tal que os pontos iterados da forma $x_k = X(k\Delta t) = X(t)$ satisfazem o esquema de Nesterov com tamanho de passo $\eta = (\Delta t)^2$. A hipótese de que $X \in C^2$ era necessária para que pudéssemos aplicar a expansão de Taylor de segunda ordem. Aqui, devemos supor que os pontos iterados satisfazem o AMD, naturalmente, e que $\Theta(x) = \nabla\Phi(x)$ também é de classe C^2 para podermos repetir o mesmo procedimento. Com isso, obtemos

$$\Theta(X(t+\Delta t)) = \Theta(X(t)) + \frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t)) \cdot \Delta t + \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Theta(X(t)) \cdot (\Delta t)^2 + o((\Delta t)^3).$$

Após algum desenvolvimento, e usando as equações oriundas da definição do AMD (mais precisamente as equações duais, cujas variáveis principais pertencem ao espaço dual), temos

$$(k+2)\frac{\Theta(X(t+\Delta t)) - \Theta(X(t))}{\Delta t} = (k-1)\frac{\Theta(X(t)) - \Theta(X(t-\Delta t))}{\Delta t} - (k+2)\Delta t\nabla f(y_k).$$

Como anteriormente substituímos (e justificamos a substituição) $\Delta t\nabla f(y_k)$ por $\Delta t\nabla f(X(t)) + o(\Delta t)$, gostaríamos de repetir esse passo aqui. Entretanto, essa passagem é um pouco mais delicada. Vamos primeiro provar que $\xi_k = \nabla\Phi(y_k) = \nabla\Phi(X(t)) + D$, onde $D = O(\Delta t)$.

Como $\Theta(x) = \nabla\Phi(x)$ é diferenciável, e $X(t)$ também, então $\Theta(X(t))$ é diferenciável. Aplicando a expansão de Taylor em $\Theta(X(t-\Delta t))$ em torno de t, temos

$$\Theta(X(t-\Delta t)) = \Theta(X(t)) - \frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t))(\Delta t) + o(\Delta t)$$

$$\implies \quad \Theta(X(t)) - \Theta(X(t-\Delta t)) = \frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t)()\Delta t) + o(\Delta t).$$

Dessa forma, obtemos

$$\xi_k = \Theta(X(t)) + \mu_k(\Theta(X(t)) - \Theta(X(t-\Delta t))) = \Theta(X(t)) + \mu_k\left(\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t))\cdot\Delta t + o(\Delta t)\right).$$

Como $\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t))$ é independente de Δt, e $\mu_k \in [0,1]$, temos que o termo $\mu_k\left(\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t))\cdot\Delta t + o(\Delta t)\right) = O(\Delta t)$. Chamando esse termo de D, temos a aproximação desejada.

Usando expansão de Taylor novamente, agora em $\nabla f((\nabla\Phi)^{-1}(\Theta(X(t)) + D))$ e em relação à variável $\Theta(X(t))$, e lembrando que $(\nabla\Phi)^{-1}\Theta(X(t)) = (\nabla\Phi)^{-1}\nabla\Phi X(t) = X(t)$, temos

$$\begin{aligned}
\nabla f(y_k) = \nabla f((\nabla\Phi)^{-1}(\xi_k)) &= \nabla f\left((\nabla\Phi)^{-1}\left(\Theta(X(t)) + D\right)\right) \\
&= \nabla f\left((\nabla\Phi)^{-1}\left(\Theta(X(t))\right)\right) + \frac{\mathrm{d}}{\mathrm{d}t}\nabla f\left((\nabla\Phi)^{-1}\left(\Theta(X(t))\right)\right)D + o(D) \\
&= \nabla f(X(t))) + \frac{\mathrm{d}}{\mathrm{d}t}\nabla f(X(t))D + o(D) \\
&= \nabla f(X(t))) + \nabla^2 f(X(t))\frac{\mathrm{d}}{\mathrm{d}t}X(t)D + o(D).
\end{aligned}$$

Como $\frac{\mathrm{d}}{\mathrm{d}t}X(t)$ é constante em relação a Δt, $D = O(\Delta t)$, e $\nabla^2 f$ é limitada (pois ∇f é L-Lipschitz), concluímos que

$$\nabla f(y_k) = \nabla f(X(t)) + O(\Delta t) \quad \Longrightarrow \quad \Delta t\nabla f(y_k) = \Delta t\nabla f(X(t)) + o(\Delta t).$$

Substituindo na equação principal, obtemos

$$\begin{aligned}
&(k+2)\left(\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t)) + \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Theta(X(t))\cdot\Delta t + o((\Delta t)^2)\right) \\
&\qquad = (k-1)\left(\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t)) - \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Theta(X(t))\cdot\Delta t + o((\Delta t)^2)\right) - (k+2)\left(\Delta t\nabla f(X(t)) + o(\Delta t)\right),
\end{aligned}$$

que, após mais algum desenvolvimento, nos leva a

$$3\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t)) + t\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Theta(X(t)) + t\nabla f(X(t)) = \Delta t\left(\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Theta(X(t)) + 2\nabla f(X(t)) + o((\Delta t)^2) + o(1)\right).$$

Tomando o limite de Δt tendendo a 0 em ambos os lados da equação, o lado esquerdo se mantém e o direito fica igual a 0. Dividindo tudo por t, chegamos enfim a

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Theta(X(t)) + \frac{3}{t}\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t)) + \nabla f(X(t)) = 0. \tag{6.1}$$

A Figura 5.6 mostra as trajetórias de $X(t)$ no espaço primal e de $\Theta(X(t))$ no espaço dual.

Como foi antecipado no Capítulo 4, existe no Seção B.3 uma derivação do AMD a partir de uma discretização da EDO 6.1, o que une, pelo menos conceitualmente, os dois objetos. No Capítulo 7, vamos provar, entre outras coisas e da mesma maneira como foi para o AGD, que a "distância" da trajetória do AMD para a solução da EDO tende a zero conforme refinamos o método, o que fecha a união dos objetos de maneira definitiva.

6.3 Aplicação do AMD em problemas de otimização no Simplexo

Vejamos agora como se traduz o AMD no caso mais notável de utilização do MD: o de otimização sobre o simplexo com mapeamento espelho induzido pela entropia negativa. Lembremo-nos que nesse caso, temos

$$(\nabla\Phi(x))_i = \log(x_i) \quad \text{e} \quad ((\nabla\Phi)^{-1}(\theta))_i = \exp(\theta_i).$$

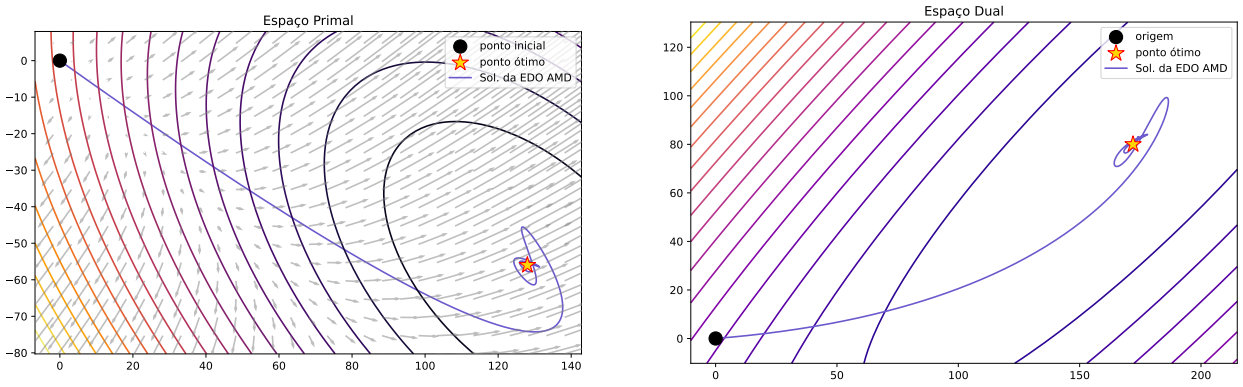

Figura 6.5: Trajetória do AMD e solução da EDO associada sobre curvas de nível da função objetivo.

Começando pela última equação que define o AMD, $\xi_k = \theta_k + \mu_k(\theta_k - \theta_{k-1})$, temos

$$\begin{aligned}(\xi_k)_i &= (\nabla\Phi(x_k))_i + \mu_k((\nabla\Phi(x_k))_i - (\nabla\Phi(x_{k-1}))_i) \\ &= \log((x_k)_i) + \mu_k(\log((x_k)_i) - \log((x_{k-1})_i)) \\ &= \log((x_k)_i) + \mu_k(\log((x_k)_i) - \log((x_{k-1})_i)) \\ &= \log\left((x_k)_i \frac{(x_k)_i^{\mu_k}}{(x_{k-1})_i^{\mu_k}}\right) = \log\left((x_k)_i \left(\frac{(x_k)_i}{(x_{k-1})_i}\right)^{\mu_k}\right)\end{aligned}$$

Usando agora a penúltima equação, $y_k = (\nabla\Phi)^{-1}(\xi_k)$, obtemos

$$(y_k)_i = ((\nabla\Phi)^{-1}(\xi_k))_i = \exp\left(\log\left((x_k)_i \left(\frac{(x_k)_i}{(x_{k-1})_i}\right)^{\mu_k}\right)\right) = (x_k)_i \left(\frac{(x_k)_i}{(x_{k-1})_i}\right)^{\mu_k}.$$

Notemos como o passo de predição dado no espaço dual se traduz como um passo de predição multiplicativo no espaço primal. Se $(x_k)_i \geq (x_{k-1})_i$, por exemplo, $\frac{(x_k)_i}{(x_{k-1})_i} \geq 1$ e aí $\left(\frac{(x_k)_i}{(x_{k-1})_i}\right)^{\mu_k} \geq 1$, fazendo com que $(y_k)_i$ seja maior do que $(x_k)_i$, o que faz sentido já que a coordenada vinha crescendo. De maneira análoga, se $(x_k)_i \leq (x_{k-1})_i$, então $(y_k)_i \leq (x_k)_i$, o que também faz sentido já que $(x_k)_i$ vinha diminuindo. Isso tudo é ainda controlado por μ_k, que começa sendo igual a 0, o que anula o efeito do passo de predição, e cresce sendo sempre limitado por 1, de maneira que o efeito multiplicativo cresça também.

Aqui, se normalizarmos y_k, o efeito de crescimento e decrescimento de cada coordenada fica sujeito aos efeitos de crescimento e decrescimento das outras coordenadas, como no MD.

Da segunda equação, $\theta_{k+1} = \xi_k - \eta\nabla f(y_k)$, temos

$$(\theta_{k+1})_i = \log\left((x_k)_i \left(\frac{(x_k)_i}{(x_{k-1})_i}\right)^{\mu_k}\right) - \eta(\nabla f(y_k))_i.$$

Da primeira equação, $x_{k+1} = (\nabla\Phi)^{-1}(\theta_{k+1})$, temos

$$\begin{aligned}(x'_{k+1})_i &= ((\nabla\Phi)^{-1}(\theta_{k+1}))_i \\ &= \exp\left(\log\left((x_k)_i \left(\frac{(x_k)_i}{(x_{k-1})_i}\right)^{\mu_k}\right) - \eta(\nabla f(y_k))_i\right) \\ &= (x_k)_i \left(\frac{(x_k)_i}{(x_{k-1})_i}\right)^{\mu_k} \exp\left(-\eta(\nabla f(y_k))_i\right).\end{aligned}$$

Finalmente, fazendo a projeção de x'_{k+1} em Δ^{n-1}, a fórmula fechada para pontos iterados pelo AMD é

$$(x_{k+1})_i = \frac{(x_k)_i \left(\frac{(x_k)_i}{(x_{k-1})_i}\right)^{\mu_k} \exp\left(-\eta(\nabla f(y_k))_i\right)}{\sum_{i=1}^n (x_k)_i \left(\frac{(x_k)_i}{(x_{k-1})_i}\right)^{\mu_k} \exp\left(-\eta(\nabla f(y_k))_i\right)}.$$

Nessas configurações, a aceleração do AMD se dá no espaço primal de maneira multiplicativa, assim como o passo do gradiente, o que já ocorria para o MD.

Capítulo 7

Uma EDO para o AMD

Neste capítulo, apresentamos e provamos teoremas que garantem 1) a existência e unicidade da solução de (6.1); e 2) a convergência do AMD para esta solução.

7.1 Existência e Unicidade

Nesta seção, provamos a existência e unicidade da solução da EDO 6.1, resultado enunciado no próximo teorema.

A existência e a unicidade da solução da EDO 6.1 estão enunciadas no teorema a seguir. Sua demonstração se dará ao longo desta seção, dividida em duas subseções (uma para a existência e outra para a unicidade), seguindo uma série de resultados auxiliares.

Teorema 7.1 (Existência e Unicidade)**.** *Sejam $f \in \mathcal{F}_L$ uma função diferenciável com gradiente ∇f L-Lipschitz, Φ uma função ν-fortemente convexa e diferenciável e $x_0 \in \mathbb{R}^n$ um ponto qualquer. Então a EDO (6.1) com condições iniciais $X(0) = x_0$ e $\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(0)) = \vec{0}$ tem uma única solução $X \in C^2((0,\infty);\mathbb{R}^n)) \cap C^1([0,\infty);\mathbb{R}^n))$.*

Demonstração. A prova se dá juntando os lemas 7.11 e 7.15, provados aos finais das próximas duas subseções, respectivamente. ■

7.1.1 Existência

A EDO 6.1 pode ser reescrita como

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\Theta(X(t))\\ Z(t)\end{pmatrix} = \begin{pmatrix}Z(t)\\ -\frac{3}{t}Z(t) - \nabla f(X(t))\end{pmatrix} = F((\Theta(X(t)), Z(t)), t),$$

com condições iniciais $\Theta(X(t_0)) = \nabla\Phi(x_0) = \theta_0$ e $\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t_0)) = \vec{0}$.

Por conta do termo $\frac{3}{t}$, a função F tem uma singularidade em $t = 0$ e não é Lipschitz em relação à variável $(\Theta(X(t)), Z(t))$, o que impede a aplicação direta do teorema de Picard-Lindelöf à EDO, que garantiria a existência e unicidade de sua solução. Para contornar isso vamos primeiro criar uma sequência de EDOs suavizadas (de forma que seja possível aplicar o teorema de Picard-Lindelöf a cada uma delas) que, no limite, se aproximam da EDO de interesse e depois tomar as soluções dessas EDOs e conseguir uma sequência de soluções cujo limite (de uma certa subsequência) deve satisfazer a EDO original. Para a segunda parte, vamos usar o teorema de Ascoli-Arzelà.

Comecemos definindo as EDOs suavizadas.

Definição. *A aproximação contínua da EDO 6.1 com parâmetro $\delta > 0$ é definida como*

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Theta(X(t)) + \frac{3}{\max(\delta, t)}\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t)) + \nabla f(X(t)) = 0 \tag{7.1}$$

com condições iniciais $X(0) = x_0 \Longrightarrow \Theta(X(0)) = \theta_0$ e $\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(0)) = 0$.

Aqui o termo singular em $t = 0$ foi substituído pelo termo $\frac{3}{\max(\delta,t)}$, de forma que ele não é mais singular para $\delta > 0$ fixo. Para cada $\delta > 0$, a EDO suavizada pode ser escrita como

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\Theta(X(t))\\ Z(t)\end{pmatrix} = \begin{pmatrix}Z(t)\\ -\frac{3}{\max(\delta,t)}Z(t) - \nabla f(X(t))\end{pmatrix} = F_\delta((\Theta(X(t)), Z(t)), t), \tag{7.2}$$

com condições iniciais $\Theta(X(t_0)) = \nabla\Phi(x_0) = \theta_0$ e $\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t_0)) = \vec{0}$.

A função $F_\delta((\theta, z), t) = \left(z, -\frac{3}{\max(\delta,t)}z - \nabla f((\nabla\Phi)^{-1}(\theta))\right)$ depende de t apenas através do termo $\frac{3}{\max(\delta,t)}$, que é contínuo em t, de forma que F_δ seja também contínua em t. Além disso,

$$\begin{aligned}
&\|F_\delta((\theta_1, z_1), t) - F_\delta((\theta_2, z_2), t)\| \\
&\qquad = \left\|\left(z_1 - z_2, -\frac{3}{\max(\delta,t)}(z_1 - z_2) - \left(\nabla f((\nabla\Phi)^{-1}(\theta_1)) - \nabla f((\nabla\Phi)^{-1}(\theta_2))\right)\right)\right\| \\
&\qquad\qquad \leq \|z_1 - z_2\| + \underbrace{\left\|-\frac{3}{\max(\delta,t)}(z_1 - z_2) - \left(\nabla f((\nabla\Phi)^{-1}(\theta_1)) - \nabla f((\nabla\Phi)^{-1}(\theta_2))\right)\right\|}_{\mathrm{N}},
\end{aligned}$$

onde, usando o fato de que $\frac{3}{\max(\delta,t)} \leq \frac{3}{\delta}$, a L-Lipschitz continuidade de ∇f em relação a $x = (\nabla\Phi)^{-1}(\theta)$ e a $\frac{1}{\nu}$-Lipschitz continuidade de $(\nabla\Phi)^{-1}(\theta)$ em relação a θ (ver corolário A.3 no Apêndice A), o termo N satisfaz

$$\begin{aligned}
N &\leq \frac{3}{\max(\delta,t)}\|z_1 - z_2\| + \left\|\left(\nabla f((\nabla\Phi)^{-1}(\theta_1)) - \nabla f((\nabla\Phi)^{-1}(\theta_2))\right)\right\|_* \\
&\leq \frac{3}{\delta}\|z_1 - z_2\| + L\left\|\left((\nabla\Phi)^{-1}(\theta_1) - (\nabla\Phi)^{-1}(\theta_2)\right)\right\| \\
&\leq \frac{3}{\delta}\|z_1 - z_2\| + \frac{L}{\nu}\|\theta_1 - \theta_2\|.
\end{aligned}$$

Seja $G = \max\left(1 + \frac{3}{\delta}, \frac{L}{\nu}\right)$. Temos então que

$$\begin{aligned}
\|F_\delta((\theta_1, z_1), t) - F_\delta((\theta_2, z_2), t)\| &\leq \|z_1 - z_2\| + \frac{3}{\delta}\|z_1 - z_2\| + \frac{L}{\nu}\|\theta_1 - \theta_2\| \\
&= \left(1 + \frac{3}{\delta}\right)\|z_1 - z_2\| + \frac{L}{\nu}\|\theta_1 - \theta_2\| \\
&\leq G\|z_1 - z_2\| + G\|\theta_1 - \theta_2\| \leq gG\|(z_1, \theta_1) - (z_2, \theta_2)\|,
\end{aligned}$$

onde g é a constante tal que $\|u\| + \|v\| \leq g\|(u, v)\|$ de acordo com a norma usada. Desta maneira, para cada $\delta > 0$, a função F_δ é Lipschitz em (θ, z).

Como F_δ é contínua na variável temporal e Lipschitz na variável espacial, podemos usar o teorema de Picard-Lindelöf e obter para cada $\delta > 0$ uma única solução X_δ para a EDO (7.1).

Gostaríamos de construir agora uma sequência de soluções com δ tendendo a zero e usar Ascoli-Arzelà para extrair a convergência de uma subsequência dela. O teorema, no entanto, exige que provemos duas propriedades sobre a sequência: que ela é equicontínua e que ela é uniformemente limitada. Essas condições estão definidas mais adiante. Para mostrar isso, consideremos o objeto auxiliar a seguir.

Definição. *Para $\delta > 0$ qualquer, definimos*

$$\alpha_\delta(t) = \sup_{u\in(0,t]} \frac{\left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(u))\right\|}{u}.$$

O termo mede o quão rapidamente a derivada $\frac{d}{dt}\Theta(X_\delta(t))$ cresce em relação ao tempo. Se $\alpha_\delta(t)$ for limitado, isso significa que o termo $-\frac{3}{t}\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t))$ permanece controlado mesmo quando $t \to 0$. Isso será fundamental para garantir que a família a ser construída seja tanto equicontínua quanto uniformemente limitada.

Os próximos fatos e lemas se dedicam a provar que $\alpha_\delta(t)$ é limitado e suas demonstrações são principalmente algébricas. Os dois primeiros fatos são bastante básicos e imediatos.

Fato 7.2. *Para $\delta > 0$ e $t > 0$ quaisquer, temos*

$$\forall u \in (0, t], \alpha_\delta(u) \leq \alpha_\delta(t).$$

Demonstração. Basta notar que os dois objetos são supremos de uma mesma expressão, apenas tomados em intervalos diferentes, sendo que um, $(0, t]$, contém o outro, $(0, u]$. ■

Fato 7.3. *Para $\delta > 0$ e $t > 0$ quaisquer, temos*

$$\forall u \in (0, t], \alpha_\delta(t)u \geq \left\| \frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(u)) \right\|.$$

Demonstração. Sejam $\delta > 0$ e $t > 0$ quaisquer. Da definição de α_δ temos que para todo $u \in (0, t]$.

$$\alpha_\delta(t) \geq \frac{\left\| \frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(u)) \right\|}{u} \quad \Longrightarrow \quad \alpha_\delta(t)u \geq \left\| \frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(u)) \right\|.$$ ■

O lema a seguir e seu corolário estabelecem a desigualdade $\|f(X_\delta(t))\| \leq \|\nabla f(x_0)\| + \frac{L}{2\nu}\alpha_\delta(t)t^2$. Vamos usá-la para chegar à desigualdade $\alpha_\delta(\delta) \leq \|\nabla f(x_0)\|_* + \frac{L}{2\nu}\alpha_\delta(\delta)\delta^2$, que nos levará a um limite superior para $\alpha_\delta(\delta)$. Esse limite, por sua vez, nos permitirá escrever um limite superior para $\alpha_\delta(t)$.

Lema 7.4. *Para $\delta > 0$ e $t > 0$ quaisquer, temos*

$$\|\nabla f(X_\delta(t)) - \nabla f(x_0)\|_* \leq \frac{L}{2\nu}\alpha_\delta(t)t^2.$$

Demonstração. Sejam $\delta > 0$ e $t > 0$ quaisquer. Como ∇f é L-Lipschitz temos

$$\|\nabla f(X_\delta(t) - \nabla f(x_0)\|_* \leq L \|X_\delta(t) - x_0\|.$$

Pelo fato A.2, encontrado no Apêndice A, obtemos

$$\|\nabla f(X_\delta(t) - \nabla f(x_0)\|_* \leq \frac{L}{\nu} \|\Theta(X(t)) - \Theta(X(0))\|. \tag{7.3}$$

Pelo Teorema Fundamental do Cálculo, temos ainda que

$$\int_0^t \frac{\mathrm{d}}{\mathrm{d}u}\Theta(X_\delta(u))\mathrm{d}u = \Theta(X_\delta(t)) - \Theta(X_\delta(0)) = \Theta(X_\delta(t)) - \theta_0.$$

Aplicando agora a desigualdade triangular para integrais e o fato 7.3, obtemos

$$\|\Theta(X_\delta(t)) - \theta_0\| \leq \int_0^t \left\| \frac{\mathrm{d}}{\mathrm{d}u}\Theta(X(u)) \right\| \mathrm{d}u \leq \int_0^t \alpha_\delta(t)u\mathrm{d}u = \alpha_\delta(t)\int_0^t u\mathrm{d}u = \alpha_\delta(t)\frac{t^2}{2}.$$

E finalmente, multiplicando ambos os lados por $\frac{L}{\nu}$ e usando a desigualdade (7.3), temos

$$\|\nabla f(X_\delta(t)) - \nabla f(x_0)\|_* \leq \frac{L}{2\nu}\alpha_\delta(t)t^2.$$ ■

Corolário 7.5. *Para $t > 0$ qualquer, temos $\|\nabla f(X_\delta(t))\|_* \leq \|\nabla f(x_0)\|_* + \frac{L}{2\nu}\alpha_\delta(t)t^2$*

Demonstração. Basta observar que para todo $t > 0$, $\|\nabla f(X_\delta(t))\|_* - \|\nabla f(x_0)\|_* \leq \|\nabla f(X_\delta(t)) - \nabla f(x_0)\|_*$ e aplicar o lema 7.4. ■

Os próximos dois lemas se dedicam a provar limites superiores para $\alpha_\delta(\delta)$ e $\alpha_\delta(t)$. O primeiro lema será usado na demonstração do segundo e o segundo será usado para limitar $\alpha_\delta\left(\sqrt{6\nu/L}\right)$ para $\delta < \sqrt{6\nu/L}$.

Lema 7.6. *Para $0 < \delta < \sqrt{6\nu/L}$ qualquer, temos*

$$\alpha_\delta(\delta) \leq \frac{\|\nabla f(x_0)\|_*}{1 - L\delta^2/6\nu}.$$

Demonstração. Sejam $0 < t \leq \delta < \sqrt{6\nu/L}$ quaisquer. Como X_δ é solução de 7.1 e $\max(\delta, t) = \delta$, vale que para $u \in [0, t]$,

$$\frac{\mathrm{d}^2}{\mathrm{d}u^2}\Theta(X_\delta(u)) + \frac{3}{\delta}\frac{\mathrm{d}}{\mathrm{d}u}\Theta(X_\delta(u)) + \nabla f(X_\delta(u)) = 0$$

ou

$$\frac{\mathrm{d}^2}{\mathrm{d}u^2}\Theta(X_\delta(u)) + \frac{3}{\delta}\frac{\mathrm{d}}{\mathrm{d}u}\Theta(X_\delta(u)) = -\nabla f(X_\delta(u)).$$

Multiplicando os dois lados por $e^{3u/\delta}$, ficamos com

$$e^{3u/\delta}\frac{\mathrm{d}^2}{\mathrm{d}u^2}\Theta(X_\delta(u)) + \frac{3}{\delta}e^{3u/\delta}\frac{\mathrm{d}}{\mathrm{d}u}\Theta(X_\delta(u)) = -e^{3u/\delta}\nabla f(X_\delta(u)).$$

Agora notemos que, pela regra da cadeia, nós temos

$$\frac{\mathrm{d}}{\mathrm{d}u}\left(e^{3u/\delta}\frac{\mathrm{d}}{\mathrm{d}u}\Theta(X_\delta(u))\right) = e^{3u/\delta}\frac{\mathrm{d}^2}{\mathrm{d}u^2}\Theta(X_\delta(u)) + \frac{3}{\delta}e^{3u/\delta}\frac{\mathrm{d}}{\mathrm{d}u}\Theta(X_\delta(u)) = -e^{3u/\delta}\nabla f(X_\delta(u)).$$

Integrando os dois lados de 0 a t, ficamos com

$$\begin{aligned}
-\int_0^t e^{3u/\delta}\nabla f(X_\delta(u))\mathrm{d}u &= \int_0^t \frac{\mathrm{d}}{\mathrm{d}u}\left(e^{3u/\delta}\frac{\mathrm{d}}{\mathrm{d}u}\Theta(X_\delta(u))\right)\mathrm{d}u \\
&= e^{3t/\delta}\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t)) - e^0\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(0))) = e^{3t/\delta}\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t)).
\end{aligned}$$

Assim,

$$\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t)) = -e^{-3t/\delta}\int_0^t e^{3u/\delta}\nabla f(X_\delta(u))\mathrm{d}u\ .$$

Tomando o módulo dos dois lados e usando a desigualdade triangular para integrais, obtemos

$$\begin{aligned}
\left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t))\right\| &= \left\|-e^{-3t/\delta}\int_0^t e^{3u/\delta}\nabla f(X_\delta(u))\,\mathrm{d}u\right\| \\
&\leq e^{-3t/\delta}\int_0^t e^{3u/\delta}\left\|\nabla f(X_\delta(u))\right\|_*\mathrm{d}u \\
&\leq e^{-3t/\delta}\int_0^t e^{3t/\delta}\left\|\nabla f(X_\delta(u))\right\|_*\mathrm{d}u = \int_0^t \left\|\nabla f(X_\delta(u))\right\|_*\mathrm{d}u.
\end{aligned}$$

Agora, pelo corolário 7.5, temos que

$$\begin{aligned}
\left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t))\right\| \leq \int_0^t \left\|\nabla f(X_\delta(u))\right\|_*\mathrm{d}u &\leq \int_0^t \left(\left\|\nabla f(x_0)\right\|_* + \frac{L}{2\nu}\alpha_\delta(u)u^2\right)\mathrm{d}u \\
&= \int_0^t \left\|\nabla f(x_0)\right\|_*\mathrm{d}u + \int_0^t \frac{L}{2\nu}\alpha_\delta(u)u^2\mathrm{d}u \\
&= \left\|\nabla f(x_0)\right\|_* t + \int_0^t \frac{L}{2\nu}\alpha_\delta(u)u^2\mathrm{d}u\ .
\end{aligned}$$

Onde, pelo fato 7.2, essa última integral é tal que

$$\int_0^t \frac{L}{2\nu}\alpha_\delta(u)u^2\mathrm{d}u \leq \int_0^t \frac{L}{2\nu}\alpha_\delta(t)u^2\mathrm{d}u = \frac{L}{2\nu}\alpha_\delta(t)\int_0^t u^2\mathrm{d}u = \frac{L}{2\nu}\alpha_\delta(t)\frac{t^3}{3} = \frac{L}{6\nu}\alpha_\delta(t)t^3.$$

Desta maneira, temos

$$\left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t))\right\| \leq \left\|\nabla f(x_0)\right\|_* t + \frac{L}{6\nu}\alpha_\delta(t)t^3.$$

Dividindo os dois lados por t e depois tomando os supremos em $t \in (0, \delta]$, obtemos

$$\begin{aligned}\alpha_\delta(\delta) = \sup_{t\in(0,\delta]} \frac{\left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t))\right\|}{t} &\leq \sup_{t\in(0,\delta]} \left\{\|\nabla f(x_0)\|_* + \frac{L}{6\nu}\alpha_\delta(t)t^2\right\} \\ &= \|\nabla f(x_0)\|_* + \frac{L}{6\nu}\alpha_\delta(\delta)\delta^2.\end{aligned}$$

Como $\sqrt{6\nu/L} > \delta \implies 1 > L\delta^2/6\nu \implies 1 - L\delta^2/6\nu \geq 0$, temos enfim que

$$\alpha_\delta(\delta)(1 - L\delta^2/6\nu) \leq \|\nabla f(x_0)\|_* \quad \iff \quad \alpha_\delta(\delta) \leq \frac{\|\nabla f(x_0)\|_*}{1 - L\delta^2/6\nu}. \qquad \blacksquare$$

Lema 7.7. *Para $\delta < \sqrt{6\nu/L}$ e $\delta \leq t < \sqrt{12\nu/L}$ quaisquer, temos*

$$\alpha_\delta(t) \leq \frac{(5 - L\delta^2/6\nu)\,\|\nabla f(x_0)\|_*}{4(1 - L\delta^2/6\nu)(1 - Lt^2/12\nu)}.$$

Demonstração. A primeira parte desta demonstração se dá de forma muito parecida com a demonstração do lema anterior.

Sejam $0 < \delta < \sqrt{6\nu/L} \leq t < \sqrt{12\nu/L}$ quaisquer. Como X_δ é solução de (7.1) e $\max(\delta, u) = u$ para $u \in [\delta, t]$, podemos escrever

$$\frac{\mathrm{d}^2}{\mathrm{d}u^2}\Theta(X_\delta(u)) + \frac{3}{u}\frac{\mathrm{d}}{\mathrm{d}u}\Theta(X_\delta(u)) = -\nabla f(X_\delta(u)).$$

Multiplicando os dois lados por u^3, ficamos com

$$u^3\frac{\mathrm{d}^2}{\mathrm{d}u^2}\Theta(X_\delta(u)) + 3u^2\frac{\mathrm{d}}{\mathrm{d}u}\Theta(X_\delta(u)) = -u^3\nabla f(X_\delta(u)).$$

Agora a regra da cadeia nos dá

$$\frac{\mathrm{d}}{\mathrm{d}u}\left(u^3\frac{\mathrm{d}}{\mathrm{d}u}\Theta(X_\delta(u))\right) = 3u^2\frac{\mathrm{d}}{\mathrm{d}u}\Theta(X_\delta(u)) + u^3\frac{\mathrm{d}^2}{\mathrm{d}u^2}\Theta(X_\delta(u)) = -u^3\nabla f(X_\delta(u)). \tag{7.4}$$

Integrando de δ a t, ficamos com

$$\begin{aligned}-\int_\delta^t u^3\nabla f(X_\delta(u))\mathrm{d}u &= \int_\delta^t \frac{\mathrm{d}}{\mathrm{d}u}\left(u^3\frac{\mathrm{d}}{\mathrm{d}u}\Theta(X_\delta(u))\right)\mathrm{d}u \\ &= t^3\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t)) - \delta^3\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(\delta)),\end{aligned}$$

o que nos leva a

$$\begin{aligned}t^3\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t)) - \delta^3\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(\delta)) &= -\int_\delta^t u^3\nabla f(X_\delta(u))\mathrm{d}u \\ &= -\int_\delta^t u^3(\nabla f(X_\delta(u)) + \nabla f(x_0) - \nabla f(x_0))\mathrm{d}u \\ &= -\int_\delta^t u^3\nabla f(x_0)\mathrm{d}u - \int_\delta^t u^3(\nabla f(X_\delta(u)) - \nabla f(x_0))\mathrm{d}u\end{aligned}$$

ou, escrevendo de outra maneira, a

$$t^3\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t)) = -\int_\delta^t u^3\nabla f(x_0)\mathrm{d}u - \int_\delta^t u^3(\nabla f(X_\delta(u)) - \nabla f(x_0))\mathrm{d}u + \delta^3\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(\delta)).$$

Agora, tomando o módulo dos dois lados e usando a desigualdade triangular, obtemos

$$\begin{aligned}
t^3\left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t))\right\| &= \left\|t^3\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t))\right\| \\
&= \left\|-\int_\delta^t u^3\nabla f(x_0)\mathrm{d}u - \int_\delta^t u^3(\nabla f(X_\delta(u)) - \nabla f(x_0))\mathrm{d}u + \delta^3\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(\delta))\right\| \\
&\leq \left\|\int_\delta^t u^3\nabla f(x_0)\mathrm{d}u\right\| + \left\|\int_\delta^t u^3(\nabla f(X_\delta(u)) + \nabla f(x_0))\mathrm{d}u\right\| + \left\|\delta^3\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(\delta))\right\| \\
&\leq \underbrace{\int_\delta^t u^3\left\|\nabla f(x_0)\right\|_*\mathrm{d}u}_{I_1} + \underbrace{\int_\delta^t u^3\left\|\nabla f(X_\delta(u)) + \nabla f(x_0)\right\|_*\mathrm{d}u}_{I_2} + \left\|\delta^3\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(\delta))\right\|,
\end{aligned}$$

onde o primeiro termo do lado direito, I_1, satisfaz

$$I_1 = \left\|\nabla f(x_0)\right\|_*\int_\delta^t u^3\mathrm{d}u = \frac{t^4-\delta^4}{4}\left\|\nabla f(x_0)\right\|_* \leq \frac{t^4}{4}\left\|\nabla f(x_0)\right\|_*,$$

o segundo termo, I_2, satisfaz

$$\begin{aligned}
I_2 &\leq \int_\delta^t u^3\frac{L}{2\nu}\alpha_\delta(u)u^2\mathrm{d}u = \int_\delta^t u^5\frac{L}{2\nu}\alpha_\delta(u)\mathrm{d}u \\
&\leq \int_\delta^t u^5\frac{L}{2\nu}\alpha_\delta(t)\mathrm{d}u = \frac{L}{2\nu}\alpha_\delta(t)\int_\delta^t u^5\mathrm{d}u \\
&= \frac{t^6-\delta^6}{6}\frac{L}{2\nu}\alpha_\delta(t) \leq \frac{L}{12\nu}\alpha_\delta(t)t^6
\end{aligned}$$

e o último termo satisfaz

$$\left\|\delta^3\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(\delta))\right\| = \delta^4\left\|\frac{\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t))}{\delta}\right\| \leq \delta^4\alpha_\delta(\delta).$$

Desta forma, juntando os últimos quatro resultados, temos enfim que

$$t^3\left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t))\right\| \leq \frac{t^4}{4}\left\|\nabla f(x_0)\right\|_* + \frac{L}{12\nu}\alpha_\delta(t)t^6 + \delta^4\alpha_\delta(\delta).$$

Dividindo os dois lados por t^4, ficamos com

$$\frac{\left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t))\right\|}{t} \leq \frac{1}{4}\left\|\nabla f(x_0)\right\|_* + \frac{L}{12\nu}\alpha_\delta(t)t^2 + \frac{\delta^4}{t^4}\alpha_\delta(\delta).$$

Lembremo-nos agora de que, pelo fato de que $t \geq \delta$ e pelo lema 7.6, $\frac{\delta^4}{t^4}\alpha_\delta(\delta) \leq \alpha_\delta(\delta) \leq \frac{\left\|\nabla f(x_0)\right\|_*}{1-L\delta^2/6\nu}$, o que nos permite escrever

$$\frac{\left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t))\right\|}{t} \leq \frac{1}{4}\left\|\nabla f(x_0)\right\|_* + \frac{L}{12\nu}\alpha_\delta(t)t^2 + \frac{\left\|\nabla f(x_0)\right\|_*}{1-L\delta^2/6\nu}. \tag{7.5}$$

Agora seja $t' \in (0,t]$. Temos dois casos:

- **Caso 1:** $t' \in (\delta, t]$
  Nesse caso, vale a Equação 7.5 para t'. Notando que a expressão à direita é crescente em t' e que $t' \leq t$, escrevemos

$$\begin{aligned}
\frac{\left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t'))\right\|}{t'} &\leq \frac{1}{4}\left\|\nabla f(x_0)\right\|_* + \frac{L}{12\nu}\alpha_\delta(t')t'^2 + \frac{\left\|\nabla f(x_0)\right\|_*}{1-L\delta^2/6\nu} \\
&\leq \frac{1}{4}\left\|\nabla f(x_0)\right\|_* + \frac{L}{12\nu}\alpha_\delta(t)t^2 + \frac{\left\|\nabla f(x_0)\right\|_*}{1-L\delta^2/6\nu}.
\end{aligned}$$

- **Caso 2:** $t' \in (0, \delta]$
  Nesse caso, pela definição de α_δ e pelo lema 7.6, temos

$$\frac{\left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t'))\right\|}{t'} \leq \alpha_\delta(\delta) \leq \frac{\|\nabla f(x_0)\|_*}{1 - L\delta^2/6\nu} \leq \frac{1}{4}\|\nabla f(x_0)\|_* + \frac{L}{12\nu}\alpha_\delta(t)t^2 + \frac{\|\nabla f(x_0)\|_*}{1 - L\delta^2/6\nu}.$$

De toda forma, isto é, para qualquer $t' \in (0, t)$, vale

$$\frac{\left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t'))\right\|}{t'} \leq \frac{1}{4}\|\nabla f(x_0)\|_* + \frac{L}{12\nu}\alpha_\delta(t)t^2 + \frac{\|\nabla f(x_0)\|_*}{1 - L\delta^2/6\nu}.$$

Tomando o supremo em $t' \in (0, t)$, ficamos com

$$\sup_{t' \in (0,t]} \frac{\left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t'))\right\|}{t'} = \alpha_\delta(t) \leq \frac{1}{4}\|\nabla f(x_0)\|_* + \frac{L}{12\nu}\alpha_\delta(t)t^2 + \frac{\|\nabla f(x_0)\|_*}{1 - L\delta^2/6\nu}.$$

Isolando $\alpha_\delta(t)$, temos

$$\begin{aligned}\alpha_\delta(t)(1 - Lt^2/12\nu) &\leq \frac{1}{4}\|\nabla f(x_0)\|_* + \frac{\|\nabla f(x_0)\|_*}{1 - L\delta^2/6\nu} \\ &\leq \left(\frac{1}{4} + \frac{1}{1 - L\delta^2/6\nu}\right)\|\nabla f(x_0)\|_* = \frac{5 - L\delta^2/6\nu}{4(1 - L\delta^2/6\nu)}\|\nabla f(x_0)\|_*.\end{aligned}$$

Como $\sqrt{12\nu/L} > t \implies 1 > Lt^2/12\nu \implies 1 - Lt^2/12\nu \geq 0$, obtemos enfim

$$\alpha_\delta(t) \leq \frac{(5 - L\delta^2/6\nu)\|\nabla f(x_0)\|_*}{4(1 - L\delta^2/6\nu)(1 - Lt^2/12\nu)}.$$ ■

O próximo fato prova a limitação de $\left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_{\delta_m}(t))\right\|$, que será essencial para a equicontinuidade da sequência a ser construída. Na verdade, a limitação da derivada serve de caracterização da equicontinuidade. Ainda assim, vamos preferir usar a definição mesmo.

Lema 7.8. *Para $0 < \delta < \sqrt{6\nu/L}$ e $0 < t \leq \sqrt{6\nu/L}$ quaisquer, temos*

$$\left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_{\delta_m}(t))\right\| \leq 5\sqrt{6\nu/L}\,\|\nabla f(x_0)\|_*.$$

Demonstração. Sejam $0 < \delta < \sqrt{6\nu/L}$ e $0 < t \leq \sqrt{6\nu/L}$. Então $\alpha_\delta(t)t \leq \alpha_\delta\left(\sqrt{6\nu/L}\right)\sqrt{6\nu/L}$. Aí, pelo fato 7.3 e pelo lema 7.7, temos

$$\begin{aligned}\left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t))\right\| \leq \alpha_\delta(t)t \leq \alpha_\delta\left(\sqrt{6\nu/L}\right)\sqrt{6\nu/L} &\leq \frac{(5 - L\delta^2/6\nu)\|\nabla f(x_0)\|_*}{4(1 - L\delta^2/6\nu)(1 - L(\sqrt{6\nu/L})^2/12\nu)}\sqrt{6\nu/L} \\ &= \frac{(5 - L\delta^2/6\nu)\|\nabla f(x_0)\|_*}{4(1 - L\delta^2/6\nu)(1 - 1/2)}\sqrt{6\nu/L} \\ &= \frac{(5 - L\delta^2/6\nu)\|\nabla f(x_0)\|_*}{2(1 - L\delta^2/6\nu)}\sqrt{6\nu/L}.\end{aligned} \tag{7.6}$$

Agora notemos que $L\delta^2/6\nu \leq \frac{L}{6\nu}(\sqrt{3\nu/L})^2 = 1/2$, o que implica em $5 - L\delta^2/6\nu \leq 5$ e $\frac{1}{1 - L\delta^2/6\nu} \leq \frac{1}{1/2}$, de forma que

$$\left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_\delta(t))\right\| \leq \frac{(5 - L\delta^2/6\nu)\|\nabla f(x_0)\|_*}{2(1 - L\delta^2/6\nu)}\sqrt{6\nu/L} \leq \frac{5\|\nabla f(x_0)\|_*}{2(1/2)}\sqrt{6\nu/L} = 5\sqrt{6\nu/L}\,\|\nabla f(x_0)\|_*.$$ ■

A seguir exibimos as definições de equicontinuidade e limitação uniforme para famílias de funções, bem como o Teorema de Ascoli-Arzelà (sem prová-lo) que as utiliza em suas hipóteses. Lembremo-nos que usaremos o teorema para provar que uma certa sequência de soluções de EDOs suavizadas tem uma subsequência convergente, cujo limite será nosso candidato a solução da EDO original.

Definição (Equicontinuidade)**.** *Uma família de funções* $\mathcal{F} = \{F_m : I \to \mathcal{D}\}_{m\in\mathbb{N}}$ *é* ***equicontínua*** *se*

$$\forall m \in \mathbb{N},\ \forall \epsilon > 0,\ \exists \gamma > 0,\ \forall t_1, t_2 \in I,\ |t_1 - t_2| < \gamma \quad \Longrightarrow \quad \|F_m(t_1) - F_m(t_2)\| < \epsilon.$$

Definição (Equicontinuidade uniforme)**.** *Uma família de funções* $\mathcal{F} = \{F_m : I \to \mathcal{D}\}_{m\in\mathbb{N}}$ *é* ***uniformemente equicontínua*** *se*

$$\forall \epsilon > 0,\ \exists \gamma > 0,\ \forall m \in \mathbb{N},\ \forall t_1, t_2 \in I,\ |t_1 - t_2| < \gamma \quad \Longrightarrow \quad \|F_m(t_1) - F_m(t_2)\| < \epsilon.$$

Equicontinuidade é uma versão mais forte da continuidade para cada função na família. Enquanto a continuidade apenas exige que cada função se comporte bem em cada ponto, a equicontinuidade exige que todas elas se comportem bem simultaneamente em cada ponto; para a continuidade γ pode depender do ponto x e da função F_m, enquanto para a equicontinuidade, ele pode depender apenas de x. A equicontinuidade uniforme é ainda mais forte: exige que todas as funções sejam bem comportadas da mesma maneira (com mesmas constantes ϵ, γ) no domínio todo; para a equicontinuidade uniforme γ não depende nem de x nem de F_m. A equicontinuidade uniforme implica na equicontinuidade, que por sua vez implica na continuidade.

Vamos provar a seguir que a sequência de soluções é uniformemente equicontínua, mas só vamos precisar da equicontinuidade para usar o teorema de Ascoli-Arzelà.

Antes, consideremos ainda uma outra definição importante para o teorema.

Definição (Limitação uniforme)**.** *Uma família de funções* $\mathcal{F} = \{F_m : I \to \mathcal{D}\}_{m\in\mathbb{N}}$ *é* ***uniformemente limitada*** *se*

$$\exists G, \forall m \in \mathbb{N}, \forall t \in I, \|F_m(t)\| \leq G.$$

A limitação uniforme é mais forte do que a limitação pontual, pois exige que uma mesma constante limite todas as funções em todos os pontos simultaneamente.

Agora podemos enunciar o teorema de Ascoli-Arzelà.

Teorema 7.9. *Seja* $\mathcal{F} = \{F_m : I \to \mathcal{D}\}_{m\in\mathbb{N}}$ *uma família de funções, equicontínua e uniformemente limitada. Então existe uma subsequência* $\{F_{m_k}\}_{k\in\mathbb{N}}$ *e uma função contínua* F *tal que essa subsequência converge uniformemente para* F*.*

No próximo lema apresentamos então a sequência de soluções e provamos que ela é equicontínua e uniformemente limitada.

Lema 7.10. *A família de funções* $\mathcal{F} = \{\Theta(X_{\delta_m}(\cdot)) : [0, \sqrt{6\nu/L}] \to \mathbb{R}^n | \delta_m = \sqrt{3/L}/2^m\}_{m\in\mathbb{N}}$ *é uniformemente limitada e equicontínua.*

Demonstração. Seja $\epsilon > 0$ qualquer e tomemos γ de forma que $\left(\leq 5\sqrt{6\nu/L}\left\|\nabla f(x_0)\right\|_*\right)\gamma = \epsilon$.

Seja $m \in \mathbb{N}$ qualquer; temos $\delta_m < \sqrt{6\nu/L}$. Seja também $m \in \mathbb{N}$ qualquer e $t_1, t_2 \in \left[0, \sqrt{6\nu/L}\right]$ tais que $|t_1 - t_2| < \gamma$. Considerando a função $\Theta(X_{\delta_m})$, pelo Teorema do Valor Médio e usando o lema 7.8, existe um certo $t' \in [t_1, t_2]$ tal que

$$\|\Theta(X_{\delta_m}(t_1)) - \Theta(X_{\delta_m}(t_2))\| \leq \left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_{\delta_m}(t'))\right\| |t_1 - t_2| < \left(5\sqrt{6\nu/L}\left\|\nabla f(x_0)\right\|_*\right)\gamma = \epsilon.$$

A família $\mathcal{F}$ é uniformemente equicontínua e, portanto, equicontínua.

Agora sejam $G = 30\nu\left\|\nabla f(x_0)\right\|_*/L + \left\|\Theta(x_0)\right\|_*$ e $m \in \mathbb{N}$ qualquer mais uma vez. Temos

$$\begin{aligned}
\left\|\Theta(X_{\delta_m}(t))\right\|_* &\leq \left\|\Theta(X_{\delta_m}(t))\right\|_* - \left\|\Theta(X_{\delta_m}(0))\right\|_* + \left\|\Theta(X_{\delta_m}(0))\right\|_* \\
&\leq \left\|\Theta(X_{\delta_m}(t)) - \Theta(X_{\delta_m}(0))\right\|_* + \left\|\Theta(X_{\delta_m}(0))\right\|_* \\
&= \left\|\int_0^t \frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_{\delta_m}(u))\mathrm{d}u\right\| + \left\|\Theta(x_0)\right\|_* \\
&\leq \int_0^t \left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_{\delta_m}(u))\right\| \mathrm{d}u + \left\|\Theta(x_0)\right\|_* \\
&\leq \int_0^{\sqrt{6\nu/L}} \left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_{\delta_m}(u))\right\| \mathrm{d}u + \left\|\Theta(x_0)\right\|_*.
\end{aligned}$$

Agora, usando o lema 7.8 novamente,

$$\begin{aligned}\|\Theta(X_{\delta_m}(t))\|_* &\leq \int_0^{\sqrt{6\nu/L}} \left\| \frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_{\delta_m}(u)) \right\| \mathrm{d}u + \|\Theta(x_0)\|_* \\ &\leq \int_0^{\sqrt{6\nu/L}} 5\sqrt{6\nu/L}\,\|\nabla f(x_0)\|_* \,\mathrm{d}u + \|\Theta(x_0)\|_* \\ &= 5\sqrt{6\nu/L}\sqrt{6\nu/L}\,\|\nabla f(x_0)\|_* + \|\Theta(x_0)\|_* \\ &= 30\nu\,\|\nabla f(x_0)\|_*/L + \|\Theta(x_0)\|_*\,.\end{aligned}$$

Assim, a família $\mathcal{F}$ é uniformemente limitada. ■

Por fim, com o próximo lema provamos o resultado mais importante dessa seção: a existência de pelo menos uma solução para 6.1 com condições iniciais $X(0) = x_0$ qualquer e $\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(0)) = \vec{0}$.

Lema 7.11. *Sejam $f \in \mathcal{F}_L$ uma função diferenciável L-Lipschitz qualquer e $x_0 \in \mathbb{R}^n$ um ponto qualquer. Então a EDO 6.1 com condições iniciais $X(0) = x_0$ e $\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(0)) = \vec{0}$ tem pelo menos uma solução $X \in C^2(0,\infty) \cup C^1[0,\infty)$*

Demonstração. Consideremos a família de funções

$$\mathcal{F} = \{\Theta(X_{\delta_m}(\cdot)) : [0, \sqrt{6\nu/L}] \to \mathbb{R}^n | \delta_m = \sqrt{3/L}/2^m\}_{m\in\mathbb{N}}$$

e notemos que a sequência $\{\delta_m\}_{m\in\mathbb{N}}$ é decrescente com $\lim_{m\to\infty}\delta_m = 0$.

Pelo Lema 7.10, essa família $\mathcal{F}$ é equicontínua e uniformemente limitada em $[0, \sqrt{6\nu/L}]$. Assim, pelo Teorema de Arzelà-Ascoli, existe uma subsequência $\{\Theta(X_{\delta_{m_i}}(\cdot))\}_{i\in\mathbb{N}}$ que converge uniformemente em $[0, \sqrt{6\nu/L}]$ para alguma função que denotamos por $\Theta(\hat{X}(\cdot))$. Vamos provar agora que $\Theta(\hat{X}(\cdot))$ satisfaz a EDO no intervalo $(0, \sqrt{6\nu/L})$.

Como visto no lema 7.10, esta família é equicontínua e uniformemente limitada, o que, pelo Teorema de Ascoli-Arzelà, garante que $\mathcal{F}$ contém uma subsequência uniformemente convergente em $[0, \sqrt{6\nu/L}]$. Denotemos esta subsequência por $\{\Theta(X_{\delta_{m_i}}(\cdot))\}_{i\in\mathbb{N}}$ e de $\hat{\Theta}$ o seu limite. Como $\Theta(\cdot) = (\nabla\Phi)(\cdot)$ é um difeomorfismo, existe inversa e podemos definir $\hat{X} = (\nabla\Phi)^{-1}(\hat{\Theta})$ e garantir que $\hat{X} \in C^2$. Visando uniformidade de notação, passaremos a escrever $\Theta(\hat{X})$ em vez de $\hat{\Theta}$.

Vamos provar que essa função satisfaz a EDO com as condições iniciais especificadas em $(0, \sqrt{6\nu/L})$. Para isso precisamos provar que ela coincide com qualquer solução local da EDO em $(0, \sqrt{6\nu/L})$ e satisfaz as condições iniciais para $t = 0$.

Seja $t_0 \in (0, \sqrt{6\nu/L})$ qualquer. Como $\left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_{\delta_{m_i}}(t_0))\right\|$ é limitado (lema 7.10), pelo Teorema de Bolzano-Weierstrass, podemos escolher uma subsequência de $\{\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_{\delta_{m_i}}(t_0))\}_{i\in\mathbb{N}}$ que converge para um certo limite que denotaremos por $Z(t_0)$. Sem perda de generalidade, vamos assumir que essa subsequência é a mesma, $\left\{\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_{\delta_{m_i}}(t_0))\right\}_{i\in\mathbb{N}}$. Assim, $\Theta(X_{\delta_{m_i}}(t_0))$ converge para $\Theta(\hat{X}(t_0))$ e $\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_{\delta_{m_i}}(t_0))$ converge para $Z(t_0)$ e podemos escrever

$$\lim_{i\to\infty}\left(\Theta(X_{\delta_{m_i}}(t_0)), \frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_{\delta_{m_i}}(t_0))\right) = \left(\Theta(\hat{X}(t_0)), Z(t_0)\right). \tag{7.7}$$

Seja $\tilde{X}$ a solução local em torno de t_0, isto é, em $(t_0 - \varepsilon_0, t_o + \varepsilon_0)$ para algum $\epsilon_0 > 0$, da EDO 6.1 com condições iniciais

$$X(t_0) = \hat{X}(t_0) \quad \text{e} \quad \frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t_0)) = Z(t_0).$$

Com i grande o suficiente, $\delta_{m_i} \leq t_0 - \varepsilon_0$ e aí as EDOs suavizadas com parâmetro δ_{m_i} são iguais à EDO 6.1 no intervalo $(t_0 - \varepsilon_0, t_o + \varepsilon_0)$. Desta maneira, pelo teorema da dependência contínua das condições iniciais, e usando 7.7, temos

$$\sup_{t\in(t_0-\varepsilon_0,t_0+\varepsilon_0)} \left\|X_{\delta_{m_i}}(t) - \tilde{X}(t)\right\| \xrightarrow[i\to\infty]{} 0.$$

Como $\Theta(\cdot)$ é contínua,

$$\sup_{t\in(t_0-\varepsilon_0,t_0+\varepsilon_0)} \left\|\Theta(X_{\delta_{m_i}}(t)) - \Theta(\tilde{X}(t))\right\|_* \xrightarrow[i\to\infty]{} 0.$$

Mas pela definição de convergência uniforme, temos também que

$$\sup_{t\in(t_0-\epsilon_0,t_0+\epsilon_0)} \left\|\Theta(X_{\delta_{m_i}}(t)) - \Theta(\hat{X}(t))\right\|_* \xrightarrow[i\to\infty]{} 0.$$

Juntando esses dois últimos limites, obtemos enfim que

$$\sup_{t\in(t_0-\epsilon_0,t_0+\epsilon_0)} \left\|\Theta(\hat{X}(t)) - \Theta(\tilde{X}(t))\right\|_* \xrightarrow[i\to\infty]{} 0,$$

Ou seja, $\Theta(\hat{X}(t))$ e $\Theta(\tilde{X}(t))$ são iguais em $(t_0-\epsilon_0, t_0+\epsilon_0)$. De novo, como Θ é um difeomorfismo e portanto inversível, $\hat{X}$ e $\tilde{X}$ devem ser iguais em $(t_0-\epsilon_0, t_0+\epsilon_0)$. Desta maneira, $\hat{X}$ satisfaz 6.1 em t_0. Como t_0 é qualquer, isto na verdade vale para todo o intervalo $(0, \sqrt{6\nu/L})$.

Como o intervalo local $(0, \sqrt{6\nu/L}]$ é independente do ponto inicial, uma extensão sucessiva padrão garante que $\hat{X}$ é solução global em $(0,\infty)$, restando verificar as condições iniciais. A primeira delas, $\hat{X}(0) = x_0$ é consequência de $X_{\delta_{m_i}}(0) = x_0$. Agora, para a segunda, consideremos $t > 0$ qualquer e notemos que

$$\begin{aligned}
\frac{\left\|\Theta(\hat{X}(t)) - \Theta(\hat{X}(0))\right\|_*}{t} &= \lim_{i\to\infty} \frac{\left\|\Theta(X_{\delta_{m_i}}(t)) - \Theta(X_{\delta_{m_i}}(0))\right\|_*}{t} \\
&= \lim_{i\to\infty} \left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_{\delta_{m_i}}(t^*))\right\| \\
&\leq \lim_{i\to\infty} t^*\alpha_{\delta_{m_i}}(t^*) \leq \lim_{i\to\infty} t\alpha_{\delta_{m_i}}(t) \leq 5t\left\|\nabla f(x_0)\right\|_*,
\end{aligned}$$

onde $t^* \in (0,t)$ é obtido via Teorema do Valor Médio e a última desigualdade é dada pelo lema 7.10. Agora, fazendo $t \to 0$, obtemos enfim que

$$\left\|\frac{\mathrm{d}}{\mathrm{d}t}\Theta(\hat{X}(0))\right\| = \lim_{t\to 0} \frac{\left\|\Theta(\hat{X}(t)) - \Theta(\hat{X}(0))\right\|_*}{t} \leq \lim_{t\to 0} 5t\left\|\nabla f(x_0)\right\|_* = 0 \quad\Longrightarrow\quad \frac{\mathrm{d}}{\mathrm{d}t}\Theta(\hat{X}(0)) = 0,$$

o que configura a segunda condição.

Isso prova que $\hat{X}$ é uma solução global da 6.1 com condições iniciais $X(0) = x_0$ e $\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(0)) = \vec{0}$. ■

7.1.2 Unicidade

Para provar a unicidade da solução, a ideia é supor duas soluções quaisquer da EDO e mostrar que elas são iguais, de forma que toda solução seja a mesma, havendo, na verdade, apenas uma solução. Para isso, mostraremos que dadas duas possíveis soluções, a diferença entre elas é nula. Começamos com a definição de um objeto auxiliar e dois fatos imediatos relativos a ele (análogos aos da seção anterior).

Definição. *Sejam X e Y duas possíveis soluções de 6.1 com mesmas condições iniciais $X(0) = Y(0) = x_0$ e $\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(0)) = \frac{\mathrm{d}}{\mathrm{d}t}\Theta(Y(0)) = \vec{0}$. Definimos*

$$\beta(t) = \sup_{u\in(0,t)} \left\|\frac{\mathrm{d}}{\mathrm{d}t}X(u) - \frac{\mathrm{d}}{\mathrm{d}t}Y(u)\right\|.$$

Aqui, $\beta(t)$ representa a maior diferença entre duas soluções da EDO. Para provar que as soluções são iguais, devemos provar que a diferença entre elas é sempre nula, ou seja, $\beta(t) = 0$.

Fato 7.12. *Para todo $u < t$, temos*

$$\beta(u) \leq \beta(t).$$

Demonstração. Esta demonstração se dá, assim como no fato 7.2, observando que os dois objetos são supremos de uma mesma expressão tomados em intervalos diferentes, um contendo o outro. ■

Fato 7.13. *Para todo $u \in (0,t)$, temos*

$$\left\| \frac{\mathrm{d}}{\mathrm{d}t} X(u) - \frac{\mathrm{d}}{\mathrm{d}t} Y(u) \right\| \leq \beta(t).$$

Demonstração. Seja $u \in (0,t)$. Então

$$\left\| \frac{\mathrm{d}}{\mathrm{d}t} X(u) - \frac{\mathrm{d}}{\mathrm{d}t} Y(u) \right\| \leq \sup_{u \in (0,t)} \left\| \frac{\mathrm{d}}{\mathrm{d}t} X(u) - \frac{\mathrm{d}}{\mathrm{d}t} Y(u) \right\| = \beta(t). \qquad \blacksquare$$

Agora provamos um rápido lema auxiliar e em seguida demonstramos o lema principal. O lema estabelece a desigualdade $\|\nabla f(X(t)) - \nabla f(Y(t))\|_* \leq Lt\beta(t)$ que futuramente deve nos levar à desigualdade $\beta(t) \leq \frac{1}{5} L^2 t \beta(t)$, que implicará, sob uma certa condição, que $\beta(t) = 0$.

Lema 7.14. *Sejam X e Y funções diferenciáveis definidas sobre um intervalo $(0,\tau)$. Então para $t \in (0,\tau)$ qualquer temos*

$$\|\nabla f(X(t)) - \nabla f(Y(t))\|_* \leq Lt\beta(t).$$

Demonstração. Seja $t \in (0,\tau)$ qualquer. Pela Lipschitz continuidade da f e usando o fato 7.13, temos

$$\begin{aligned}
\|\nabla f(X(t)) - \nabla f(Y(t))\|_* &\leq L \|X(t) - Y(t)\| \\
&= L \|X(t) - Y(t) - X(0) + Y(0) + X(0) - Y(0)\| \\
&= L \left\| (X(u) - Y(u))\big|_0^t + x_0 - x_0 \right\| \\
&= L \left\| \int_0^t \frac{\mathrm{d}}{\mathrm{d}v} X(u) - \frac{\mathrm{d}}{\mathrm{d}v} Y(u) \mathrm{d}u \right\| \\
&\leq L \int_0^t \left\| \frac{\mathrm{d}}{\mathrm{d}v} X(u) - \frac{\mathrm{d}}{\mathrm{d}v} Y(u) \right\| \mathrm{d}u \\
&\leq L \int_0^t \beta(u) \mathrm{d}u \leq L \int_0^t \beta(t) \mathrm{d}u = Lt\beta(t). \qquad \blacksquare
\end{aligned}$$

Lema 7.15. *Sejam $f \in \mathcal{F}_L$ uma função diferenciável L-Lipschitz qualquer e $x_0 \in \mathbb{R}^n$ um ponto qualquer. Então a EDO 6.1 com condições iniciais $X(0) = x_0$ e $\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(0)) = \vec{0}$ tem no máximo uma solução $X \in C^2(0,\tau) \cup C^1[0,\tau)$, $0 < \tau < \sqrt{5/L}$.*

Demonstração. Suponhamos que haja duas soluções $X, Y \in C^2(0,\tau) \cup C^1[0,\tau)$ definidas em $(0,\tau)$ para algum $\tau > 0$. Seja $t \in (0,\tau)$.

Como X e Y são satisfazem 6.1, de maneira quase igual à feita na demonstração do lema 7.7 para derivar a equação 7.4, podemos obter

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}u} u^3 \left(\frac{\mathrm{d}}{\mathrm{d}u}\Theta(X(u)) - \frac{\mathrm{d}}{\mathrm{d}u}\Theta(Y(u)) \right) &= \frac{\mathrm{d}}{\mathrm{d}u} \left(u^3 \frac{\mathrm{d}}{\mathrm{d}u}\Theta(X(u)) \right) - \frac{\mathrm{d}}{\mathrm{d}u} \left(u^3 \frac{\mathrm{d}}{\mathrm{d}u}\Theta(Y(u)) \right) \\
&= -u^3 \nabla f(X(u)) + u^3 \nabla f(Y(u)) \\
&= -u^3 (\nabla f(X(u) - \nabla f(Y(u))) \qquad .
\end{aligned}$$

Integrando os dois lados de 0 a t ficamos com

$$\begin{aligned}
-\int_0^t u^3 (\nabla f(X(u) - \nabla f(Y(u))) \mathrm{d}u &= \int_0^t \frac{\mathrm{d}}{\mathrm{d}u} u^3 \left(\frac{\mathrm{d}}{\mathrm{d}u}\Theta(X(u)) - \frac{\mathrm{d}}{\mathrm{d}u}\Theta(Y(u)) \right) \mathrm{d}u \\
&= t^3 \left(\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t)) - \frac{\mathrm{d}}{\mathrm{d}t}\Theta(Y(t)) \right).
\end{aligned}$$

Agora usando o lema 7.14, desigualdade triangular para integrais e o fato 7.12, obtemos

$$\begin{aligned} t^3 \left\| \frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t)) - \frac{\mathrm{d}}{\mathrm{d}t}\Theta(Y(t)) \right\| &= \left\| t^3 \left(\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t)) - \frac{\mathrm{d}}{\mathrm{d}t}\Theta(Y(t)) \right) \right\| \\ &= \left\| \int_0^t u^3 (\nabla f(X(u) + \nabla f(Y(u))) \mathrm{d}u \right\| \\ &\leq \int_0^t \left\| u^3 (\nabla f(X(u) + \nabla f(Y(u))) \right\|_* \mathrm{d}u \\ &= \int_0^t u^3 \left\| \nabla f(X(u) + \nabla f(Y(u)) \right\|_* \mathrm{d}u \\ &\leq \int_0^t Lu^3 u\beta(u) \mathrm{d}u \leq \int_0^t Lu^3 u\beta(t) \mathrm{d}u \\ &= L\beta(t) \int_0^t u^4 \mathrm{d}u = \frac{1}{5} Lt^5 \beta(t), \end{aligned}$$

ou ainda, dividindo tudo por t^3,

$$\left\| \frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t)) - \frac{\mathrm{d}}{\mathrm{d}t}\Theta(Y(t)) \right\| \leq \frac{1}{5} Lt^2 \beta(t).$$

Seja $t' \leq t$. Usando novamente o fato 7.12, temos ainda que

$$\left\| \frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t')) - \frac{\mathrm{d}}{\mathrm{d}t}\Theta(Y(t')) \right\| \leq \frac{1}{5} Lt'^2 \beta(t') \leq \frac{1}{5} Lt^2 \beta(t).$$

Agora tomando o supremo em $t' \in (0, t)$, obtemos

$$\sup_{t' \in (0,t)} \left\| \frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t')) - \frac{\mathrm{d}}{\mathrm{d}t}\Theta(Y(t')) \right\| \leq \sup_{t' \in (0,t)} \frac{1}{5} Lt^2 \beta(t) \quad \Longrightarrow \quad \beta(t) \leq \frac{1}{5} Lt^2 \beta(t).$$

Se $\beta(t) \neq 0$, então $\beta(t) \leq \frac{1}{5} Lt^2 \beta(t) \Longrightarrow 1 \leq \frac{1}{5} Lt^2 \Longrightarrow t^2 \geq 5/L$. Assim, como $t \leq \tau < \sqrt{5/L}$, temos que ter que $\beta(t) = 0$. Mas isso quer dizer que $\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t))$ e $\frac{\mathrm{d}}{\mathrm{d}t}\Theta(Y(t))$ são iguais em $(0, \tau)$. Com o mesmo ponto inicial $\Theta(X(0)) = \Theta(Y(0)) = \Theta(x_0) = \theta_0$ e mesmo gradiente, temos

$$\begin{aligned} \Theta(X(t)) &= \theta_0 + \int_0^\tau \frac{\mathrm{d}}{\mathrm{d}u}\Theta(X(u)) \mathrm{d}u \\ &= \theta_0 + \int_0^\tau \frac{\mathrm{d}}{\mathrm{d}u}\Theta(Y(u)) \mathrm{d}u = \Theta(Y(t)). \end{aligned}$$

Daí concluimos que $\Theta(X)$ e $\Theta(Y)$ devem ser iguais em $(0, \tau)$ e, como $\Theta(\cdot)$ é um difeomorfismo, X e Y também devem ser iguais em $(0, \tau)$. Como X e Y são soluções quaisquer de 6.1 em $(0, \tau)$, concluimos que toda solução deve ser a mesma em em $(0, \tau)$. ■

7.2 Convergência do AMD

Nesta seção provamos a convergência da trajetória produzida pelo método AMD para a solução da EDO (6.1), resultado enunciado pelo teorema a seguir.

Teorema 7.16 (Convergência do AMD para a solução da EDO associada)**.** *Sejam $f \in \mathcal{F}_L$ uma função diferenciável L-Lipschitz com um mínimo global $y^* \in \mathbb{R}^n$ e $x_0 \in \mathbb{R}^n$ um ponto qualquer. Sejam ainda $X : \mathbb{R}^n_+ \to \mathbb{R}^n$ a solução da EDO (6.1) com condições iniciais $X(0) = x_0$ e $\frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(0)) = \vec{0}$ e $\{x_k\}_\mathbb{N}$ a sequência de pontos gerada pelo AMD com ponto inicial x_0 e passo s. Então para qualquer $T \in R_+$,*

$$\lim_{s \to 0} \sup_{k \in [T/\sqrt{s}]} \left\| x_k - X(k\sqrt{s}) \right\| = 0 \ .$$

Demonstração. A demonstração se dá seguindo os resultados a partir daqui até o fim deste capítulo, que culminam no lema 7.37. ■

A ideia principal da demonstração é explorar:

1. a convergência uniforme quando $\delta \to 0$ dos pontos $\{X_\delta(k\sqrt{s})\}_{k\in\mathbb{N}}$ gerados pelas soluções suavizadas para os pontos $\{X(k\sqrt{s})\}_{k\in\mathbb{N}}$ gerados pela solução real;
2. a convergência uniforme quando $\sqrt{s} \to 0$ dos pontos gerados pelas soluções suavizadas para os pontos $\{X_k^\delta\}_{k\in\mathbb{N}}$ gerados pelas suas discretivações obtidas pelo método de Euler Foward com passo $\sqrt{s}$; e
3. a convergência uniforme quando $\sqrt{s} \to 0$ e $\delta \to 0$ dos pontos gerados por essas discretizações para os pontos $\{x_k\}_{k\in\mathbb{N}}$ gerados pelo AMD,

concluindo que deve haver uma convergência uniforme dos pontos gerados pelo AMD para os gerados pela solução da EDO. As duas primeiras convergências já estão garantidas: a primeira pela demonstração da seção anterior e a segunda pelo método de Euler Forward. Vamos nos dedicar a provar a terceira.

Relembremo-nos que a EDO suavizada com parâmetro δ pode ser escrita, como em 7.2, como

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\Theta(X(t))\\ Z(t)\end{pmatrix} = \begin{pmatrix} Z(t)\\ -\frac{3}{\max(\delta,t)}Z(t) - \nabla f(X(t))\end{pmatrix},$$

onde $Z(t) = \frac{\mathrm{d}}{\mathrm{d}t}\Theta(X(t))$. Aplicando a ela um método de Euler Forward com passo $\sqrt{s}$, obtemos

$$\begin{cases}\Theta(X_{k+1}^\delta) & = \Theta(X_k^\delta) + \sqrt{s}Z_k^\delta \\ Z_{k+1}^\delta & = Z_k^\delta - \sqrt{s}\left(\frac{3\sqrt{s}}{\max(\delta,k\sqrt{s})}Z_k^\delta - \nabla f(X_k^\delta)\right) = \left(1 - \frac{3\sqrt{s}}{\max(\delta,k\sqrt{s})}\right)Z_k^\delta - \sqrt{s}\nabla f(X_k^\delta)\ .\end{cases} \tag{7.8}$$

O próximo lema introduz uma nova formulação para o AMD com passo s, provando sua equivalência à dada pela equação (6.1), apresentada na seção 6.1. Essa nova formulação introduz uma variável z_k que fará paralelo com Z_k^δ e nos permitirá definir logo depois um objeto que vai nos permitir explorar melhor algumas propriedades oportunas sobre as diferenças de X_k^δ e x_k.

Lema 7.17. *Seja $s > 0$ qualquer. Com $z_k = \frac{\theta_{k+1}-\theta_k}{\sqrt{s}}$ e $z_0 = \vec{0}$, o esquema do AMD com passo s que usamos é equivalente a*

$$\begin{cases} x_{k+1} & = (\nabla\Phi)^{-1}(\theta_{k+1}) \\ \theta_{k+1} & = \theta_k + \sqrt{s}z_k \\ z_{k+1} & = \frac{k-1}{k+2}z_k - \sqrt{s}\nabla f(y_k) \\ y_k & = (\nabla\Phi)^{-1}\left(\theta_k + \frac{k-1}{k+2}\sqrt{s}z_k\right)\ .\end{cases} \tag{7.9}$$

Demonstração. Pela definição, o método do AMD com passo s é dado por

$$\begin{cases} x_{k+1} & = (\nabla\Phi)^{-1}(\theta_{k+1}) \\ \theta_{k+1} & = \xi_k - s\nabla f(y_k) \\ y_k & = (\nabla\Phi)^{-1}(\xi_k) \\ \xi_k & = \theta_k + \frac{k-1}{k+2}(\theta_k - \theta_{k-1}).\end{cases}$$

Substituindo a última equação na segunda, obtemos

$$\theta_{k+1} = \theta_k + \frac{k-1}{k+2}(\theta_k - \theta_{k-1}) - s\nabla f(y_k) \quad\Longrightarrow\quad \theta_{k+1} - \theta_k = \frac{k-1}{k+2}(\theta_k - \theta_{k-1}) - s\nabla f(y_k).$$

Seja $z_k = \frac{\theta_{k+1}-\theta_k}{\sqrt{s}} \iff \sqrt{s}z_k = \theta_{k+1} - \theta_k$. Temos então que

$$\sqrt{s}z_k = \frac{k-1}{k+2}\sqrt{s}z_{k-1} - s\nabla f\left(y_k\right) \quad\Longrightarrow\quad z_k = \frac{k-1}{k+2}z_{k-1} - \sqrt{s}\nabla f(y_k).$$

Desta maneira, o AMD é equivalente a

$$\begin{cases} x_{k+1} &= (\nabla\Phi)^{-1}(\theta_{k+1}) \\ \theta_{k+1} &= \theta_k + \sqrt{s}z_k \\ z_{k+1} &= \frac{k-1}{k+2}z_k - \sqrt{s}\nabla f(y_k) \\ y_k &= (\nabla\Phi)^{-1}\left(\theta_k + \frac{k-1}{k+2}(\theta_k - \theta_{k-1})\right) = (\nabla\Phi)^{-1}\left(\theta_k + \frac{k-1}{k+2}\sqrt{s}z_{k-1}\right). \end{cases}$$

Quando $k = 1$, não importa muito o que é z_k pois o termo $\frac{k-1}{k+2}$, que vai aparecer multiplicando-o é igual a 0. De qualquer forma, observando que o termo pode ser visto e pensado com uma espécie de vetor velocidade, que na interpretação de inércia deve ser nula no começo do método, vamos fazer $z_0 = \vec{0}$ por consistência. Essa convenção também será mais útil num corolário adiante. ■

Aqui vale lembrar que da derivação informal da EDO do AGD, dada por (4.1), descobrimos que a EDO e e o esquema AGD "andam" em escalas de tempo diferentes. Teríamos descoberto isso também para o AMD se tivéssemos feito a derivação informal completa para sua EDO. De qualquer forma, podemos observar esse fenômeno aqui agora. Enquanto a discretização via Euler Forward, que segue a mesma escala da EDO, anda com um tamanho de passo s o AMD anda com um tamanho de passo $\sqrt{s}$; aqui as direções dos passos são dadas por Z_k^δ e z_k.

Definidos os dois métodos dessas maneiras, partimos agora para a definição de alguns objetos auxiliares e alguns fatos sobre eles.

Definição. *Sejam $\{x_k\}_{k\in\mathbb{N}}$ e $\{z_k\}_{k\in\mathbb{N}}$ sequências geradas pelo AMD a partir da formulação 7.9 e $\{X_k^\delta\}_{k\in\mathbb{N}}$ e $\{Z_k^\delta\}_{k\in\mathbb{N}}$ sequências geradas pelo método de Euler Forward 7.8. Definimos as discrepâncias iterativas relativas à posição e à velocidade entre os dois métodos no passo k como*

$$p_k = \left\|\Theta(X_k^\delta) - \theta_k\right\|_* \qquad e \qquad v_k = \left\|Z_k^\delta - z_k\right\|$$

respectivamente. Definimos também a discrepância acumulada relativa à velocidade como

$$S_k = v_0 + v_1 + \cdots + v_k = \sum_{i=0}^{k} v_i.$$

Fato 7.18. *Dadas as definições anteriores, temos que para $\delta, s > 0$ quaisquer e para todo $k \in \mathbb{N}$ vale*

$$p_k \leq p_{k-1} + \sqrt{s}v_{k-1} \quad e \quad p_k \leq \sqrt{s}S_{k-1}.$$

Demonstração. Seja $k \in \mathbb{N}$ qualquer. Provamos o primeiro resultado fazendo

$$\begin{aligned} p_k &= \left\|\Theta(X_k^\delta) - \theta_k\right\| \\ &= \left\|\Theta(X_{k-1}^\delta) + \sqrt{s}Z_{k-1}^\delta - \theta_{k-1} - \sqrt{s}z_{k-1}\right\| \\ &= \left\|\Theta(X_{k-1}^\delta) - \theta_{k-1} + \sqrt{s}(Z_{k-1}^\delta - z_{k-1})\right\| \\ &\leq \left\|\Theta(X_{k-1}^\delta) - \theta_{k-1}\right\| + \sqrt{s}\left\|Z_{k-1}^\delta - z_{k-1}\right\| \\ &= p_{k-1} + \sqrt{s}v_{k-1}. \end{aligned}$$

O segundo resultado é provado por indução fraca em k. Para o caso base ($k = 1$), usando o primeiro resultado e o fato de que $p_0 = \left\|\Theta(X_0^\delta) - \theta_0\right\| = \|\theta_0 - \theta_0\| = 0$, temos

$$p_1 \leq p_0 + \sqrt{s}v_0 = \sqrt{s}v_0 = \sqrt{s}S_0.$$

Para o passo indutivo, suponhamos que valha que $p_k \leq \sqrt{s}S_{k-1}$ para algum $k \in \mathbb{N}$. Assim, usando novamente o primeiro resultado, temos

$$p_{k+1} \leq p_k + \sqrt{s}v_k \leq \sqrt{s}S_{k-1} + \sqrt{s}v_k = \sqrt{s}S_k,$$

o que conclui nossa prova. ■

p_k representa a distância entre as trajetórias feitas pelo método de Euler e pelo AMD e v_k representa uma espécie de diferença (a menos de uma constante $\sqrt{s}$) entre os vetores velocidade desses métodos.

Se não fossem as normas nas definições de p_k e v_k, S_k representaria exatamente a diferença entre $\Theta(X_k^\delta)$ e θ_k; para ver isso basta notar que na demonstração anterior, sem as normas a desigualdade pode ser substituída por uma igualdade. Com a norma, no entanto, esse resultado é mais útil, pois para cada k, S_k funciona como uma cota superior não apenas para a diferença entre as trajetórias na k-ésima iteração, mas para todas as iterações até a k-ésima. Isso é provado no corolário a seguir.

Corolário 7.19. *Para $\delta, s > 0$ quaisquer e $K \in \mathbb{N}$, temos*

$$\max_{k\in K} p_k \leq \max_{k\in K} \sqrt{s} S_{k-1} = \sqrt{s} S_{K-1}.$$

Demonstração. Para a desigualdade, basta tomar o máximo dos dois lados da equação obtida pelo fato anterior. Para a igualdade, basta notar que $\{S_k\}_{k\in\mathbb{N}}$ é não decrescente. ■

Agora, seja $T > 0$ um tempo qualquer. Para o restante das definições e dos resultados consideremos as variáveis

$$k^*_{\delta,s} = \lfloor \delta/\sqrt{s} \rfloor \quad \text{e} \quad K^*_s = \lfloor T/\sqrt{s} \rfloor.$$

Aqui $k^*_{\delta,s}$ é o número da última iteração (do esquema de Euler) para a qual a EDO suavizada com parâmetro δ é diferente da EDO original e K^*_s é a última iteração para a qual os pontos gerados correspondem a pontos da solução da EDO em $R^n \times (0, T]$.

Com essas definições, observemos que provar a terceira convergência se reduz a provar que o limite de $\max_{k\in K^*_s} p_k$ quando $s \to 0$ e $\delta \to 0$ é 0. Com isso e o corolário anterior em mente, o próximo corolário garante que podemos nos concentrar em provar a convergência de $\max_{k\in K^*_s} \sqrt{s} S_k$.

Corolário 7.20. *Para $\delta, s > 0$ quaisquer, temos*

$$\max_{k\in K^*_s} p_k \leq \max_{k\in K^*_s} \sqrt{s} S_{k-1} = S_{K^*_s - 1}.$$

Demonstração. Basta aplicar o corolário 7.19 fazendo $K = K^*_s$. ■

Agora, elencamos uma nova definição e dois resultados auxiliares para depois desenvolver a desigualdade principal desta demonstração. É no primeiro desses resultados que usamos a hipótese de que f tem um mínimo global y^*.

Definição. *Seja $f^* = f(x^*)$ o valor de f em seu ponto ótimo. Definimos*

$$C_1 = \sup_{\substack{s>0 \\ k\in[K^*_s]}} \sqrt{2L(f(y_k) - f^*)}.$$

No Apêndice B,Subseção B.2.1, provamos que $C_1 < +\infty$ para $0 < s < \varsigma$. Intuitivamente, podemos esperar isso pois o passo do gradiente é o mesmo que leva o MD a convergir e o passo da inércia tem um fator $\mu_k < 1$, que "dissipa a energia" do método, impedindo que ele divirja.

Lema 7.21. *Sejam $f : \mathbb{R}^n \to \mathbb{R}$ uma função convexa e L-Lipschitz e $x \in \mathbb{R}^n$ quaisquer. Então*

$$\|\nabla f(x)\| \leq \sqrt{2L(f(x) - f^*)}.$$

A demonstração deste lema se encontra no final de Apêndice B,Subseção B.2.2. Juntando o lema com a definição de C_1 e o fato deste ser limitado, temos que para $s < \varsigma$, $\|\nabla f(y_k\| \leq C_1$. Usamos isso no lema a seguir.

Lema 7.22. *Sejam $0 < s < \varsigma$ e $k \in [K^*_s]$ quaisquer. Então*

$$\|z_k\| \leq \|z_{k-1}\| + C_1\sqrt{s}.$$

Demonstração. Basta ver que para qualquer $s > 0$ e qualquer $k \in [K_s^*]$,

$$\|z_k\| = \left\| \frac{k-1}{k+2} z_{k-1} + \sqrt{s}\nabla f(y_{k-1}) \right\| \leq \frac{k-1}{k+2} \|z_{k-1}\| + \sqrt{s} \|\nabla f(y_{k-1})\| \leq \|z_{k-1}\| + C_1\sqrt{s}. \qquad \blacksquare$$

Corolário 7.23. *Sejam $0 < s < \varsigma$ e $k \in [K_s^*]$ quaisquer. Então*

$$\|z_k\| \leq C_1 k\sqrt{s}.$$

Demonstração. Esta demonstração se dá por indução fraca em k. Para o caso base ($k = 0$), temos que $\|z_0\| = \|\vec{0}\| = 0 \leq C_1 0\sqrt{s}$ por definição.

Para o passo indutivo, supondo que vale que $\|z_k\| \leq C_1 k\sqrt{s}$, temos

$$\|z_{k+1}\| \leq \|z_k\| + C_1\sqrt{s} = C_1 k\sqrt{s} + C_1\sqrt{s} = C_1(k+1)\sqrt{s}. \qquad \blacksquare$$

O próximo lema estabelece um limite superior para os termos da sequência $\{v_k\}_{k\in\mathbb{N}}$, o que vai nos levar a um limite superior para termos da série $\{S_k\}_{k\in\mathbb{N}}$ e consequentemente para termos da sequência $\{p_k\}_{k\in\mathbb{N}}$.

A partir daqui vamos passar a usar a hipótese de que $s < 1$; podemos fazer isso porque vamos nos interessar pelos próximos resultados apenas no limite com $s \to 0$. Também vamos usar a hipótese de que $\delta > \frac{3}{2}\sqrt{s} \Leftrightarrow s < \frac{4}{9}\delta^2$. Também podemos fazer isso porque vamos tomar o limite com $s \to 0$ antes de tomar o limite com $\delta \to 0$; a razão disso ficará mais clara em breve, ao final desta seção. Tomamos essas hipóteses para assegurar que $\max(\delta, k\sqrt{s}) = \delta$ quando $k = 1$, o que vai permitir uma simplificação no próximo fato. Com essas novas hipóteses e a anterior, escreveremos $s < \min(\varsigma, 1, \frac{4}{9}\delta^2)$

Fato 7.24. *Para $\delta > 0$, $0 < s < \min(\varsigma, 1, \frac{4}{9}\delta^2)$ e $k \in [K_s^*]$ quaisquer, vale que*

$$v_{k+1} \leq v_k + \frac{L\sqrt{s}}{\nu} p_k + \left(\frac{3}{k+2} - \frac{3\sqrt{s}}{\max(\delta, k\sqrt{s})} + \frac{Ls}{\nu} \right) \|z_k\|.$$

Demonstração. Sejam $\delta > 0$, $0 < s < \min(\varsigma, 1, \frac{4}{9}\delta^2)$ e $k \in \mathbb{N}$ quaisquer. Sejam também $M_{\delta,s,k} = \left(1 - \frac{3\sqrt{s}}{\max(\delta, k\sqrt{s})}\right)$ e $\mu_k = \frac{k-1}{k+2}$.

Para $\delta > k\sqrt{s}$ (o que inclui o caso em que $k = 1$), temos $M_{\delta,s,k} = 1 - \frac{3\sqrt{s}}{\delta} \geq 1 - 2 = -1$. Nesse caso, temos também que $M_{\delta,s,k} \leq 0$.

Para $\delta \leq k\sqrt{s}$ e $k > 1$, temos $M_{\delta,s,k} = 1 - \frac{3\sqrt{s}}{k\sqrt{s}} = 1 - \frac{3}{k} \in [-1/2, 1)$.

Dessa maneira, para todo $k \in [K_s^*]$, temos $|M_{\delta,s,k}| \leq 1$. Além disso, para todo $k \in [K_s^*]$ também temos $|\mu_k| \leq 1$.

Notemos também que

$$\begin{aligned} M_{\delta,s,k} - \mu_k &= 1 - \frac{3\sqrt{s}}{\max(\delta, k\sqrt{s})} - \frac{k-1}{k+2} = \frac{k+2}{k+2} - \frac{k-1}{k+2} - \frac{3\sqrt{s}}{\max(\delta, k\sqrt{s})} \\ &= \frac{k+2-k+1}{k+2} - \frac{3\sqrt{s}}{\max(\delta, k\sqrt{s})} = \frac{3}{k+2} - \frac{3\sqrt{s}}{\max(\delta, k\sqrt{s})}. \end{aligned}$$

Além disso, com a introdução dessas variáveis, podemos escrever

$$Z_{k+1}^\delta = M_{\delta,s,k} Z_k^\delta - \sqrt{s}\nabla f(X_k^\delta) \quad , \quad z_{k+1} = \mu_k z_k - \sqrt{s}\nabla f(y_k) \quad \text{e} \quad y_k = (\nabla\Phi)^{-1}(\theta + \mu_k\sqrt{s} z_{k-1})$$

de forma que

$$\begin{aligned}
v_{k+1} &= \left\|Z_{k+1}^{\delta} - z_{k+1}\right\| \\
&= \left\|M_{\delta,s,k}Z_k^{\delta} - \sqrt{s}\nabla f(X_k^{\delta}) - \mu_k z_k + \sqrt{s}\nabla f(y_k))\right\| \\
&\leq \left\|M_{\delta,s,k}Z_k^{\delta} - \mu_k z_k\right\| + \sqrt{s}\left\|\nabla f(y_k) - \nabla f(X_k^{\delta})\right\| \\
&= \left\|M_{\delta,s,k}Z_k^{\delta} - M_{\delta,s,k}z_k + M_{\delta,s,k}z_k - \mu_k z_k\right\| + \sqrt{s}L\left\|y_k - X_k^{\delta}\right\| \\
&\leq \left\|M_{\delta,s,k}Z_k^{\delta} - M_{\delta,s,k}z_k\right\| + \|(M_{\delta,s,k} - \mu_k)z_k\| + \frac{L\sqrt{s}}{\nu}\left\|\theta_k + \mu_k\sqrt{s}z_{k-1} - \Theta(X_k^{\delta})\right\| \\
&\leq |M_{\delta,s,k}|v_k + |M_{\delta,s,k} - \mu_k|\,\|z_k\| + \frac{L\sqrt{s}}{\nu}\left\|\theta_k - \Theta(X_k^{\delta})\right\| + \frac{\mu_k Ls}{\nu}\|z_k\| \\
&\leq v_k + |M_{\delta,s,k} - \mu_k|\,\|z_k\| + \frac{L\sqrt{s}}{\nu}p_k + \frac{Ls}{\nu}\|z_k\| \\
&= v_k + \frac{L\sqrt{s}}{\nu}p_k + \left(|M_{\delta,s,k} - \mu_k| + \frac{Ls}{\nu}\right)\|z_k\| \\
&= v_k + \frac{L\sqrt{s}}{\nu}p_k + \left(\left|\frac{3}{k+2} - \frac{3\sqrt{s}}{\max(\delta, k\sqrt{s})}\right| + \frac{Ls}{\nu}\right)\|z_k\|\,. \qquad \blacksquare
\end{aligned}$$

O próximo fato fornece uma análise individual do último termo, simplificando-o.

Fato 7.25. *Sejam $\delta > 0$, $0 < s < \min(\varsigma, 1, \frac{4}{9}\delta^2)$ e $k \in [K_s^*]$ quaisquer. Então*

$$\left(\left|\frac{3}{k+2} - \frac{3\sqrt{s}}{\max(\delta, k\sqrt{s})}\right| + Ls\right)\|z_k\| \leq \begin{cases} C_2\sqrt{s}, & \text{se } \delta \geq k\sqrt{s}, \\ \frac{C_2\sqrt{s}}{k}, & \text{se } \delta < k\sqrt{s}. \end{cases}$$

Demonstração. Sejam $\delta > 0$, $0 < s < \min(\varsigma, 1, \frac{4}{9}\delta^2)$ e $k \in [K_s^*]$ quaisquer. Nessas circunstâncias, $k\sqrt{s} \leq T$.

Se $\delta \geq k\sqrt{s}$, temos que $\max(\delta, k\sqrt{s}) = \delta$ e $k \leq \frac{\delta}{\sqrt{s}} \Longrightarrow \frac{1}{k} \geq \frac{\sqrt{s}}{\delta}$. Logo existe um $C_2' > 0$ tal que

$$\begin{aligned}
\left|\frac{3}{k+2} - \frac{3\sqrt{s}}{\max(\delta, k\sqrt{s})}\right| &= \left|\frac{3}{k+2} - \frac{3\sqrt{s}}{\delta}\right| \leq \left|\frac{3}{k+2}\right| + \left|\frac{3\sqrt{s}}{\delta}\right| \\
&\leq \left|\frac{3}{k+2}\right| + \left|\frac{3}{k}\right| \leq \frac{3}{k+2} + \frac{3}{k} \leq C_2'\frac{1}{k}.
\end{aligned}$$

Dessa forma, existe um outro $C_2''' > 0$ de modo que

$$\begin{aligned}
\left(\left|\frac{3}{k+2} - \frac{3\sqrt{s}}{\max(\delta, k\sqrt{s})}\right| + Ls\right)\|z_k\| &\leq (C_2'\sqrt{s} + Ls)\,\|z_k\| \leq (C_2'\sqrt{s} + Ls)C_1 k\sqrt{s} \\
&= C_2'\sqrt{s}C_1 k\sqrt{s} + LsC_1 k\sqrt{s} \leq C_2'C_1 T\sqrt{s} + LC_1 T\sqrt{s} = C_2'''\sqrt{s}.
\end{aligned}$$

Agora se $\delta < k\sqrt{s}$, temos que $\max(\delta, k\sqrt{s}) = k\sqrt{s}$ e aí existe um certo $C_2'' > 0$ tal que

$$\begin{aligned}
\left|\frac{3}{k+2} - \frac{3\sqrt{s}}{\max(\delta, k\sqrt{s})}\right| &= \left|\frac{3}{k+2} - \frac{3\sqrt{s}}{k\sqrt{s}}\right| = \left|\frac{3}{k+2} - \frac{3}{k}\right| \\
&= \left|\frac{3k - 3(k+2)}{k(k+2)}\right| = \left|\frac{-6}{k(k+2)}\right| = \frac{6}{k(k+2)} \leq \frac{C_2''}{k^2}.
\end{aligned}$$

Assim, existe um $C_2'''' > 0$ tal que

$$\begin{aligned}
\left(\left|\frac{3}{k+2} - \frac{3\sqrt{s}}{\max(\delta, k\sqrt{s})}\right| + Ls\right)\|z_k\| &\leq \left(\frac{C_2''}{k^2} + Ls\right)\|z_k\| \leq \left(\frac{C_2''}{k^2} + L\sqrt{s}\sqrt{s}\right)C_1 k\sqrt{s} \\
&= \left(\frac{C_2''}{k^2}C_1 k + LC_1 sk\right)\sqrt{s} = \left(C_2''C_1 + LC_1\sqrt{s}k\sqrt{s}k\right)\frac{\sqrt{s}}{k} \\
&\leq (C_2''C_1 + LC_1 TT)\,\frac{\sqrt{s}}{k} \leq C_2''''\frac{\sqrt{s}}{k}.
\end{aligned}$$

Fazendo $C_2 = \max(C_2''', C_2'''')$, temos enfim que

$$\left(\left|\frac{3}{k+2} - \frac{3\sqrt{s}}{\max(\delta, k\sqrt{s})}\right| + Ls\right)\|z_k\| \leq \begin{cases} C_2\sqrt{s}, & \text{se } \delta \geq k\sqrt{s}, \\ \frac{C_2\sqrt{s}}{k}, & \text{se } \delta < k\sqrt{s}. \end{cases}$$ ■

O corolário a seguir é central para a demonstração da convergência. Ele fornece uma desigualdade para v_k. Mais adiante, vamos dividir a prova nos dois casos da desigualdade e analisar cada um individualmente.

Corolário 7.26. *Sejam $\delta > 0$, $0 < s < \min(\varsigma, 1, \frac{4}{9}\delta^2)$ e $k \in [K_s^*]$ quaisquer. Temos*

$$v_{k+1} \leq \begin{cases} v_k + LsS_{k-1} + C_2\sqrt{s}, & \text{se } \delta \geq k\sqrt{s}, \\ v_k + LsS_{k-1} + \frac{C_2\sqrt{s}}{k}, & \text{se } \delta < k\sqrt{s} \end{cases}.$$

Demonstração. A prova se dá substituindo as fórmulas dadas pelos fatos 7.18 e 7.25 na expressão dada pelo fato 7.24. ■

Tendo esses limites superiores para v_k, podemos atacar os termos S_k, procurando um limite superior para cada um deles. Para os próximos resultados, em nome da legibilidade, definimos

$$a_s = \frac{C_1Ls + C_2}{2\sqrt{L}} \quad \text{e} \quad b_s = \frac{C_1 + C_2}{2}\sqrt{s}.$$

Definimos também

$$A_s = a_s + b_s \quad \text{e} \quad B_s = a_s - b_s.$$

Por fim definimos

$$c_s = 1 + \sqrt{Ls} \quad \text{e} \quad d_s = 1 - \sqrt{Ls}.$$

De maneira imediata, temos

$$c_s + d_s = 2 \quad , \quad c_s - d_s = 2\sqrt{Ls} \quad \text{e} \quad c_s d_s = 1 - Ls.$$

Usaremos esses resultados diversas vezes. Elencamos a seguir mais dois fatos auxiliares não tão imediatos. As demonstrações desses fatos são puramente algébricas e serão omitidas nesse capítulo, mas podem ser encontradas no Apêndice B, na Subseção B.2.3.

Fato 7.27. *Para $s > 0$ e $k \in \mathbb{N}$ quaisquer, temos*

$$A_s c_s^k + B_s d_s^k = \frac{C_1\sqrt{s}}{2}(c_s^{k+1} - d_s^{k+1}) + \frac{C_2}{2\sqrt{L}}(c_s^{k+1} + d_s^{k+1}).$$

Fato 7.28. *Para qualquer $s > 0$, vale que*

$$\frac{A_s}{c_s} + \frac{B_s}{d_s} = \frac{C_2}{\sqrt{L}}.$$

A partir daqui vamos analisar o comportamento de $\sqrt{Ls}S_k$ em dois casos: quando $k \leq k_{\delta,s}^*$ e quando $k > k_{\delta,s}^*$. A ideia é que no primeiro caso, o método de Euler discretiza uma EDO diferente da EDO que o AMD discretiza (pelo menos supostamente, já que é o que estamos provando), de forma que seja natural que o máximo das diferenças entre as trajetórias não tenda a 0 necessariamente, mesmo partindo de um mesmo ponto. No segundo caso, as EDOs discretizadas pelos dois métodos são a mesma, mas eles partem de pontos diferentes: cada um parte do ponto final da trajetória do método até a iteração $k_{\delta,s}^*$.

No entanto, como no segundo caso a EDO é a mesma e é contínua e Lipschitz, pela dependência contínua das condições iniciais, esperaríamos que o máximo das diferenças entre as trajetórias dependesse continuamente da distância das condições iniciais $X^{\delta}_{k^*_{\delta,s}}$ e $x_{k^*_{\delta,s}}$. Ao mesmo tempo, pela convergência uniforme das soluções suavizadas (pensando que as EDOs suavizadas "tendem" à EDO original quando $\delta \to 0$), essas mesmas condições deveriam se aproximar de 0 quando $\delta \to 0$. Aqui se justifica por que o limite com $s \to 0$ deve ser tomado antes do limite com $\delta \to 0$.

7.2.1 Caso 1: $k \leq k^*_{\delta,s}$ ou onde há suavização

Fato 7.29. *Sejam $\delta > 0$, $0 < s < \min(\varsigma, 1, \frac{4}{9}\delta^2)$ e $k \in [k^*_{\delta,s}]$ quaisquer. Então*

$$\sqrt{Ls}S_{k-1} \leq A_s c_s^{k-1} + B_s d_s^{k-1} - \frac{C_2}{\sqrt{L}} \qquad e \qquad v_k \leq A_s c_s^{k-1} - B_s d_s^{k-1}.$$

Demonstração. A demonstração se dá por indução forte em k. Nos atentemos de que estamos provando duas proposições simultaneamente por indução. A hipótese de indução consiste em supor que apenas a segunda proposição é válida para todo $i \leq k$. Inicialmente, essa hipótese será usada para demonstrar a primeira proposição no caso $i = k+1$. Em seguida, empregamos novamente a hipótese de indução, agora em conjunto com o caso já estabelecido da primeira proposição para $i = k+1$, para concluir a segunda proposição nesse mesmo índice. Em certo sentido, as duas proposições funcionam como auxiliares uma da outra no processo indutivo.

Esta demonstração se dá por indução forte em k. Atentemo-nos de que estamos provando duas proposições de uma vez por indução. A hipótese de indução será de que vale apenas a segunda para $i \leq k \in \mathbb{N}$. A hipótese será usada primeiro para provar a primeira proposição no caso $i = k+1$. Depois, será usada novamente, junto do caso recém provado da primeira preposição com $i = k+1$, para provar a segunda no caso $i = k+1$. Em um certo sentido, as duas proposições são auxiliares uma da outra.

Para o caso base ($k = 1$) temos

$$\begin{aligned}\sqrt{Ls}S_0 = \sqrt{Ls}v_0 = 0 \leq C_1\sqrt{L}s = C_1\sqrt{L}s + \frac{C_2}{\sqrt{L}} - \frac{C_2}{\sqrt{L}} \\ = 2\frac{C_1 Ls + C_2}{2\sqrt{L}} - \frac{C_2}{\sqrt{L}} = 2a - \frac{C_2}{\sqrt{L}} = A_s + B_s - \frac{C_2}{\sqrt{L}} = A_s c_s^0 + B_s d_s^0 - \frac{C_2}{\sqrt{L}}.\end{aligned}$$

e

$$v_1 \leq \frac{C_1 + C_2}{2}\sqrt{s} = b_s = A_s - B_s = A_s c_s^0 - B_s d_s^0.$$

Para o passo indutivo, suponhamos que para todo $i \in [k]$ vale que $v_i \leq A_s c_s^{i-1} - B_s d_s^{i-1}$; esta é a hipótese da indução. Usando a fórmula para somas finitas de progressões geométricas e o fato 7.28, temos então que

$$\begin{aligned}\sqrt{Ls}S_{k-1} &= \sqrt{Ls}\sum_{i=0}^{k-1} v_i \leq \sqrt{Ls}\sum_{i=0}^{k-1}\left[A_s c_s^{i-1} - B_s d_s^{i-1}\right] \\ &= \sqrt{Ls}A_s\sum_{i=0}^{k-1} c_s^{i-1} - \sqrt{Ls}B_s\sum_{i=0}^{k-1} d_s^{i-1} = \sqrt{Ls}\frac{A_s}{c_s}\sum_{i=0}^{k-1} c_s^i - \sqrt{Ls}B_s\frac{1}{d_s}\sum_{i=0}^{k-1} d_s^i \\ &= \sqrt{Ls}\frac{A_s}{c_s}\left(\frac{1-c_s^k}{1-c_s}\right) - \sqrt{Ls}\frac{B_s}{d_s}\left(\frac{1-d_s^k}{1-d_s}\right) \\ &= \sqrt{Ls}\frac{A_s}{c_s}\left(\frac{1-c_s^k}{-\sqrt{Ls}}\right) - \sqrt{Ls}\frac{B_s}{d_s}\left(\frac{1-d_s^k}{\sqrt{Ls}}\right) \\ &= -\frac{A_s}{c_s}(1-c_s^k) - \frac{B_s}{d_s}(1-d_s^k) = -\frac{A_s}{c_s} + A_s c_s^{k-1} - \frac{B_s}{d_s} + B_s d_s^{k-1} \\ &= A_s c_s^{k-1} + B_s d_s^{k-1} - \left(\frac{A_s}{c_s} + \frac{B_s}{d_s}\right) = A_s c_s^{k-1} + B_s d_s^{k-1} - \frac{C_2}{\sqrt{L}}.\end{aligned}$$

Lembremo-nos também de que $v_{k+1} \leq v_k + LsS_{k-1} + C_2\sqrt{s}$, já que $k \leq k^*_{\delta,s}$. Desta maneira, usando a hipótese da indução e a equação que acabamos de derivar, obtemos

$$\begin{aligned}v_{k+1} &\leq v_k + LsS_{k-1} + C_2\sqrt{s} \\ &\leq A_s c_s^{k-1} - B_s d_s^{k-1} + \sqrt{Ls}A_s c_s^{k-1} + \sqrt{Ls}B_s d_s^{k-1} - C_2\sqrt{s} + C_2\sqrt{s} \\ &= A_s c_s^{k-1}(1 + \sqrt{Ls}) - B_s d_s^{k-1}(1 - \sqrt{Ls}) \\ &= A_s c_s^{k-1} c_s - B_s d_s^{k-1} d_s = A_s c_s^k - B_s d_s^k.\end{aligned}$$

■

Fato 7.30. *Sejam $\delta > 0$, $0 < s < \min(\varsigma, 1, \frac{4}{9}\delta^2)$ e $k \in [k^*_{\delta,s}]$ quaisquer. Então*

$$S_{k-1} \leq \frac{C_1}{2\sqrt{L}}(c_s^k - d_s^k) + \frac{C_2}{2L\sqrt{s}}(c_s^k + d_s^k) - \frac{C_2}{L\sqrt{s}}.$$

Demonstração. Usando os fatos 7.29 e 7.27, temos

$$\begin{aligned} S_k = \frac{\sqrt{Ls}S_k}{\sqrt{Ls}} &\leq \frac{A_s c_s^k + B_s d_s^k}{\sqrt{Ls}} - \frac{C_2}{L\sqrt{s}} \\ &= \frac{C_1\sqrt{s}}{2\sqrt{Ls}}(c_s^{k+1} - d_s^{k+1}) + \frac{C_2}{2\sqrt{L}\sqrt{Ls}}(c_s^{k+1} + d_s^{k+1}) - \frac{C_2}{L\sqrt{s}} \\ &= \frac{C_1}{2\sqrt{L}}(c_s^{k+1} - d_s^{k+1}) + \frac{C_2}{2L\sqrt{s}}(c_s^{k+1} + d_s^{k+1}) - \frac{C_2}{L\sqrt{s}}. \end{aligned}$$

■

A partir de agora, para resultados relacionados a c_s e d_s — especialmente d_s e suas potências —, vamos passar a supor que $s \leq 1/L$ de forma que $1 - \sqrt{Ls} \geq 0$ e vamos escrever $s < \min(\varsigma, 1, \frac{4}{9}\delta^2, 1/L)$ (exceto no fato logo a seguir, quando só precisamos de $s < \min(1, 1/L)$); de novo, podemos fazer isso pois estamos interessados nesses resultados no limite com $s \to 0$.

Afirmação 7.31. *Para $0 < s < \min(1, 1/L)$, as funções f e g definidas por $f(t) = c_s^t - d_s^t$ e $g(t) = c_s^t + d_s^t$, com $c_s = 1 + \sqrt{Ls}$ e $d_s = 1 - \sqrt{Ls}$, são crescentes em t.*

Usaremos essa afirmação para majorar, por exemplo, $c_s^{k^*_{\delta,s}} - d_s^{k^*_{\delta,s}}$ por $c_s^{\delta/\sqrt{s}} - d_s^{\delta/\sqrt{s}}$ no próximo fato e n'outro mais adiante. A sua demonstração se encontra no Apêndice B, Subseção B.2.4.

Fato 7.32. *Para $\delta > 0$, $0 < s < \min(\varsigma, 1, \frac{4}{9}\delta^2, 1/L)$ e $k \in [k^*_{\delta,s}]$ quaisquer, temos*

$$\sqrt{s}S_{k^*_{\delta,s}-1} \leq \frac{C_1}{2\sqrt{L}}\left(c_s^{\delta/\sqrt{s}} - d_s^{\delta/\sqrt{s}}\right)\sqrt{s} + \frac{C_2}{2L}\left(c_s^{\delta/\sqrt{s}} + d_s^{\delta/\sqrt{s}}\right) - \frac{C_2}{L}.$$

Demonstração. Usando o fato 7.30 e a afirmação 7.31,

$$\begin{aligned} \sqrt{s}S_{k^*_{\delta,s}-1} &\leq \sqrt{s}\left(\frac{C_1}{2\sqrt{L}}\left(c_s^{k^*_{\delta,s}} - d_s^{k^*_{\delta,s}}\right) + \frac{C_2}{2L\sqrt{s}}\left(c_s^{k^*_{\delta,s}} + d_s^{k^*_{\delta,s}}\right) - \frac{C_2}{L\sqrt{s}}\right) \\ &\leq \sqrt{s}\left(\frac{C_1}{2\sqrt{L}}\left(c_s^{\delta/\sqrt{s}} - d_s^{\delta/\sqrt{s}}\right) + \frac{C_2}{2L\sqrt{s}}\left(c_s^{\delta/\sqrt{s}} + d_s^{\delta/\sqrt{s}}\right) - \frac{C_2}{L\sqrt{s}}\right) \\ &= \frac{C_1}{2\sqrt{L}}\left(c_s^{\delta/\sqrt{s}} - d_s^{\delta/\sqrt{s}}\right)\sqrt{s} + \frac{C_2}{2L}\left(c_s^{\delta/\sqrt{s}} + d_s^{\delta/\sqrt{s}}\right) - \frac{C_2}{L}. \end{aligned}$$

■

Este fato fornece uma majoração de $\sqrt{s}S_{k^*_{\delta,s}-1}$ em função apenas de δ e s. Usaremos isso para majorar a diferença máxima das condições $X^\delta_{k^*_{\delta,s}}$ e $x_{k^*_{\delta,s}}$.

7.2.2 Caso 2: $k > k^*_{\delta,s}$ ou onde não há suavização

Fato 7.33. *Para $\delta > 0$, $0 < s < \min(\varsigma, 1, \frac{4}{9}\delta^2, 1/L)$ e $k \in [K^*_s]$ quaisquer, temos*

$$\sqrt{s}S_{K^*_\delta} \leq \sqrt{s}\left(\frac{S_{k^*_{\delta,s}}}{2} + \frac{C_2}{2\delta L}\right)\left(c_s^{K^*_s - k^*_{\delta,s}} + d_s^{K^*_s - k^*_{\delta,s}}\right) + \frac{v_{k^*_{\delta,s}+1}}{2\sqrt{L}}\left(c_s^{K^*_s - k^*_{\delta,s}} - d_s^{K^*_s - k^*_{\delta,s}}\right) - \frac{C_2\sqrt{s}}{\delta L}.$$

Demonstração. Sejam $\delta > 0$, $0 < s < \min(\varsigma, 1, \frac{4}{9}\delta^2, 1/L)$. Seja $k \geq k^*_{\delta,s}$. Temos então que $k \geq \delta/\sqrt{s} \implies 1/k \leq \sqrt{s}/\delta \implies \sqrt{s}/k \leq s/\delta$ e aí, usando o corolário 7.26,

$$v_k \leq v_{k-1} + LsS_{k-2} + C_2\sqrt{s}/k \leq v_{k-1} + LsS_{k-2} + C_2 s/\delta. \tag{7.10}$$

Consideremos as variáveis auxiliares w_k e R_k definidas para $k \geq k^*_{\delta,s} + 1$ pela recorrência

$$\begin{cases} w_k &= w_{k-1} + LsR_{k-2} + C_2 s/\delta \\ R_k &= R_{k-1} + w_k \end{cases}$$

com condições iniciais $w_{k^*_{\delta,s}+1} = v_{k^*_{\delta,s}+1}$ e $R_{k^*_{\delta,s}} = S_{k^*_{\delta,s}}$. Isto é, w_k e R_k representam o que consideramos o "pior caso" para v_k e S_k respectivamente (quando vale a igualdade para a Equação 7.10).

Por indução podemos provar que para $k \geq k^*_{\delta,s} + 1$, vale que

$$w_k \geq v_k \quad \text{e} \quad R_{k-1} \geq S_{k-1}.$$

O caso base da indução ($i = k^*_{\delta,s} + 1$) se dá pela própria definição das condições iniciais. supondo agora que $w_i \geq v_i$ e $R_{i-1} \geq S_{i-1}$ para algum $i \geq k^*_{\delta,s} + 1$, temos

$$w_{i+1} = w_i + LsR_{i-1} + C_2 s/\delta \geq v_i + LsS_{i-1} + C_2 s\delta \geq v_i$$

e, consequentemente,

$$R_i = R_{i-1} + w_i \geq S_{i-1} + v_i = S_i,$$

o que configura o passo indutivo.

Usando $R_k = R_{k-1} + w_k$ e $R_{k-1} = R_{k-2} + w_{k-1}$, para $k \geq k^*_{\delta,s} + 1$ podemos escrever

$$\begin{aligned} R_k - R_{k-1} &= R_{k-1} - R_{k-2} + LsR_{k-2} + C_2 s/\delta \\ &= R_{k-1} - (1 - Ls)R_{k-2} + C_2 s/\delta. \end{aligned}$$

Por fim, rearranjando os termos, poderíamos escrever

$$R_k = 2R_{k-1} - (1 - Ls)R_{k-2} + C_2 s/\delta \quad \text{ou} \quad R_k = 2R_{k-1} - c_s d_s R_{k-2} + C_2 s/\delta.$$

Esta equação junto das condições iniciais $R_{k^*_{\delta,s}} = S_{k^*_{\delta,s}}$ e $R_{k^*_{\delta,s}+1} = R_{k^*_{\delta,s}} + w_{k^*_{\delta,s}+1} = R_{k^*_{\delta,s}} + v_{k^*_{\delta,s}+1}$ caracteriza uma definição alternativa para a sequência $\{R_k\}$. Com essa definição, podemos provar que para $k \in [k^*_{\delta,s}, K^*_\delta]$, vale

$$R_k = \left(\frac{S_{k^*_{\delta,s}}}{2} + \frac{C_2}{2\delta L}\right)(c_s^{k-k^*_{\delta,s}} + d_s^{k-k^*_{\delta,s}}) + \frac{v_{k^*_{\delta,s}+1}}{2\sqrt{Ls}}(c_s^{k-k^*_{\delta,s}} - d_s^{k-k^*_{\delta,s}}) - \frac{C_2}{\delta L};$$

essa demonstração se dá por indução fraca e se encontra no Apêndice B, Subseção B.2.5. Assim

$$\sqrt{s}S_{K^*_s} \leq \sqrt{s}R_{K^*_s} = \sqrt{s}\left(\frac{S_{k^*_{\delta,s}}}{2} + \frac{C_2}{2\delta L}\right)\left(c_s^{K^*_s - k^*_{\delta,s}} + d_s^{K^*_s - k^*_{\delta,s}}\right) + \frac{v_{k^*_{\delta,s}+1}}{2\sqrt{L}}\left(c_s^{K^*_s - k^*_{\delta,s}} - d_s^{K^*_s - k^*_{\delta,s}}\right) - \frac{C_2\sqrt{s}}{\delta L}. \blacksquare$$

7.2.3 Parte final

Tomaremos agora os limites de $\sqrt{s}S_{K^*_s}$ com $s \to 0^+$ e $\delta \to 0^+$.

Para tomar o limite em s, consideremos antes os limites

i) $\lim_{s\to 0^+} a_s = \frac{C_2}{2\sqrt{L}}$;

ii) $\lim_{s\to 0^+} b_s = 0$;

iii) $\lim_{s\to 0^+}\left(c_s^{k^*_{\delta,s}-1} + d_s^{k^*_{\delta,s}-1}\right) = e^{\sqrt{L}\delta} + e^{-\sqrt{L}\delta}$; e

iv) $\lim_{s\to 0^+}\left(c_s^{k^*_{\delta,s}-1} - d_s^{k^*_{\delta,s}-1}\right) = e^{\sqrt{L}\delta} - e^{-\sqrt{L}\delta}$.

Os desenvolvimentos desses limites se encontram no Apêndice B, Subseção B.2.6. Notemos que para $\delta > 0$ fixo, cada um desses limites é finito. Além disso, pela segunda desigualdade do fato 7.29, com $k = k^*_{\delta,s}$, temos

$$\begin{aligned}
\lim_{s\to 0^+} v_{k^*_{\delta,s}+1} &\leq \lim_{s\to 0^+} \left(A_s c_s^{k^*_{\delta,s}} - B_s d_s^{k^*_{\delta,s}}\right) \\
&= \lim_{s\to 0^+} \left((a_s+b_s)c_s^{k^*_{\delta,s}} - (a_s-b_s)d_s^{k^*_{\delta,s}}\right) \\
&= \lim_{s\to 0^+} \left(a_s\left(c_s^{k^*_{\delta,s}} - d_s^{k^*_{\delta,s}}\right) + b_s\left(c_s^{k^*_{\delta,s}} + d_s^{k^*_{\delta,s}}\right)\right) \\
&= \lim_{s\to 0^+} a_s \lim_{s\to 0^+}\left(c_s^{k^*_{\delta,s}} - d_s^{k^*_{\delta,s}}\right) + \lim_{s\to 0^+} b_s \lim_{s\to 0^+}\left(c_s^{k^*_{\delta,s}} + d_s^{k^*_{\delta,s}}\right) = \frac{C_2}{2\sqrt{L}}\left(e^{\sqrt{L}\delta} - e^{-\sqrt{L}\delta}\right).
\end{aligned}$$

Consideremos ainda os limites

v) $\lim_{s\to 0^+}\left(c_s^{\delta/\sqrt{s}+1} + d_s^{\delta/\sqrt{s}+1}\right) = e^{\sqrt{L}\delta} + e^{-\sqrt{L}\delta}$;

vi) $\lim_{s\to 0^+}\left(c_s^{\delta/\sqrt{s}+1} - d_s^{\delta/\sqrt{s}+1}\right) = e^{\sqrt{L}\delta} - e^{-\sqrt{L}\delta}$;

vii) $\lim_{s\to 0^+}\left(c_s^{K^*_s - k^*_{\delta,s}} + d_s^{K^*_s - k^*_{\delta,s}}\right) \leq e^{(T-\delta)\sqrt{L}} + e^{-(T-\delta)\sqrt{L}}$; e

viii) $\lim_{s\to 0^+}\left(c_s^{K^*_s - k^*_{\delta,s}} - d_s^{K^*_s - k^*_{\delta,s}}\right) \leq e^{(T-\delta)\sqrt{L}} - e^{-(T-\delta)\sqrt{L}}$.

Os desenvolvimentos desses limites também se encontram no Apêndice B, Subseção B.2.6. Para cada $\delta > 0$ fixo, esses limites também são finitos.

O próximo fato apresenta o limite com $s \to 0^+$ de $\sqrt{s}S_{k^*_{\delta,s}-1}$. Vamos precisar desse limite pois esse termo aparece na fórmula que usaremos para tomar limite de $\sqrt{s}S_{K^*_s}$.

Fato 7.34. *Para $\delta > 0$ qualquer, vale*

$$\lim_{s\to 0^+} \sqrt{s}S_{k^*_{\delta,s}-1} \leq \frac{C_2}{2L}\left(e^{\sqrt{L}\delta} + e^{-\sqrt{L}\delta} - 2\right).$$

Demonstração. Seja $\delta > 0$ qualquer. Usando o fato da sequência S_k ser não decrescente, o fato 7.32 e os limites v) e vi), temos

$$\begin{aligned}
\lim_{s\to 0^+} \sqrt{s}S_{k^*_{\delta,s}} &\leq \lim_{s\to 0^+} \frac{C_1}{2\sqrt{L}}(c_s^{\delta/\sqrt{s}+1} - d_s^{\delta/\sqrt{s}+1})\sqrt{s} + \frac{C_2}{2L}(c_s^{\delta/\sqrt{s}+1} + d_s^{\delta/\sqrt{s}+1}) - \frac{C_2}{L} \\
&= \frac{C_1}{2\sqrt{L}} \lim_{s\to 0^+}(c_s^{\delta/\sqrt{s}+1} - d_s^{\delta/\sqrt{s}+1})\sqrt{s} + \frac{C_2}{2L}\lim_{s\to 0^+}(c_s^{\delta/\sqrt{s}+1} + d_s^{\delta/\sqrt{s}+1}) - \frac{C_2}{L} \\
&= 0 + \frac{C_2}{2L}(e^{\sqrt{L}\delta} + e^{-\sqrt{L}\delta}) - \frac{C_2}{L} \\
&= \frac{C_2}{2L}\left(e^{\sqrt{L}\delta} + e^{-\sqrt{L}\delta} - 2\right).
\end{aligned}$$

■

Tomamo agora o limite de $\sqrt{s}S_{K^*_s}$ com $s \to 0^+$.

Fato 7.35. *Para $\delta > 0$ qualquer, vale*

$$\lim_{s\to 0^+} \max_{k\in[K^*_s]} \sqrt{s}S_{k-1} \leq \frac{C_2}{2L}\left(\left(e^{T/\sqrt{L}} - e^{(T-\delta)/\sqrt{L}}\right) + \left(e^{-T/\sqrt{L}} - e^{-(T-\delta)/\sqrt{L}}\right)\right).$$

Demonstração.

$$\begin{aligned}
\lim_{s\to 0^+}\max_{k\in[K_s^*]}\sqrt{s}S_{k-1} &= \lim_{s\to 0^+}\sqrt{s}S_{K_s^*-1} \le \lim_{s\to 0^+}\sqrt{s}S_{K_s^*} \le \lim_{s\to 0^+}\sqrt{s}R_{K_s^*}\\
&= \lim_{s\to 0^+}\left[\sqrt{s}\left(\frac{S_{k^*_{\delta,s}}}{2}+\frac{C_2}{2\delta L}\right)\left(c_s^{K_s^*-k^*_{\delta,s}}+d_s^{K_s^*-k^*_{\delta,s}}\right)+\frac{v_{k^*_{\delta,s}+1}}{2\sqrt{L}}\left(c_s^{K_s^*-k^*_{\delta,s}}-d_s^{K_s^*-k^*_{\delta,s}}\right)-\frac{C_2\sqrt{s}}{\delta L}\right]\\
&= \frac{1}{2}\lim_{s\to 0^+}\sqrt{s}S_{k^*_{\delta,s}}\lim_{s\to 0^+}\left(c_s^{K_s^*-k^*_{\delta,s}}+d_s^{K_s^*-k^*_{\delta,s}}\right)+\frac{C_2}{2\delta L}\lim_{s\to 0^+}\sqrt{s}\left(c_s^{K_s^*-k^*_{\delta,s}}+d_s^{K_s^*-k^*_{\delta,s}}\right)\\
&\qquad+\frac{1}{2\sqrt{L}}\lim_{s\to 0^+}v_{k^*_{\delta,s}+1}\lim_{s\to 0^+}\left(c_s^{K_s^*-k^*_{\delta,s}}-d_s^{K_s^*-k^*_{\delta,s}}\right)-\frac{C_2}{\delta L}\lim_{s\to 0^+}\sqrt{s}\\
&\le \frac{C_2}{4L}\left(e^{\sqrt{L}\delta}+e^{-\sqrt{L}\delta}-2\right)\left(e^{(T-\delta)\sqrt{L}}+e^{-(T-\delta)\sqrt{L}}\right)+\frac{C_2}{4L}\left(e^{\sqrt{L}\delta}-e^{-\sqrt{L}\delta}\right)\left(e^{(T-\delta)\sqrt{L}}-e^{-(T-\delta)\sqrt{L}}\right).
\end{aligned}$$

A partir daí, com mais algum desenvolvimento, podemos concluir enfim que

$$\lim_{s\to 0^+}\max_{k\in[K_s^*]}\sqrt{s}S_{k-1} \le \frac{C_2}{2L}\left(\left(e^{T/\sqrt{L}}-e^{(T-\delta)/\sqrt{L}}\right)+\left(e^{-T/\sqrt{L}}-e^{-(T-\delta)/\sqrt{L}}\right)\right). \qquad \blacksquare$$

Agora, para tomar o limite em δ, consideremos os limites

ix) $\lim_{\delta\to 0^+}\left(e^{-T/\sqrt{L}}-e^{-(T-\delta)/\sqrt{L}}\right)=e^{-T/\sqrt{L}}-e^{-(T)/\sqrt{L}}=0$; e

x) $\lim_{\delta\to 0^+}\left(e^{T/\sqrt{L}}-e^{(T-\delta)/\sqrt{L}}\right)=e^{T/\sqrt{L}}-e^{T/\sqrt{L}}=0.$

Esses limites são obtidos pela simples substituição de δ por 0, já que a exponencial é contínua. Com isso, podemos provar o próximo resultado.

Lema 7.36.

$$\lim_{\delta\to 0^+}\lim_{s\to 0^+}\max_{k\in[K_s^*]}p_k=0.$$

Demonstração.

$$\begin{aligned}
\lim_{\delta\to 0^+}\lim_{s\to 0^+}\max_{k\in[K_s^*]}p_k &\le \lim_{\delta\to 0^+}\lim_{s\to 0^+}\max_{k\in[K_s^*]}\sqrt{s}S_{k-1}\\
&\le \lim_{\delta\to 0^+}\left(\frac{C_2}{2L}\left(\left(e^{T/\sqrt{L}}-e^{(T-\delta)/\sqrt{L}}\right)+\left(e^{-T/\sqrt{L}}-e^{-(T-\delta)/\sqrt{L}}\right)\right)\right)\\
&= \frac{C_2}{2L}\left(\lim_{\delta\to 0^+}\left(e^{T/\sqrt{L}}-e^{(T-\delta)/\sqrt{L}}\right)+\lim_{\delta\to 0^+}\left(e^{-T/\sqrt{L}}-e^{-(T-\delta)/\sqrt{L}}\right)\right)=0. \qquad \blacksquare
\end{aligned}$$

Isso prova a terceira convergência, como gostaríamos! Podemos agora finalizar a demonstração da convergência do AMD para a solução da EDO (6.1).

Lema 7.37.

$$\lim_{s\to 0}\max_{k\in[K_s^*]}\left\|x_k-X(k\sqrt{s})\right\|=0$$

Demonstração. Tomemos a sequência $\delta_m \searrow 0$ do lema 7.11, que satisfaz $\lim_{m\to\infty}\sup_{t\in[0,T]}\|X_\delta(t)-X(t)\|=0$ ou, de maneira equivalente,

$$\lim_{m\to\infty}\lim_{s\to 0}\max_{k\in[K_s^*]}\left\|X_\delta(k\sqrt{s})-X(k\sqrt{s})\right\|=0.$$

Como nem x_k nem $X(k\sqrt{s})$ dependem de m,

$$\lim_{s\to 0}\max_{k\in[K_s^*]}\left\|x_k-X(k\sqrt{s})\right\|=\lim_{m\to\infty}\lim_{s\to 0}\max_{k\in[K_s^*]}\left\|x_k-X(k\sqrt{s})\right\|.$$

Agora, pela desigualdade triangular, obtemos

$$\begin{aligned}\left\|x_k - X(k\sqrt{s})\right\| &= \left\|x_k - X_k^{\delta_m} + X_k^{\delta_m} - X_{\delta_m}(k\sqrt{s}) + X_{\delta_m}(k\sqrt{s}) - X(k\sqrt{s})\right\| \\ &\leq \left\|x_k - X_k^{\delta_m}\right\| + \left\|X_k^{\delta_m} - X_{\delta_m}(k\sqrt{s})\right\| + \left\|X_{\delta_m}(k\sqrt{s}) - X(k\sqrt{s})\right\| \\ &\leq \frac{1}{\nu}\left\|\theta_k - \Theta(X_k^{\delta_m})\right\| + \left\|X_k^{\delta_m} - X_{\delta_m}(k\sqrt{s})\right\| + \left\|X_{\delta_m}(k\sqrt{s}) - X(k\sqrt{s})\right\| \\ &= \frac{1}{\nu}p_k + \left\|X_k^{\delta_m} - X_{\delta_m}(k\sqrt{s})\right\| + \left\|X_{\delta_m}(k\sqrt{s}) - X(k\sqrt{s})\right\|.\end{aligned}$$

Dessa forma

$$\begin{aligned}\lim_{m\to\infty}\lim_{s\to 0}\max_{k\in[K_s^*]}\left\|x_k - X(k\sqrt{s})\right\| &= \lim_{\delta_m\to 0}\lim_{s\to 0}\max_{k\in[K_s^*]}\left\|x_k - X(k\sqrt{s})\right\| \\ &= \lim_{\delta_m\to 0}\lim_{s\to 0}\max_{k\in[K_s^*]}\left(\frac{1}{\nu}p_k + \left\|X_k^{\delta_m} - X_{\delta_m}(k\sqrt{s})\right\| + \left\|X_{\delta_m}(k\sqrt{s}) - X(k\sqrt{s})\right\|\right) \\ &= \frac{1}{\nu}\lim_{\delta_m\to 0}\lim_{s\to 0}\max_{k\in[K_s^*]} p_k \\ &\quad + \lim_{\delta_m\to 0}\lim_{s\to 0}\max_{k\in[K_s^*]}\left\|X_k^{\delta_m} - X_{\delta_m}(k\sqrt{s})\right\| \\ &\quad + \lim_{\delta_m\to 0}\lim_{s\to 0}\max_{k\in[K_s^*]}\left\|X_{\delta_m}(k\sqrt{s}) - X(k\sqrt{s})\right\|.\end{aligned}$$

O primeiro termo é igual a zero pelo lema 7.36. O segundo é igual a zero pois X_k^{δ} foi construído a partir do método Euler-Forward justamente para aproximar $X_{\delta_m}(k\sqrt{s})$. Por fim, o último termo é igual a zero pois $\{\delta_m\}_{m\in\mathbb{N}}$ foi tomada para isso. Portanto,

$$\lim_{s\to 0}\max_{k\in[K_s^*]}\left\|x_k - X(k\sqrt{s})\right\| = 0. \qquad \blacksquare$$

Capítulo 8

Experimento numérico

Neste capítulo consideramos o problema de mínimos quadrados restrito ao simplexo n-dimensional (como visto em 5.5),

$$\min_{x\in\Delta^{n-1}} f(x) = \|Ax-b\|^2,$$

e comparamos numericamente quatro algoritmos de otimização abordados: Gradient Descent projetado, Accelerated Gradient Descent Projetado, Projected Mirror Descent usando entropia negativa e Projected Accelerated Mirror Descent, também com entropia negativa. O experimento foi feito usando Python ([Pyt23]) e os códigos podem ser encontrados em `https://github.com/josearthursouza/Dissertacao`.

A construção das escolhas de passo η segue diretamente da desigualdade-canônica de *regret* para Mirror Descent. A seguir, transcrevemos o resultado ([Bub15b, Teorema 4.3]) que usamos para as escolhas dos tamanhos de passo, e derivamos a escolha ótima de η para duas geometrias distintas:

- $\Psi(x) = \frac{1}{2}\|x\|_2^2$, a função geradora da norma euclidiana, para o GD e AGD;
- $\Phi(x) = \sum_{i=1}^n x_i \log x_i$, a função geradora da entropia negativa, para o MD e AMD.

Teorema 8.1 (Bubeck, 2015)**.** *Seja* Φ *um* mirror map *ν-fortemente convexo sobre $\mathcal{X}\cap\mathcal{D}$ com respeito à norma $\|\cdot\|$ Seja $x_0 = \arg\min_{x\in\mathcal{X}\cap\mathcal{D}}\Phi(x)$, e defina $R^2 = \sup_{x\in\mathcal{X}\cap\mathcal{D}}\Phi(x)-\Phi(x_0)$. Se f é convexa e L-Lipschitz com respeito à norma $\|\cdot\|$, o algoritmo de* dual averaging *com $\eta = \frac{R}{L}\sqrt{\frac{\nu}{2T}}$ satisfaz*

$$f\left(\frac{1}{T}\sum_{t=1}^T x_t\right) - f(x^*) \leq 2RL\sqrt{\frac{2}{\nu T}}.$$

O *Dual Averaging* (DA) é definido por

$$\begin{cases} \theta_{k+1} &= \theta_0 - \eta\sum_{i=0}^k \nabla f(x_i), \\ x_{k+1} &= (\nabla\Phi)^{-1}(\theta_{k+1}) \end{cases}$$

onde $\theta_0 = \vec{0}$ ([FHPF22, Algoritmo 1]). A primeira equação nos leva a

$$\begin{aligned} \theta_{k+1} &= \theta_0 - \eta\nabla f(x_0) - \eta\nabla f(x_1) - \eta\nabla f(x_2) - \cdots - \eta\nabla f(x_k) \\ &= \theta_1 - \eta\nabla f(x_1) - \eta\nabla f(x_2) - \cdots - \eta\nabla f(x_k) \\ &= \theta_2 - \eta\nabla f(x_2) - \cdots - \eta\nabla f(x_k) \\ &\vdots \\ &= \theta_k - \eta\nabla f(x_k), \end{aligned}$$

que é igual à equação do passo do gradiente do MD. Como a segunda equação também é igual à primeira equação do MD (5.2), temos que os dois métodos são equivalentes quando $\theta_0 = \vec{0}$, de forma que o uso do teorema se justifica (vamos escolher x_0 satisfazendo essa condição).

Para o caso euclidiano, a norma usada é $\|\cdot\|_2$ e a função geradora é $\Psi(x) = \frac{1}{2}\|x\|_2^2$, que é 1-fortemente convexa com respeito à norma euclidiana, de modo que $\nu = 1$. Como $\mathcal{X}$ é o simplexo, o valor de R^2 é obtido com

$$R^2 = \sup_{x \in \mathcal{X}} \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|x_0\|_2^2 .$$

Esse cálculo pode ser feito analiticamente. No simplexo, $\|x\|_2^2$ é maximizado em um vértice e_i, com $\|e_i\|_2^2 = 1$, e minimizado em $(1/n, \ldots, 1/n)$, que é a escolha natural para x_0. Assim, como $\|x_0\|_2^2 = 1/n^2 + 1/n^2 + \cdots + 1/n^2 = n/n^2 = 1/n$, temos

$$R^2 = \frac{1}{2} - \frac{1}{2n} = \frac{n-1}{2n}.$$

O parâmetro L_2 (constante de Lipschitz de f com respeito à norma ℓ^2) é tomado como o maior valor de $\|\nabla f(x)\|_2$ no simplexo, o que foi calculado numericamente por varredura em malha fina. Ressaltamos que esse método não é o mais prático de maneira geral.

O passo usado para o GD e para o AGD foi

$$\eta_{\mathrm{GD}} = \frac{\tilde{\eta}}{L_2}\sqrt{\frac{n-1}{4nT}} = \frac{\tilde{\eta}}{2L_2}\sqrt{\frac{n-1}{nT}},$$

onde T é o número de iterações (200) e $\tilde{\eta} = 0{,}001$ a fim de melhorar a visualização dos comportamentos; esse mesmo fator $\tilde{\eta}$ foi usado para escalar o passo dos métodos MD e AMD.

No caso da entropia negativa, a norma usada é a ℓ_1 e a função geradora é $\Phi(x) = \sum_{i=1}^n x_i \log x_i$, que é 1-fortemente convexa com respeito à norma $\|\cdot\|_1$ (pelo menos no simplexo), de modo que $\nu = 1$ também. Para calcular R^2 analiticamente, notamos que o valor máximo de $\Phi(x)$ no simplexo ocorre em um vértice ($x = e_i$, com $\Phi(e_i) = 0$), e o mínimo em $x_0 = (1/n, \ldots, 1/n)$, de novo, a escolha natural para x_0, onde $\Phi(x_0) = \sum_{i=1}^n \frac{1}{n}\log\frac{1}{n} = -\log n$. Portanto,

$$R^2 = 0 - (-\log n) = \log n.$$

Já L_∞ é a constante de Lipschitz de f em relação à norma ℓ_1, isto é,

$$L = \sup_{x \in \mathcal{X}} \|\nabla f(x)\|_\infty ,$$

que nos experimentos também foi estimado numericamente (de novo, não a maneira mais prática de se fazer isso).

O passo usado para o MD e AMD foi, portanto

$$\eta_{\mathrm{MD}} = \frac{\tilde{\eta}}{L_\infty}\sqrt{\frac{\log n}{2T}}.$$

A Tabela 8.1 mostra os valores reais calculados. Notemos que os passos para o MD e AMD são consideravelmente maiores que os passos para o GD e AGD. Isso é possível e justo porque os passos dados pelo MD são dados "sem restrição" no espaço dual, uma vez que o mapeamento é sobrejetivo e coercivo, permitindo que os pontos duais divirjam. Na verdade, como em geral a solução estará no bordo do simplexo (por exemplo, em um dos vértices) a trajetória dual irá divergir. É como se a função a ser minimizada no espaço dual, $f((\nabla\Phi)^{-1}(\theta))$, não tivesse mínimo. Dessa maneira, dar passos maiores não só faz sentido como é útil para a convergência do algoritmo.

Dimensão	$\boldsymbol{\eta_{\mathrm{GD}}}$	$\boldsymbol{\eta_{\mathrm{MD}}}$
100	$1{,}08 \times 10^{-7}$	$2{,}49 \times 10^{-4}$
10.000	$3{,}46 \times 10^{-8}$	$2{,}01 \times 10^{-4}$

Tabela 8.1: Tamanhos de passo para GD (Gradient Descent) e MD (Mirror Descent) em diferentes dimensões.

Comparamos os métodos em dois casos: dimensão 100 e dimensão 10.000. O *Mirror Descent* sabidamente se sai melhor em alta dimensão. Isso se dá em parte porque nas configurações do simplexo, a projeção

usada pelo MD é a de Kullback-Leibler, que tem custo computacional $\mathcal{O}(n)$, onde n é a dimensão do espaço, enquanto a projeção euclidiana tem custo $\mathcal{O}(n \log n)$, normalmente por conta de uma ordenação dos elementos do vetor a ser projetado ([WCP13]). Existem maneiras de calcular a projeção euclidiana de um ponto no simplexo com custo teórico $\mathcal{O}(n^2)$, mas com um desempenho prático comumente melhor, $\mathcal{O}(n)$. Uma maneira é usar o método de Newton para resolver o problema de mínimos quadrados com restrição no simplexo ([CMS14]). A Figura 8.1 mostra as médias de tempo (em escala logarítmica) necessário para realizar cada uma das projeções em função da dimensão (também em escala logarítmica); as médias foram tomadas de 50 amostras de projeções de vetores aleatórios no simplexo.

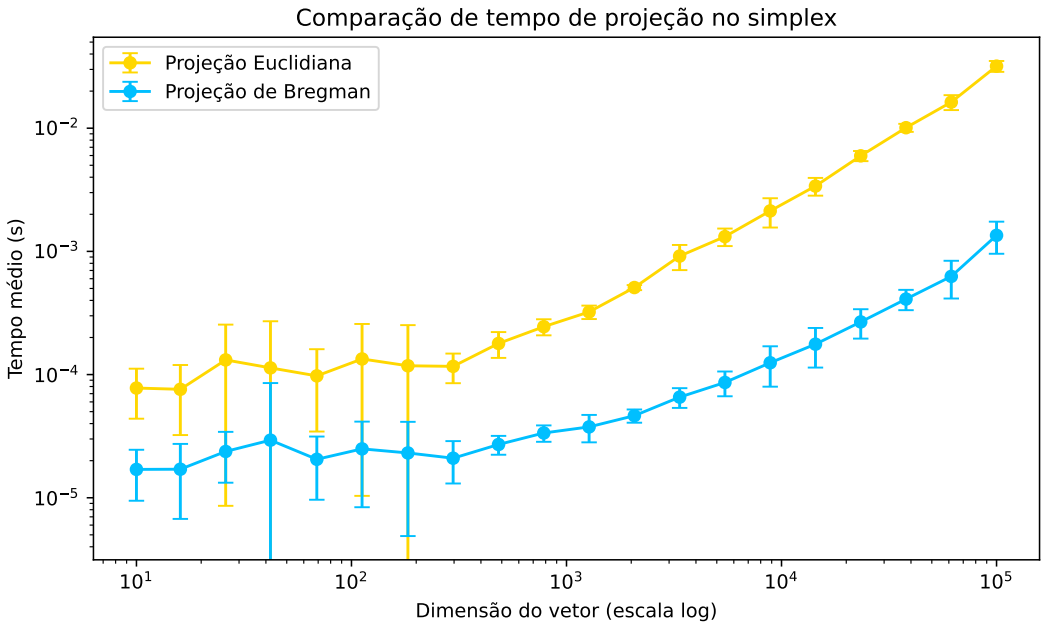

Figura 8.1: Comparação de custo (em tempo) das projeções euclidianas e de Kullback-Leibler em função da dimensão do espaço (escala loglog).

As figuras 8.2 e 8.3 mostram os erros em escala logarítmica simétrica (as figuras correspondentes em escala log normal podem ser encontradas no Apêndice C) no valor da f dos quatro métodos em função do tempo de execução. Escolhemos usar o tempo de execução em vez das iterações para notar que alguns métodos convergem mais rápido, não somente em iterações, mas em tempo mesmo, evidenciando o efeito dos custos das projeções. As iterações podem ser visualizadas pelos pontos nas curvas: cada ponto $(t_k, f(x_k) - f^*)$ representa uma iteração. O efeito dos custos de projeção, apesar de não ter sido grande, pode ser notado observando que os tempos finais, isto é, da última iteração, para os métodos MD e AMD são menores (mais à esquerda) do que os tempos finais para os GD e AGD.

Notemos que as convergências dos métodos são bem parecidas em baixa dimensão (100), mas têm diferenças expressivas em alta dimensão (10.000). Como era esperado, os métodos acelerados representam uma melhora em relação aos não acelerados. Além disso, os métodos *mirror* também representam uma melhora em relação aos métodos mais usuais. Desta forma, o GD acaba sendo o pior dentre os quatro e o AMD, o melhor. Por fim, o MD ainda apresentou um desempenho melhor do que o do AGD, mostrando que a adaptação à geometria do problema é bastante importante, principalmente para problemas em maior dimensão.

No Apêndice C se encontram também as figuras que apresentam a convergência dos pontos $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_n$ gerados pelos métodos, o que dialoga mais com o Teorema 8.1. Além disso, no Apêndice D, apresentamos as taxas de convergência empíricas relativas a este experimento numérico.

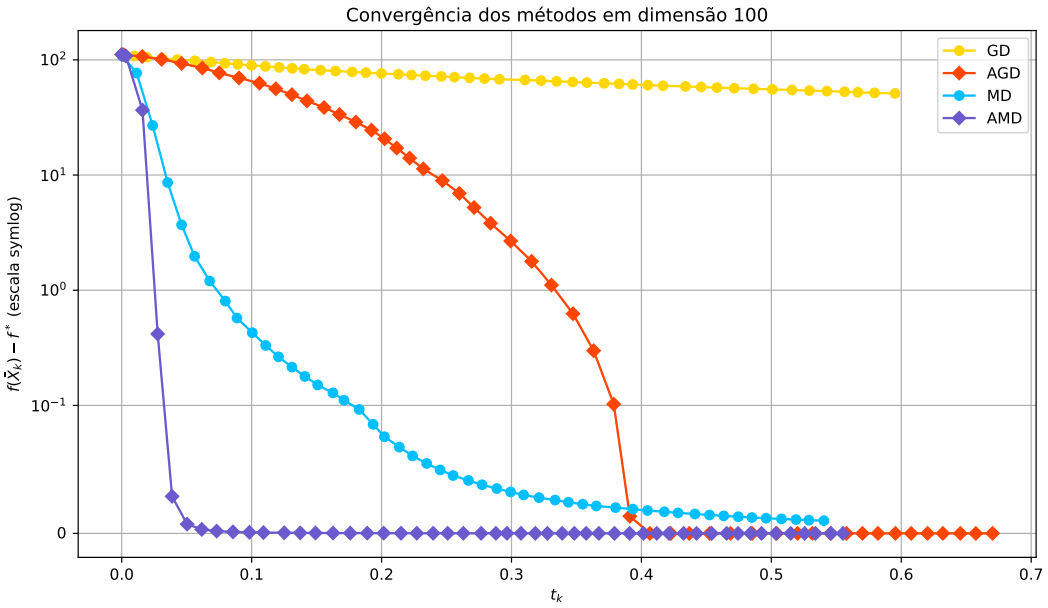

Figura 8.2: Comparação das convergências dos métodos GD, AGD, MD e AMD para um problema em dimensão 100.

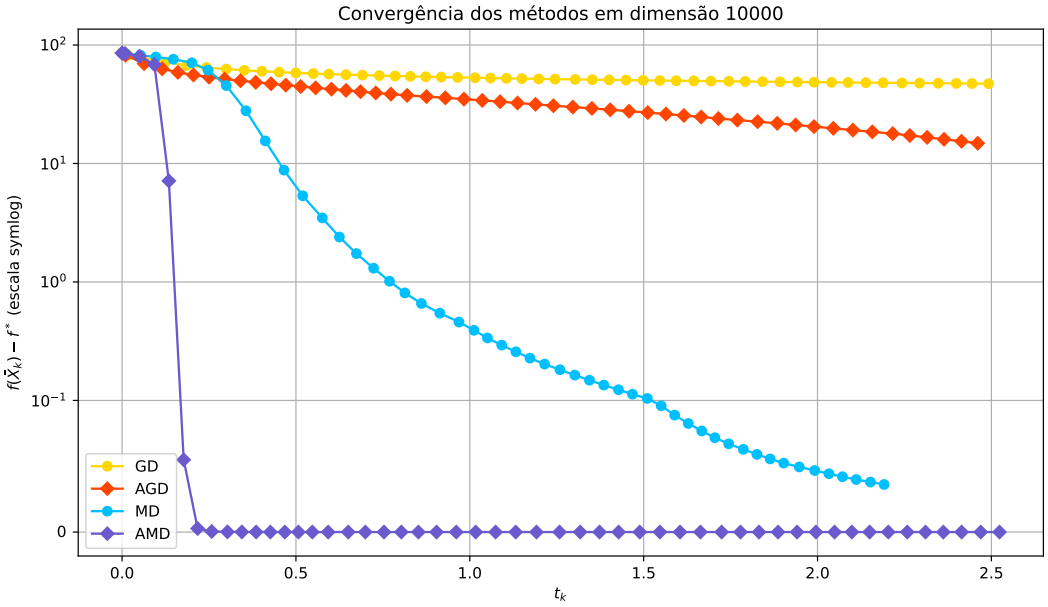

Figura 8.3: Comparação das convergências dos métodos GD, AGD, MD e AMD para um problema em dimensão 10000.

Capítulo 9

Conclusão

Neste trabalho estudamos três métodos de primeira ordem — *Gradient Descent* (GD), *Accelerated Gradient Descent* (AGD) e *Mirror Descent* (MD) —, já clássicos em otimização convexa, bem como suas versões projetadas. Propusemos ainda um quarto método, o *Accelerated Mirror Descent* (AMD), que combina as ideias de aceleração e de mapeamento espelho.

Para os três primeiros métodos, discutimos suas taxas de convergência quando conhecidas, assim como suas conexões com equações diferenciais ordinárias (EDOs). As soluções dessas EDOs podem ser vistas como casos limite das trajetórias discretizadas pelos métodos. Assim, a interpretação contínua permite transferir resultados sobre taxas de convergência de soluções de EDOs para os algoritmos discretos, desde que se mantenham as mesmas condições iniciais.

No caso do AMD, apresentamos a EDO correspondente e estabelecemos dois resultados centrais: i) existência e unicidade da solução sob certas condições iniciais e ii) convergência da trajetória produzida pelo algoritmo para a solução da EDO associada.

Exploramos também a aplicação do MD e do AMD em otimização sobre o simplexo, derivando expressões explícitas para os iterados e permitindo uma visualização clara do efeito de cada passo.

A análise numérica concentrou-se em um experimento simples, voltado a avaliar: i) os efeitos da aceleração no GD e no MD; ii) o impacto da escolha de um mapeamento espelho; e iii) a influência da dimensão do problema. Mostramos ainda que, no simplexo, os métodos baseados em espelho apresentam vantagens de custo devido à projeção de Kullback–Leibler, mais eficiente que a projeção euclidiana.

Quanto a trabalhos futuros, há diversas direções possíveis. Uma primeira linha é investigar a aplicação do AMD em outros contextos em que o MD já é conhecido por ser vantajoso, como no caso do espectraedro (*spectrahedron*). Nessa direção, seria relevante verificar empiricamente — ou idealmente provar de forma analítica — se a aceleração traz ganhos significativos também nesse cenário.

Além disso, dado que a conexão entre o método e sua EDO associada está bem estabelecida, abre-se a possibilidade de estudar taxas de convergência do AMD por meio da análise contínua, por exemplo utilizando funções de Lyapunov.

Outra perspectiva interessante é considerar variantes estocásticas do AMD, que poderiam ser aplicadas em problemas de larga escala ou de aprendizado de máquina, onde já há amplo uso de GD estocástico e suas variantes aceleradas. Também seria valioso estudar versões adaptativas, em que o tamanho de passo se ajusta automaticamente ao longo das iterações.

Do ponto de vista prático, seria útil avaliar o desempenho do AMD em aplicações concretas de otimização convexa de alta dimensão, como problemas em ciência de dados, otimização sobre matrizes positivas semidefinidas, ou mesmo em contextos de teoria dos jogos (como algoritmos de pesos multiplicativos).

Referências Bibliográficas

[AB06] Charalambos D Aliprantis and Kim C Border. *Infinite dimensional analysis: a hitchhiker's guide*. Springer, 2006.

[AHK12] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of computing*, 8(1):121–164, 2012.

[Anu20] Gupta Anupam. Advanced algorithms (notes for cmu 15-850, fall 2020). Carnegie Mellon University Lecture Notes, 2020.

[AP95] Antonio Ambrosetti and Giovanni Prodi. *A primer of nonlinear analysis*. Cambridge University Press, 1995.

[Arm66] Larry Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of mathematics*, 16(1):1–3, 1966.

[Bat93] Sean M Bates. Toward a precise smoothness hypothesis in sard's theorem. *Proceedings of the American Mathematical Society*, pages 279–283, 1993.

[BBB89] Andrew A Brown and Michael C Bartholomew-Biggs. Some effective methods for unconstrained optimization based on the solution of systems of ordinary differential equations. *Journal of Optimization Theory and Applications*, 62(2):211–224, 1989.

[BBF16] Annette Burden, Richard Burden, and J. Faires. *Numerical Analysis, 10th ed.* Cengage Learning, 01 2016.

[BCB+12] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[Blo94] Anthony Bloch. *Hamiltonian and gradient flows, algorithms and control*, volume 3. American Mathematical Soc., 1994.

[BT03] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[BT09] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[BTMN01] Aharon Ben-Tal, Tamar Margalit, and Arkadi Nemirovski. The ordered subsets mirror descent optimization method with applications to tomography. *SIAM Journal on Optimization*, 12(1):79–108, 2001.

[BTN01] Aharon Ben-Tal and Arkadi Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. SIAM, 2001.

[Bub15a] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[Bub15b] Sébastien Bubeck. *Convex optimization: Algorithms and complexity*, volume 8. Foundations and Trends® in Machine Learning, 2015.

[Cal20] Anderson de Oliveira Calixto. Métodos de otimização aplicados à estatística. Master's thesis, Universidade Federal do Rio de Janeiro, 2020.

[Car12] Henri Cartan. *Differential forms*. Courier Corporation, 2012.

[CBL06] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[CMS14] Roberto Cominetti, Walter F Mascarenhas, and Paulo JS Silva. A newton's method for the continuous quadratic knapsack problem. *Mathematical Programming Computation*, 6(2):151–169, 2014.

[Die11] Jean Dieudonné. *Foundations of modern analysis*. Read Books Ltd, 2011.

[Eva18] LawrenceCraig Evans. *Measure theory and fine properties of functions*. Routledge, 2018.

[FHPF22] Huang Fang, Nicholas JA Harvey, Victor S Portella, and Michael P Friedlander. Online mirror descent and dual averaging: keeping pace in the dynamic case. *Journal of Machine Learning Research*, 23(121):1–38, 2022.

[GG23] Guillaume Garrigos and Robert M Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

[Gol65] Allen A Goldstein. On steepest descent. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 3(1):147–151, 1965.

[Haz16] Elad Hazan. *Introduction to online convex optimization*. Foundations and Trends® in Optimization, 2 edition, 2016.

[HM12] Uwe Helmke and John B Moore. *Optimization and dynamical systems*. Springer Science & Business Media, 2012.

[Hun07] John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[Ius95] A. Iusem. Método de ponto proximal em otimização. Livro do 20º Colóquio Brasileiro de Matemática, IMPA - CNPq, 1995.

[JN11] Anatoli Juditsky and Arkadi Nemirovski. First-order methods for nonsmooth convex large-scale optimization, i: General purpose methods. *Optimization for Machine Learning*, pages 121–148, 2011.

[JNT11] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[Kar39] William Karush. Minima of functions of several variables with inequalities as side constraints. *M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago*, 1939.

[KBB15] Walid Krichene, Alexandre M. Bayen, and Peter L. Bartlett. Accelerated mirror descent in continuous and discrete time. *Advances in Neural Information Processing Systems*, 28:2845–2853, 2015.

[KKB15] Walid Krichene, Syrine Krichene, and Alexandre Bayen. Efficient bregman projections onto the simplex. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 3291–3298. IEEE, 2015.

[KL51] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[KT51] Harold W. Kuhn and Albert W. Tucker. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492. University of California Press, 1951.

[Lan12] Serge Lang. *Real and functional analysis*, volume 142. Springer Science & Business Media, 2012.

[Mar70] Bernard Martinet. Régularisation d'inéquations variationnelles par approximations successives. rev. française informat. *Recherche Opérationnelle*, 4:154–158, 1970.

[Mat25] MathCha.io. Mathcha editor. `https://www.mathcha.io/`, 2025. Acesso em: 15 ago. 2025.

[Nes83] Yurii E. Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547, 1983.

[Nes05] Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[Nes13] Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013.

[Nes18] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[NY83a] Arkadi Nemirovski and David Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

[NY83b] Arkadii Semenovich Nemirovskij and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

[Pol64] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[Pyt23] Python Software Foundation, https://www.python.org. *Python: A dynamic, open source programming language*, 2023. Version 3.x.

[Roc76] R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[SBC16] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

[SH98] Andrew M Stuart and Anthony R Humphries. *Dynamical systems and numerical analysis*, volume 2. Cambridge university press, 1998.

[SS00] Johannes Schropp and I Singer. A dynamical systems approach to constrained minimization. *Numerical functional analysis and optimization*, 21(3-4):537–551, 2000.

[vS15] Sebastian van Strien. MA2AA1 (ODE's): Lecture Notes. `https://www.ma.imperial.ac.uk/~svanstri/Files/de-4th.pdf`, 2015. Imperial College London, Lecture Notes on ODEs, Theorem 5.5: Continuous Dependence on Initial Conditions and Functions.

[WCP13] Weiran Wang and Miguel A Carreira-Perpinán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *arXiv preprint arXiv:1309.1541*, 2013.

[Wik25] Wikipedia contributors. Grönwall's inequality — Wikipedia, the free encyclopedia, 2025. [Online; accessed 29-August-2025].

[Wol69] Philip Wolfe. Convergence conditions for ascent methods. *SIAM review*, 11(2):226–235, 1969.

[Wol71] Philip Wolfe. Convergence conditions for ascent methods. ii: Some corrections. *SIAM review*, 13(2):185–188, 1971.

[WWJ16] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

[Zei12] Eberhard Zeidler. *Applied functional analysis: applications to mathematical physics*, volume 108. Springer Science & Business Media, 2012.

Apêndice A

Mirror Descent

Neste apêndice provamos alguns resultados auxiliares para o estudo do *Mirror Descent* e *Accelerated Mirror Descent*, com foco no mapeamento espelho usado por ambos.

Teorema A.1. $\forall k \in \mathbb{N}, (\nabla\Phi(x_{k+1}) - \nabla\Phi(x_k))(x_{k+1} - x_k) \geq 0$

Demonstração. Como Φ é convexa, temos que

$$\Phi(y) \geq \Phi(x) + \nabla\Phi(x)(y - x)$$

ou

$$\Phi(y) - \Phi(x) \geq \nabla\Phi(x)(y - x).$$

Fazendo $y = x_k$ e $x = x_{k+1}$, obtemos

$$\Phi(x_k) - \Phi(x_{k+1}) \geq \nabla\Phi(x_{k+1})(x_k - x_{k+1}).$$

Agora, fazendo $y = x_{k+1}$ e $x = x_k$, obtemos

$$\Phi(x_{k+1}) - \Phi(x_k) \geq \nabla\Phi(x_k)(x_{k+1} - x_k).$$

Somando as duas equações, ficamos com

$$\begin{aligned} 0 = \Phi(x_k) - \Phi(x_{k+1}) + \Phi(x_{k+1}) - \Phi(x_k) &\geq \nabla\Phi(x_{k+1})(x_k - x_{k+1}) + \nabla\Phi(x_k)(x_{k+1} - x_k) \\ &= \nabla\Phi(x_{k+1})(x_k - x_{k+1}) - \nabla\Phi(x_k)(x_k - x_{k+1}) \\ &= (\nabla\Phi(x_{k+1}) - \nabla\Phi(x_k))(x_k - x_{k+1}) \end{aligned}$$

Desta maneira

$$(\nabla\Phi(x_{k+1}) - \nabla\Phi(x_k))(x_k - x_{k+1}) \leq 0$$

ou ainda

$$-(\nabla\Phi(x_{k+1}) - \nabla\Phi(x_k))(x_k - x_{k+1}) = (\nabla\Phi(x_{k+1}) - \nabla\Phi(x_k))(x_{k+1} - x_k) \geq 0. \qquad \blacksquare$$

Observação. *Este resultado é conhecido como monotonicidade do gradiente para funções convexas. Ele garante que o passo dado pelo Mirror Descent seja positivamente correlacionado ao passo que seria dado pelo Gradient Descent. Isso pode ser visualizado notando que o campo gradiente da função Φ funciona como uma fonte, mais ou menos da mesma maneira que o campo gradiente identidade do espaço dual.*

Fato A.2. *Seja $\Phi : \mathbb{R}^n \to \mathbb{R}$ uma função ν-fortemente convexa. Então*

$$\|\nabla\Phi(x) - \nabla\Phi(y)\| \geq \nu \|x - y\|.$$

Demonstração. De maneira parecida com a prova do teorema A.1, pela ν-convexidade, temos que

$$\Phi(x) - \Phi(y) \geq \nabla\Phi(x)(y-x) + \frac{\nu}{2}\|x-y\|^2$$

e

$$\Phi(y) - \Phi(x) \geq \nabla\Phi(y)(x-y) + \frac{\nu}{2}\|x-y\|^2 .$$

Somando as duas inequações, obtemos

$$0 \geq (\nabla\Phi(x) - \nabla\Phi(y))(y-x) + \nu\|x-y\|^2$$

ou ainda

$$(\nabla\Phi(x) - \nabla\Phi(y))(x-y) \geq \nu\|x-y\|^2 .$$

Agora, pela desigualdade de Cauchy-Schwartz (2.2), temos então que

$$\|\nabla\Phi(x) - \nabla\Phi(y)\|\,\|x-y\| \geq (\nabla\Phi(x) - \nabla\Phi(y))(x-y) \geq \nu\|x-y\|^2 .$$

Por fim, dividindo ambos os lados por $\|x-y\|$, obtemos

$$\|\nabla\Phi(x) - \nabla\Phi(y)\| \geq \nu\|x-y\| . \qquad \blacksquare$$

Corolário A.3. *Seja $\Phi : \mathbb{R}^n \to \mathbb{R}$ uma função ν-fortemente convexa. Então $(\nabla\Phi)^{-1}(\cdot)$ é $\frac{1}{\nu}$-Lipschitz.*

Demonstração. Usando o fato A.2 e escrevendo $\theta = \nabla\Phi(x)$ e $\xi = \nabla\Phi(y)$ e consequentemente $x = (\nabla\Phi)^{-1}(\theta)$ e $y = (\nabla\Phi)^{-1}(\xi)$, temos

$$\|\theta - \xi\| \geq \nu\left\|(\nabla\Phi)^{-1}(\theta) - (\nabla\Phi)^{-1}(\xi)\right\| \quad \Leftrightarrow \quad \left\|(\nabla\Phi)^{-1}(\theta) - (\nabla\Phi)^{-1}(\xi)\right\| \leq \frac{1}{\nu}\|\theta - \xi\| ,$$

o que caracteriza $(\nabla\Phi)^{-1}(\cdot)$ como uma função $\frac{1}{\nu}$-Lipschitz. $\blacksquare$

Observação. *Usando a notação introduzida pela EDO suavizada 7.1, este corolário é escrito como*

$$\|\Theta(X(t)) - \Theta(X(0))\| \geq \nu\|X(t) - X(0)\| = \nu\|X(t) - x_0\|$$

ou mesmo

$$\|X(t) - x_0\| \leq \frac{1}{\nu}\|\Theta(X(t)) - \Theta(X(0))\| .$$

Apêndice B

EDO

Neste apêndice trazemos resultados auxiliares usados no Capítulo 7.

B.1 Teoria das EDOs

O teorema da dependência contínua nas condições iniciais se encontra abaixo e uma versão mais forte, bem como sua demonstração, pode ser encontrada em [vS15, Teorema 5.5]. Ele foi usado para mostrar que, com $\delta \to 0$, as soluções da EDO (6.1) com condições se aproximando também se aproximam.

Teorema B.1 (Dependência Contínua nas Condições Iniciais)**.** *Considere a EDO:*

$$\frac{\mathrm{d}y}{\mathrm{d}t} = F(t, y), \quad y(t_0) = y_0,$$

onde F é contínua em t e Lipschitz em y (uniformemente em t). Seja $\{y_{0,n}\}_{n\in\mathbb{N}}$ uma sequência com $y_{0,n} \to y_0$, e denote por $y_n(t)$ a solução única correspondente à condição inicial $y_n(t_0) = y_{0,n}$. Então, em qualquer intervalo compacto $[t_0 - \tau, t_0 + \tau]$ contido no domínio da solução $y(t)$, a sequência $\{y_n\}$ converge uniformemente para y.

B.2 Convergência do método

Nesta seção, nos concentramos nos resultados usados na demonstração da convergência do AMD para a solução da EDO associada.

B.2.1 Sobre a limitação de $f(y_k)$

Vamos provar que $f(y_k)$, onde y_k são os pontos auxiliares primais produzidos pelo AMD, é limitado para todo $s < \varsigma$ e todo $k \in [K_s^*]$. Para isso, consideremos primeiro o Teorema de Grönwall a seguir.

Teorema B.2 (Desigualdade de Grönwall)**.** *Seja $u : [0, T] \to [0, \infty)$ contínua e suponha que existam constantes $A, B \geq 0$ tais que*

$$u(t) \leq A + B \int_0^t u(\tau)\, \mathrm{d}\tau, \quad \forall t \in [0, T].$$

Então

$$u(t) \leq A\, e^{Bt}, \quad \forall t \in [0, T].$$

Mais geralmente, se

$$u(t) \leq A + Ct + B \int_0^t u(\tau)\, \mathrm{d}\tau,$$

onde $C \geq 0$ é constante, então

$$u(t) \leq (A + Ct)e^{Bt}.$$

A demonstração deste teorema pode ser encontrada em [Wik25, Integral form for continuous functions (especialmente o item (b))]. Com ele podemos provar o fato que gostaríamos.

Fato B.3. *Dado $T > 0$ fixo, existem $\varsigma > 0$ e uma constante M (dependendo apenas de constantes como T, L, ν e das condições iniciais, também vistas como constantes) tais que para todo $0 < s \leq \varsigma$ a sequência $\{y_k\}_{k=0}^{K_s^*}$ está contida em uma bola fechada de raio dependendo apenas de T, L, ν e dados iniciais.*

Aí para todo $0 < s \leq \varsigma$ e todo índice k com $0 \leq k \leq K_s^*$, temos que

$$f(y_k) \leq M,$$

e consequentemente,

$$C_1 = \sup_{\substack{s>0 \\ k\in[K_s^*]}} \sqrt{2L(f(y_k) - f^*)} < +\infty.$$

Demonstração. A prova é dividida em cinco etapas. A primeira delas consiste de introduções de novas variáveis e identidades básicas relativas e estas.

Definimos $d_k := \theta_k - \theta_{k-1}$ para $k \geq 1$. Assim, podemos escrever $\xi_k = \theta_k + \mu_k d_k$. Aí, da definição do AMD com passo s,

$$d_{k+1} = \theta_{k+1} - \theta_k = \xi_k - s\nabla f(y_k) - \theta_k = \mu_k d_k - s\nabla f(y_k). \tag{B.1}$$

Para as variáveis ξ_k, a diferença é calculada como

$$\begin{aligned}\xi_{k+1} - \xi_k &= \theta_{k+1} + \mu_{k+1}d_{k+1} - \theta_k - \mu_k d_k \\ &= d_{k+1} + \mu_{k+1}d_{k+1} - \mu_k d_k = (1 + \mu_{k+1})d_{k+1} - \mu_k d_k.\end{aligned}$$

Substituindo (B.1) nessa última equação e reorganizando termos, obtemos

$$\begin{aligned}\xi_{k+1} - \xi_k &= (1 + \mu_{k+1})(\mu_k d_k - s\nabla f(y_k)) - \mu_k d_k \\ &= \mu_k d_k + \mu_{k+1}\mu_k d_k - (1 + \mu_{k+1})s\nabla f(y_k) - \mu_k d_k \\ &= \mu_{k+1}\mu_k d_k - (1 + \mu_{k+1})s\nabla f(y_k).\end{aligned} \tag{B.2}$$

Usando $0 \leq \mu_k \leq 1$ e $1 + \mu_{k+1} \leq 2$, chegamos à desigualdade

$$\|\xi_{k+1} - \xi_k\| \leq \|d_k\| + 2s \|\nabla f(y_k)\|. \tag{B.3}$$

Seja y^* mínimo global de f; então $f^* = f(y^*)$. Na segunda etapa que se inicia agora, buscamos controlar d_k através de somas de termos como $\|y_i - y^*\|$.

Como $\nabla f(y_*) = 0$ e ∇f é L-Lipschitz, temos

$$\|\nabla f(y_k)\| = \|\nabla f(y_k) - \nabla f(y_*)\| \leq L \|y_k - y_*\|.$$

Usando (B.1) e usando a desigualdade triangular, obtemos

$$\begin{aligned}\|d_{k+1}\| &= \|\mu_k d_k - s\nabla f(y_k)\| \\ &\leq |\mu_k| \|d_k\| + s \|\nabla f(y_k)\| \\ &\leq \|d_k\| + s \|\nabla f(y_k)\| \leq \|d_k\| + sL \|y_k - y^*\|.\end{aligned}$$

Iterando de 1 a $k - 1$,

$$\|d_k\| \leq \|d_1\| + sL \sum_{i=1}^{k-1} \|y_i - y^*\|. \tag{B.4}$$

Substituindo (B.4) em (B.3) e usando $\|\nabla f(y_k)\| \leq L \|y_k - y_*\|$ e $\|y_k - y_*\| \leq \sum_{i=1}^{k} \|y_i - y_*\|$, temos

$$\begin{aligned}\|\xi_{k+1} - \xi_k\| &\leq \|d_1\| + sL \sum_{i=1}^{k-1} \|y_i - y_*\| + 2sL \|y_k - y_*\| \\ &\leq \|d_1\| + sL \sum_{i=1}^{k} \|y_i - y_*\| + sL \|y_k - y_*\| \leq \|d_1\| + 2sL \sum_{i=1}^{k} \|y_i - y_*\|.\end{aligned}$$

Na terceira etapa, passamos a trabalhar somente nas coordenadas primais e definimos uma curva que interpola os pontos que nos será útil.

Usando A.3

$$\|y_{k+1} - y_k\| \leq \frac{1}{\nu} \|\xi_{k+1} - \xi_k\| . \tag{B.5}$$

Combinando com a desigualdade anterior, obtemos

$$\|y_{k+1} - y_k\| \leq \frac{1}{\nu} \|d_1\| + \frac{2Ls}{\nu} \sum_{i=1}^{k} \|y_i - y_*\| \quad \overset{\div\sqrt{s}}{\Longrightarrow} \quad \frac{\|y_{k+1} - y_k\|}{\sqrt{s}} \leq \frac{1}{\nu} \frac{\|d_1\|}{\sqrt{s}} + \frac{2L\sqrt{s}}{\nu} \sum_{i=1}^{k} \|y_i - y_*\| . \tag{B.6}$$

Agora definamos $t_k := k\sqrt{s}$ e a interpolação linear por partes $y^s : [0, T] \to \mathbb{R}^n$ tal que $y^s(t_k) = y_k$ e y^s é afim em cada intervalo $[t_k, t_{k+1}]$. Isto é, y^s é a curva contínua formada pelo segmentos de reta unindo y_k e y_{k+1} para cada $k \in [K_s^* - 1]$. Definamos ainda

$$\gamma_s(t) := \sup_{0 \leq \tau \leq t} \|y^s(\tau) - y_*\| .$$

Queremos provar que γ_s é finito para qualquer s.

Para todo $i \in [K_s^* - 1]$ e $t \in [t_i, t_{i+1}]$, temos

$$\|y_i - y_*\| = \|y^s(t_i) - y_*\| \leq \gamma_s(t) \quad \Longrightarrow \quad \sum_{i=1}^{k} \|y_i - y_*\| \leq k\, \gamma_s(t) = \frac{t_k}{\sqrt{s}} \gamma_s(t) \leq \frac{T}{\sqrt{s}} \gamma_s(t).$$

Substituindo isso em (B.6), temos que para $t \in [0, T]$,

$$\frac{\|y_{k+1} - y_k\|}{\sqrt{s}} \leq \frac{1}{\nu} \frac{\|d_1\|}{\sqrt{s}} + \frac{2LT}{\nu} \gamma_s(t).$$

Para $t \in (t_k, t_{k+1})$, por ser uma função linear por partes, a derivada de y^s satisfaz $\left\| \frac{\mathrm{d}}{\mathrm{d}t} y^s(t) \right\| = \frac{\|y_{k+1} - y_k\|}{\sqrt{s}}$ e aí, como γ_s é não decrescente,

$$\begin{aligned} \left\| \frac{\mathrm{d}}{\mathrm{d}t} y^s(t) \right\| &\leq \frac{1}{\nu} \frac{\|d_1\|}{\sqrt{s}} + \frac{2L}{\nu} \int_0^t \gamma_s(\tau) \mathrm{d}\tau \\ &\leq \frac{1}{\nu} \frac{\|d_1\|}{\sqrt{s}} + \frac{2L}{\nu} \int_0^t \gamma_s(t) \mathrm{d}\tau \\ &= \frac{1}{\nu} \frac{\|d_1\|}{\sqrt{s}} + \frac{2L}{\nu} \gamma_s(t) t \leq \frac{1}{\nu} \frac{\|d_1\|}{\sqrt{s}} + \frac{2LT}{\nu} \gamma_s(t). \end{aligned} \tag{B.7}$$

Na quarta etapa, buscamos o controle do termo $\frac{\|d_1\|}{\sqrt{s}}$ e aplicamos a desigualdade de Grönwall.

Da inicialização do método, temos $d_1 = -s\nabla f(y_0)$, de modo que

$$a_s := \frac{\|d_1\|}{\sqrt{s}} = \sqrt{s}\, \|\nabla f(y_0)\| \xrightarrow[s \to 0]{} 0.$$

Do Teorema Fundamental do Cálculo, temos

$$\int_0^t \frac{\mathrm{d}}{\mathrm{d}u} y^s(u) \mathrm{d}u = y^s(t) - y_0 = (y^s(t) - y^*) - (y_0 - y^*),$$

o que, usando a desigualdade triangular e a desigualdade triangular para integrais, nos leva a

$$\|y^s(t) - y^*\| \leq \|y_0 - y^*\| + \int_0^t \left\| \frac{\mathrm{d}}{\mathrm{d}u} y^s(\tau) \right\| \mathrm{d}\tau.$$

Substituindo B.7 nessa desigualdade, obtemos

$$\|y^s(t) - y_*\| \leq \|y_0 - y_*\| + \int_0^t \left(\frac{1}{\nu} a_s + \frac{2LT}{\nu} \gamma_s(\tau) \right) \mathrm{d}\tau = \|y_0 - y_*\| + \frac{1}{\nu} a_s\, t + \frac{2LT}{\nu} \int_0^t \gamma_s(\tau)\, \mathrm{d}\tau.$$

Tomando supremo em $[0,t]$ no lado esquerdo, temos

$$\gamma_s(t) \leq R_0 + \frac{1}{\nu} a_s\, t + \frac{2LT}{\nu} \int_0^t \gamma_s(\tau)\, \mathrm{d}\tau,$$

onde $R_0 := \|y_0 - y_*\|$. Pela desigualdade de Grönwall (vide B.2), para todo $0 \leq t \leq T$ vale que

$$\gamma_s(t) \leq \left(R_0 + \frac{1}{\nu} a_s\, t \right) \exp\left(\frac{2LT}{\nu}\, t \right) \leq \left(R_0 + \frac{1}{\nu} a_s\, T \right) \exp\left(\frac{2LT^2}{\nu} \right).$$

Como $a_s = \sqrt{s}\, \|\nabla f(y_0)\| \xrightarrow{s \to 0} 0$, existe $\varsigma > 0$ tal que para todo $0 < s \leq \varsigma$, valha $a_s \leq \frac{\nu}{T} R_0$. Assim, para $s \leq \varsigma$ e $t \in [0,T]$, temos, enfim

$$\begin{aligned}
\gamma_s(t) &\leq \left(R_0 + \frac{T}{\nu} \frac{\nu}{T} R_0 \right) \exp\left(\frac{2LT^2}{\nu} \right) \\
&= (R_0 + R_0) \exp\left(\frac{2LT^2}{\nu} \right) \\
&= 2R_0 \exp\left(\frac{2LT^2}{\nu} \right) =: R(T) < \infty.
\end{aligned}$$

Por fim, como f é contínua, a imagem do conjunto compacto $\{y \in \mathbb{R}^n : \|y - y_*\| \leq R(T)\}$ é limitada. Portanto deve existir

$$M_f(T) := \max_{\|y - y_*\| \leq R(T)} f(y) < +\infty,$$

e aí para todo $k \in [K_s^*]$, temos $f(y_k) \leq M_f(T)$ e, por consequência, $C_1 = \sup_{\substack{s>0 \\ k \in [K_s^*]}} \sqrt{2L(f(y_k) - f^*)} < +\infty$ se $s \in (0, \varsigma)$. ■

B.2.2 Um pouco mais sobre funções Lipschitz

Aqui, demonstramos alguns resultados que serão usados para provar uma limitação para $\|\nabla f(y_k)\|$, onde y_k são os pontos auxiliares primais produzidos pelo AMD. Para os próximos resultados, consideremos a definição a seguir.

Definição. *Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função convexa com gradientes L-Lipschitz e $x \in \mathbb{R}^n$ qualquer. O modelo quadrático de f em x é definido pela função $h : \mathbb{R}^n \to \mathbb{R}$ com*

$$h(y) = f(x) + \nabla f(x)(y - x) + \frac{L}{2} \|y - x\|^2.$$

Fato B.4. *Sejam $f : \mathbb{R}^n \to \mathbb{R}$ uma função convexa com gradiente L-Lipschitz e $x \in \mathbb{R}^n$ qualquer. Seja também $h : \mathbb{R}^n \to \mathbb{R}$ o modelo quadrático de f em x. Então, para todo $y \in \mathbb{R}^n$,*

$$f(y) \leq h(y).$$

Demonstração. Sejam $x, y \in \mathbb{R}^n$ quaisquer e $g(t) = f(x + t(y - x)), t \in [0,1]$. Temos que

$$\frac{\mathrm{d}}{\mathrm{d}t} g(t) = \nabla f(x + t(y - x))(y - x).$$

Além disso, $g(0) = f(x)$ e $g(1) = f(y)$. Portanto

$$\begin{aligned}
f(y) - f(x) = g(1) - g(0) &= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} g(t) \mathrm{d}t = \int_0^1 \nabla f(x + t(y - x))(y - x) \mathrm{d}t \\
\implies \quad f(y) &= f(x) + \int_0^1 \nabla f(x + t(y - x))(y - x) \mathrm{d}t.
\end{aligned}$$

Subtraindo $\nabla f(x)(y-x)$ dos dois lados e usando a desigualdade de Cauchy-Schwarz (2.2) e o fato de ∇f ser L-Lipschitz, ficamos com

$$\begin{aligned}
f(y) - \nabla f(x)(y-x) &= f(x) + \int_0^1 \nabla f(x + t(y-x))(y-x)\mathrm{d}t - \nabla f(x)(y-x) \\
&= f(x) + \int_0^1 \nabla f(x + t(y-x))(y-x)\mathrm{d}t - \int_0^1 \nabla f(x)(y-x)\mathrm{d}t \\
&= f(x) + \int_0^1 (\nabla f(x + t(y-x)) - \nabla f(x))(y-x)\mathrm{d}t \\
&\leq f(x) + \int_0^1 \|\nabla f(x + t(y-x)) - \nabla f(x)\| \cdot \|y-x\| \,\mathrm{d}t \\
&\leq f(x) + \|y-x\| \int_0^1 L \|x + t(y-x) - x\| \,\mathrm{d}t \\
&= f(x) + L\|y-x\| \int_0^1 \|t(y-x)\| \,\mathrm{d}t = f(x) + L\|y-x\| \int_0^1 |t| \cdot \|y-x\| \,\mathrm{d}t \\
&= f(x) + L\|y-x\|^2 \int_0^1 t\mathrm{d}t = f(x) + \frac{L}{2}\|y-x\|^2 .
\end{aligned}$$

Somando novamente $\nabla f(x)(y-x)$ dos dois lados, obtemos enfim

$$f(y) \leq f(x) + \nabla f(x)(y-x) + \frac{L}{2}\|y-x\|^2 = h(y). \qquad \blacksquare$$

Fato B.5. *Sejam $f : \mathbb{R}^n \to \mathbb{R}$ uma função convexa e L-Lipschitz e $x \in \mathbb{R}^n$ quaisquer. Sejam também $h : \mathbb{R}^n \to \mathbb{R}$ o modelo quadrático de f em x e y^* o seu minimizador. Então*

$$\|\nabla f(x)\|^2 = 2L(f(x) - h(y^*)).$$

Demonstração. Pela própria definição de h, temos que $\nabla h(y) = \nabla f(x) + L(y-x)$. Dessa maneira, o ponto y^* deve ser tal que

$$\nabla f(x) + L(y^* - x) = \vec{0} \quad \iff \quad y^* = x - \nabla f(x)/L.$$

Assim, avaliando h em y^*, temos

$$\begin{aligned}
h(y^*) = h\left(x - \frac{\nabla f(x)}{L}\right) &= f(x) + \nabla f(x)\left(x - \frac{\nabla f(x)}{L} - x\right) + \frac{L}{2}\left\|x - x - \frac{\nabla f(x)}{L}\right\|^2 \\
&= f(x) + \nabla f(x)\left(-\frac{\nabla f(x)}{L}\right) + \frac{L}{2}\left\|-\frac{\nabla f(x)}{L}\right\|^2 \\
&= f(x) - \frac{1}{L}\|\nabla f(x)\|^2 + \frac{1}{2L}\|\nabla f(x)\|^2 \\
&= f(x) - \frac{1}{2L}\|\nabla f(x)\|^2 .
\end{aligned}$$

Um rápido rearranjo dos termos nos leva a $\|\nabla f(x)\|^2 = 2L(f(x) - h(y^*))$. $\blacksquare$

Lema B.6. *Sejam $f : \mathbb{R}^n \to \mathbb{R}$ uma função convexa e L-Lipschitz e $x \in \mathbb{R}^n$ quaisquer. Então*

$$\|\nabla f(x)\| \leq \sqrt{2L(f(x) - f^*)}.$$

Demonstração. Sejam $x, y \in \mathbb{R}^n$ quaisquer. Consideremos h, o modelo quadrático de f em x e y^* seu minimizador. Pelo fato B.4, $f^* \leq f(y^*) \leq h(y^*)$. Agora, usando o fato B.5,

$$\|\nabla f(x)\|^2 = 2L(f(x) - h(y^*)) \leq 2L(f(x) - f^*).$$

Tomando a raiz quadrada dos dois lados, temos $\|\nabla f(x)\| \leq \sqrt{2L(f(x) - f^*)}$. $\blacksquare$

B.2.3 Dois fatos auxiliares

Nesta subseção, tratamos de dois fatos auxiliares relativos aos objetos A_s, B_s, c_s e d_s. Os fatos são bem algébricos e não contribuem tanto para a demonstração a ponto de serem incluídos no Capítulo 7, que já é bastante extenso.

Fato B.7. *Para $s > 0$ e $k \in \mathbb{N}$ quaisquer, temos*

$$A_s c_s^k + B_s d_s^k = \frac{C_1\sqrt{s}}{2}(c_s^{k+1} - d_s^{k+1}) + \frac{C_2}{2\sqrt{L}}(c_s^{k+1} + d_s^{k+1}).$$

Demonstração. A demonstração segue do desenvolvimento a seguir.

$$\begin{aligned}
A_s c_s^k + B_s d_s^k &= (a_s + b_s)c_s^k + (a_s - b_s)d_s^k \\
&= a_s(c_s^k + d_s^k) + b_s(c_s^k - d_s^k) \\
&= \left(\frac{C_1 Ls + C_2}{2\sqrt{L}}\right)(c_s^k + d_s^k) + \left(\frac{C_1 + C_2}{2}\sqrt{s}\right)(c_s^k - d_s^k) \\
&= \frac{C_1\sqrt{L}s}{2}(c_s^k + d_s^k) + \frac{C_2}{2\sqrt{L}}(c_s^k + d_s^k) + \frac{C_1\sqrt{s}}{2}(c_s^k - d_s^k) + \frac{C_2\sqrt{s}}{2}(c_s^k - d_s^k) \\
&= \frac{C_1\sqrt{s}}{2}\left(\sqrt{Ls}(c_s^k + d_s^k) + (c_s^k - d_s^k)\right) + \frac{C_2}{2\sqrt{L}}\left((c_s^k + d_s^k) + \sqrt{Ls}(c_s^k - d_s^k)\right) \\
&= \frac{C_1\sqrt{s}}{2}\left(c_s^k(1 + \sqrt{Ls}) - d_s^k(1 - \sqrt{Ls})\right) + \frac{C_2}{2\sqrt{L}}\left(c_s^k(1 + \sqrt{Ls}) + d_s^k(1 - \sqrt{Ls})\right) \\
&= \frac{C_1\sqrt{s}}{2}(c_s^{k+1} - d_s^{k+1}) + \frac{C_2}{2\sqrt{L}}(c_s^{k+1} + d_s^{k+1}).
\end{aligned}$$

■

Observação. *Também poderíamos provar de maneira análoga que*

$$A_s c_s^k - B_s d_s^k = \frac{C_1\sqrt{s}}{2}(c_s^{k+1} + d_s^{k+1}) + \frac{C_2}{2\sqrt{L}}(c_s^{k+1} - d_s^{k+1}).$$

Fato B.8. *Para qualquer $s > 0$, vale que*

$$\frac{A_s}{c_s} + \frac{B_s}{d_s} = \frac{C_2}{\sqrt{L}}.$$

Demonstração. Basta seguir o seguinte desenvolvimento.

$$\begin{aligned}
\frac{A_s}{c_s} + \frac{B_s}{d_s} &= \frac{A_s d_s + B_s c_s}{c_s d_s} = \frac{A_s d_s + B_s c_s}{1 - Ls} \\
&= \frac{(a_s + b_s)d_s + (a_s - b_s)c_s}{1 - Ls} = \frac{a_s(d_s + c_s) + b_s(d_s - c_s)}{1 - Ls} = \frac{2a_s - 2\sqrt{Ls}b_s}{1 - Ls} \\
&= \frac{1}{1 - Ls}\left(2\frac{C_1 Ls + C_2}{2\sqrt{L}} - 2\sqrt{Ls}\frac{C_1 + C_2}{2}\sqrt{s}\right) \\
&= \frac{1}{1 - Ls}\left(\frac{C_1 Ls + C_2}{\sqrt{L}} - Ls\frac{C_1 + C_2}{\sqrt{L}}\right) \\
&= \frac{1}{1 - Ls}\frac{1}{\sqrt{L}}\left(C_1 Ls + C_2 - Ls(C_1 + C_2)\right) \\
&= \frac{1}{\sqrt{L}}\left(\frac{C_1 Ls + C_2 - LsC_1 - LsC_2}{1 - Ls}\right) \\
&= \frac{1}{\sqrt{L}}\left(\frac{C_2 - LsC_2}{1 - Ls}\right) = \frac{1}{\sqrt{L}}\left(\frac{(1 - Ls)C_2}{1 - Ls}\right) = \frac{C_2}{\sqrt{L}}.
\end{aligned}$$

■

B.2.4 Funções não descrescentes

Nesta subseção, demnstramos que duas funções são não decrescentes. Lembremo-nos de que $c_s = 1 + \sqrt{Ls}$ e $d_s = 1 - \sqrt{Ls}$.

Fato B.9. $\forall s \in (0, 1/L),\ f(t) = c_s^t + d_s^t$ *é cresente em* t *para* $t \geq 1$.

Demonstração. Como $s < 1/L$, c_s e d_s são maiores do que 0. Consideremos a função $f(t) = c_s^t + d_s^t$ e suas derivadas:

$$f'(t) = \ln(c_s) \cdot c_s^t + \ln(d) \cdot d_s^t,$$
$$f''(t) = \ln^2(c_s) \cdot c_s^t + \ln^2(d_s) \cdot d_s^t \geq 0.$$

Para mostrar que f é crescente em $t \geq 1$, vamos provar que $f'(t) > 0$ para $t \geq 1$. Como $f''(t) \geq 0$ para todo $t \geq 0$, a derivada $f'(t)$ é não-decrescente. Portanto, se mostrarmos que $f'(1) > 0$, seguirá que $f'(t) \geq f'(1) > 0$ para todo $t \geq 1$.

Definimos a função auxiliar

$$\begin{aligned} g(s) = f'(1) &= \ln(c_s)c_s + \ln(d_s)d_s \\ &= \ln(1 + \sqrt{Ls})(1 + \sqrt{Ls}) + \ln(1 - \sqrt{Ls})(1 - \sqrt{Ls}). \end{aligned}$$

Queremos provar que $g(s) \geq 0$ para todo $s \in (0, 1/L)$. Para isso vamos mostrar que $g(0) > 0$ e que $g'(s) \leq 0$.

Observemos que em $s = 0$, $g(0) = \ln(1) \cdot 1 + \ln(1) \cdot 1 = 0$ e que

$$1 + \sqrt{Ls} > 1 - \sqrt{Ls} \Longrightarrow \ln(1 + \sqrt{Ls}) > \ln(1 - \sqrt{Ls}) \Longrightarrow \ln(1 + \sqrt{Ls}) - \ln(1 - \sqrt{Ls}) > 0.$$

Agora, para $s \in (0, 1/L)$, temos

$$\begin{aligned} g'(s) &= \frac{\mathrm{d}}{\mathrm{d}s}\left(\ln(1 + \sqrt{Ls})(1 + \sqrt{Ls}) + \ln(1 - \sqrt{Ls})(1 - \sqrt{Ls})\right) \\ &= \frac{L}{2\sqrt{Ls}}\left(\frac{1 + \sqrt{Ls}}{1 + \sqrt{Ls}} + \ln(1 + \sqrt{Ls}) - \frac{1 - \sqrt{Ls}}{1 - \sqrt{Ls}} - \ln(1 - \sqrt{Ls})\right) \\ &= \frac{L}{2\sqrt{Ls}}\left(1 + \ln(1 + \sqrt{Ls}) - 1 - \ln(1 - \sqrt{Ls})\right) \\ &= \frac{L}{2\sqrt{Ls}}\left(\ln(1 + \sqrt{Ls}) - \ln(1 - \sqrt{Ls})\right) > 0\ . \end{aligned}$$

Assim, mostramos que $g(s) \geq 0$ para todo $s \in (0, 1/L)$, o que conclui nossa demonstração. ■

Fato B.10. $\forall s \in (0, 1/L),\ f(x) = c_s^t - d_s^t$ *é cresente em* t *para* $t \geq 1$.

Demonstração. Como $s < 1/L$, c_s e d_s são maiores do que 0. Além disso, $c_s > 1$ e $d_s < 1$. Isso significa que c_s^t é não decrescente em t e d_s^t não crescente (o que é o mesmo que dizer que $-d_s^t$ é não decrescenteem t). Assim $c_s^t - d_s^t$ é a soma de duas funções não decrescentes, logo, não decrescente também. ■

B.2.5 Recorrência

Nesta subseção vamos provar que se R_k é definida pela recorrência

$$R_k = 2R_{k-1} - c_s d_s R_{k-2} + C_2 s/\delta$$

com condições iniciais $R_{k^*_{\delta,s}} = S_{k^*_{\delta,s}}$ e $R_{k^*_{\delta,s}+1} = R_{k^*_{\delta,s}} + v_{k^*_{\delta,s}+1}$, então para todo $k \in [k^*_{\delta,s}, K^*_\delta]$, vale

$$R_k = \left(\frac{S_{k^*_{\delta,s}}}{2} + \frac{C_2}{2\delta L}\right)(c_s^{k-k^*_{\delta,s}} + d_s^{k-k^*_{\delta,s}}) + \frac{v_{k^*_{\delta,s}+1}}{2\sqrt{Ls}}(c_s^{k-k^*_{\delta,s}} - d_s^{k-k^*_{\delta,s}}) - \frac{C_2}{\delta L},$$

onde $c_s = 1 + \sqrt{Ls}$ e $d_s = 1 - \sqrt{Ls}$. Provaremos isso por indução fraca em k.

Para os casos bases ($i = k^*_{\delta,s}$ e $i = k^*_{\delta,s} + 1$), temos

$$\begin{aligned}
R_{k^*_{\delta,s}} &= \left(\frac{S_{k^*_{\delta,s}}}{2} + \frac{C_2}{2\delta L}\right)(c_s^{k^*_{\delta,s}-k^*_{\delta,s}} + d_s^{k^*_{\delta,s}-k^*_{\delta,s}}) + \frac{v_{k^*_{\delta,s}+1}}{2\sqrt{Ls}}(c_s^{k^*_{\delta,s}-k^*_{\delta,s}} - d_s^{k^*_{\delta,s}-k^*_{\delta,s}}) - \frac{C_2}{\delta L} \\
&= \left(\frac{S_{k^*_{\delta,s}}}{2} + \frac{C_2}{2\delta L}\right)(c_s^0 + d_s^0) + \frac{v_{k^*_{\delta,s}+1}}{2\sqrt{Ls}}(c_s^0 - d_s^0) - \frac{C_2}{\delta L} \\
&= \left(\frac{S_{k^*_{\delta,s}}}{2} + \frac{C_2}{2\delta L}\right)2 - \frac{C_2}{\delta L} = S_{k^*_{\delta,s}} + \frac{C_2}{\delta L} - \frac{C_2}{\delta L} = S_{k^*_{\delta,s}}.
\end{aligned}$$

e

$$\begin{aligned}
R_{k^*_{\delta,s}+1} &= \left(\frac{S_{k^*_{\delta,s}}}{2} + \frac{C_2}{2\delta L}\right)(c_s^{k^*_{\delta,s}+1-k^*_{\delta,s}} + d_s^{k^*_{\delta,s}+1-k^*_{\delta,s}}) + \frac{v_{k^*_{\delta,s}+1}}{2\sqrt{Ls}}(c_s^{k^*_{\delta,s}+1-k^*_{\delta,s}} - d_s^{k^*_{\delta,s}+1-k^*_{\delta,s}}) - \frac{C_2}{\delta L} \\
&= \left(\frac{S_{k^*_{\delta,s}}}{2} + \frac{C_2}{2\delta L}\right)(c_s^1 + d_s^1) + \frac{v_{k^*_{\delta,s}+1}}{2\sqrt{Ls}}(c_s^1 - d_s^1) - \frac{C_2}{\delta L} \\
&= \left(\frac{S_{k^*_{\delta,s}}}{2} + \frac{C_2}{2\delta L}\right)2 + \frac{v_{k^*_{\delta,s}+1}}{2\sqrt{Ls}}(2\sqrt{Ls}) - \frac{C_2}{\delta L} = S_{k^*_{\delta,s}} + \frac{C_2}{\delta L} + v_{k^*_{\delta,s}+1} - \frac{C_2}{\delta L} = S_{k^*_{\delta,s}} + v_{k^*_{\delta,s}+1}.
\end{aligned}$$

Agora, para o passo indutivo, suponhamos que valha a igualdade para $i = k^*_{\delta,s} - 2$ e $i = k^*_{\delta,s} - 1$. Então para $i = k^*_{\delta,s}$, temos

$$\begin{aligned}
R_i &= 2R_{i-1} - c_s d_s R_{i-2} + C_2 s/\delta \\
&= 2\left(\left(\frac{S_{k^*_{\delta,s}}}{2} + \frac{C_2}{2\delta L}\right)(c_s^{i-1-k^*_{\delta,s}} + d_s^{i-1-k^*_{\delta,s}}) + \frac{v_{k^*_{\delta,s}+1}}{2\sqrt{Ls}}(c_s^{i-1-k^*_{\delta,s}} - d_s^{i-1-k^*_{\delta,s}}) - \frac{C_2}{\delta L}\right) \\
&\quad - c_s d_s\left(\left(\frac{S_{k^*_{\delta,s}}}{2} + \frac{C_2}{2\delta L}\right)(c_s^{i-2-k^*_{\delta,s}} + d_s^{i-2-k^*_{\delta,s}}) + \frac{v_{k^*_{\delta,s}+1}}{2\sqrt{Ls}}(c_s^{i-2-k^*_{\delta,s}} - d_s^{i-2-k^*_{\delta,s}}) - \frac{C_2}{\delta L}\right) \\
&\quad + C_2 s/\delta \\
&= \left(\frac{S_{k^*_{\delta,s}}}{2} + \frac{C_2}{2\delta L}\right)(2c_s^{i-1-k^*_{\delta,s}} + 2d_s^{i-1-k^*_{\delta,s}} - c_s d_s c_s^{i-2-k^*_{\delta,s}} - c_s d_s d_s^{i-2-k^*_{\delta,s}}) \\
&\quad + \frac{v_{k^*_{\delta,s}+1}}{2\sqrt{Ls}}(2c_s^{i-1-k^*_{\delta,s}} - 2d_s^{i-1-k^*_{\delta,s}} - c_s d_s c_s^{i-2-k^*_{\delta,s}} + c_s d_s d_s^{i-2-k^*_{\delta,s}}) \\
&\quad + \frac{C_2}{\delta}\left(s - \frac{2}{L} + \frac{c_s d_s}{L}\right).
\end{aligned}$$

Mas nós temos que

$$\begin{aligned}
2c_s^{i-1-k^*_{\delta,s}} + 2d_s^{i-1-k^*_{\delta,s}} - c_s d_s c_s^{i-2-k^*_{\delta,s}} - c_s d_s d_s^{i-2-k^*_{\delta,s}} &= 2c_s^{i-1-k^*_{\delta,s}} + 2d_s^{i-1-k^*_{\delta,s}} - d_s c_s^{i-1-k^*_{\delta,s}} - c_s d_s^{i-1-k^*_{\delta,s}} \\
&= (2 - d_s)c_s^{i-1-k^*_{\delta,s}} + (2 - c_s)d_s^{i-1-k^*_{\delta,s}} \\
&= c_s c_s^{i-1-k^*_{\delta,s}} + d_s d_s^{i-1-k^*_{\delta,s}} = c_s^{i-k^*_{\delta,s}} + d_s^{i-k^*_{\delta,s}}
\end{aligned}$$

e que

$$\begin{aligned}2c_s^{i-1-k^*_{\delta,s}} - 2d_s^{i-1-k^*_{\delta,s}} - c_s d_s c_s^{i-2-k^*_{\delta,s}} + c_s d_s d_s^{i-2-k^*_{\delta,s}} &= 2c_s^{i-1-k^*_{\delta,s}} - 2d_s^{i-1-k^*_{\delta,s}} - d_s c_s^{i-1-k^*_{\delta,s}} + c_s d_s^{i-1-k^*_{\delta,s}} \\ &= (2-d_s)c_s^{i-1-k^*_{\delta,s}} - (2-c_s)d_s^{i-1-k^*_{\delta,s}} \\ &= c_s c_s^{i-1-k^*_{\delta,s}} - d_s d_s^{i-1-k^*_{\delta,s}} = c_s^{i-k^*_{\delta,s}} - d_s^{i-k^*_{\delta,s}}.\end{aligned}$$

Alem disso, temos

$$s - \frac{2}{L} + \frac{c_s d_s}{L} = \frac{Ls}{L} - \frac{2}{L} + \frac{1-Ls}{L} = \frac{Ls-2+1-Ls}{L} = -\frac{1}{L}.$$

Juntando tudo isso, obtemos enfim,

$$R_i = \left(\frac{S_{k^*_{\delta,s}}}{2} + \frac{C_2}{2\delta L}\right)(c_s^{i-k^*_{\delta,s}} + d_s^{i-k^*_{\delta,s}}) + \frac{vk^*_{\delta,s}+1}{2\sqrt{Ls}}(c_s^{i-k^*_{\delta,s}} - d_s^{i-k^*_{\delta,s}}) - \frac{C_2}{\delta L}.$$

B.2.6 Limites

Nesta última subseção, calculamos alguns limites usados na demonstração da convergência do AMD para a solução de sua EDO associada.

Para todos $\alpha, \beta \in \mathbb{R}$ fixos, temos

1) $\lim\limits_{s\to 0^+} a_s = \lim\limits_{s\to 0^+} \dfrac{C_1 Ls + C_2}{2\sqrt{L}} = \dfrac{C_2}{2\sqrt{L}}$;

2) $\lim\limits_{s\to 0^+} b_s = \lim\limits_{s\to 0^+} \dfrac{C_1 + C_2}{2}\sqrt{s} = 0$;

3) $\lim\limits_{s\to 0^+} c_s^\beta = \lim\limits_{s\to 0^+} (1+\sqrt{Ls})^\beta = 1$;

4) $\lim\limits_{s\to 0^+} d_s^\beta = \lim\limits_{s\to 0^+} (1-\sqrt{Ls})^\beta = 1$;

5) $\lim\limits_{s\to 0^+} c_s^{\alpha/\sqrt{s}+\beta} = \lim\limits_{s\to 0^+} c_s^\beta (1+\sqrt{Ls})^{\alpha/\sqrt{s}} = \lim\limits_{s\to 0^+} c_s^\beta \lim\limits_{s\to 0^+} (1+\sqrt{Ls})^{\sqrt{L}\alpha/\sqrt{Ls}} = e^{\sqrt{L}\delta}$;

6) $\lim\limits_{s\to 0^+} d_s^{\alpha/\sqrt{s}+\beta} = \lim\limits_{s\to 0^+} d_s^\beta (1-\sqrt{Ls})^{\alpha/\sqrt{s}} = \lim\limits_{s\to 0^+} d_s^\beta \lim\limits_{s\to 0^+} (1-\sqrt{Ls})^{-\sqrt{L}\alpha/-\sqrt{Ls}} = e^{-\sqrt{L}\delta}$;

7) $\lim\limits_{\delta\to 0^+} e^{\sqrt{L}\delta} + e^{-\sqrt{L}\delta} = e^0 + e^0 = 2$;

8) $\lim\limits_{\delta\to 0^+} e^{\sqrt{L}\delta} - e^{-\sqrt{L}\delta} = e^0 - e^0 = 0$.

Agora, como $c_s^t + d_s^t$ é não decrescente em t e usando o limite 5) acima com $\alpha = T-\delta$ e $\beta = 1$, temos

$$\begin{aligned}\lim_{s\to 0^+}\left(c_s^{K^*_s - k^*_{\delta,s}} + d_s^{K^*_s - k^*_{\delta,s}}\right) &\le \lim_{s\to 0^+}\left(c_s^{(T-\delta)/\sqrt{s}+1} + d_s^{(T-\delta)/\sqrt{s}+1}\right) \\ &= \lim_{s\to 0^+} c_s^{(T-\delta)/\sqrt{s}+1} + \lim_{s\to 0^+} d_s^{(T-\delta)/\sqrt{s}+1} = e^{(T-\delta)\sqrt{L}} + e^{-(T-\delta)\sqrt{L}}.\end{aligned}$$

Da mesma forma, agora com $\alpha = \delta$ e $\beta = 0$, temos

$$\lim_{s\to 0^+}\left(c_s^{k^*_{\delta,s}+1} + d_s^{k^*_{\delta,s}+1}\right) \le \lim_{s\to 0^+}\left(c_s^{\delta/\sqrt{s}+1} + d_s^{\delta/\sqrt{s}+1}\right) = \lim_{s\to 0^+} c_s^{\delta/\sqrt{s}+1} + \lim_{s\to 0^+} d_s^{\delta/\sqrt{s}+1} = e^{\sqrt{L}\delta} + e^{-\sqrt{L}\delta}.$$

Por fim, como $c_s^t - d_s^t$ também é não decrescente em t e usando de novo o limite 6) com $\alpha = \delta$ e $\beta = 1$, temos

$$\lim_{s\to 0^+}\left(c_s^{k^*_{\delta,s}+1} - d_s^{k^*_{\delta,s}+1}\right) \le \lim_{s\to 0^+}\left(c_s^{\delta/\sqrt{s}+1} - d_s^{\delta/\sqrt{s}+1}\right) \le \lim_{s\to 0^+} c_s^{\delta/\sqrt{s}+1} - \lim_{s\to 0^+} d_s^{\delta/\sqrt{s}+1} = e^{\sqrt{L}\delta} - e^{-\sqrt{L}\delta}.$$

B.3 Sobre a discretização da EDO associada ao AMD

Neste capítulo, derivamos o AMD (6.1) a partir das equações que definem a EDO 6.1, mostrando especificamente a qual discretização o método corresponde.

Como visto na subseção 7.1.1, a EDO 6.1 pode ser reescrita como

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\Theta(X(t))\\ Z(t)\end{pmatrix} = \begin{pmatrix}Z(t)\\ -\frac{3}{t}Z(t) - \nabla f(X(t))\end{pmatrix}. \tag{B.8}$$

Aplicando o método de Euler progressivo com passo $\sqrt{s}$ à primeira equação desta nova formulação, obtemos

$$\frac{\theta_{k+1} - \theta_k}{\sqrt{s}} = z_k \quad \Longrightarrow \quad \theta_{k+1} = \theta_k + \sqrt{s}z_k,$$

o que corresponde à segunda equação da formulação 7.9 para o AMD.

Agora, aplicando o método de Euler progressivo à segunda equação da EDO e substituindo t por $(k+2)\sqrt{s}$, obtemos

$$\begin{aligned}\frac{z_{k+1} - z_k}{\sqrt{s}} = -\frac{3}{(k+2)\sqrt{s}}z_k - \nabla f(x_k) \quad &\Longrightarrow \quad z_{k+1} = \left(1 - \frac{3}{k+2}\right)z_k - \nabla f(x_k)\\ &\Longrightarrow \quad z_{k+1} = \mu_k z_k - \nabla f(x_k).\end{aligned}$$

De maneira análoga, aplicando o método de Euler **regressivo** à mesma equação, obteríamos

$$z_{k+1} = \mu_k z_k - \nabla f(x_{k+1}).$$

Agora, consideremos a combinação convexa de θ_k e θ_{k+1} definida por $\theta' = \theta_k + \mu_k(\theta_{k+1} - \theta_k)$. Temos que θ' é igual a θ_k quando $k = 1$ e se aproxima de θ_{k+1} conforme k aumenta. Como $\theta_{k+1} = \theta_k + \sqrt{s}z_k$, podemos reescrever a combinação como

$$\theta' = \theta_k + \mu_k\sqrt{s}z_k.$$

Seja $y_k = (\nabla\Phi)^{-1}(\theta')$. Definimos uma discretização da segunda equação da EDO dada por

$$z_{k+1} = \mu_k z_k - \nabla f(y_k),$$

que chamamos de método de Euler **ponderado**. Essa discretização gera as duas últimas equações da formulação 7.9 para o AMD. Com isso, concluímos que o AMD corresponde a uma discretização, ainda que não tão óbvia, da equação encontrada.

Apêndice C

Experimento Numérico

Neste capítulo, trazemos figuras complementares às produzidas pelo experimento numérico. Abaixo, as Figura C.1 e Figura C.2 mostram os dados apresentados no Capítulo 8 em escala log.

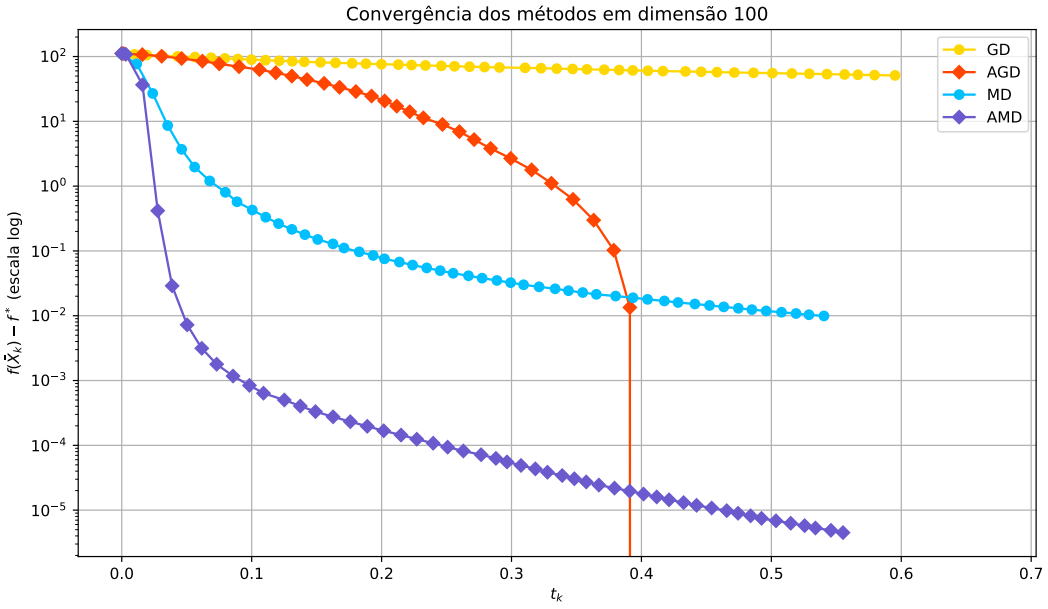

Figura C.1: Comparação das convergências dos métodos GD, AGD, MD e AMD para um problema em dimensão 100.

Observamos que o erro do método AGD é exatamente zero a partir de certo ponto, isto é, que ele atinge o mínimo da função objetivo. Isso ocorre porque, nas configurações do problema, a solução é um dos vértices do simplexo e, com a projeção, o AGD acaba tendo uma vantagem, já que pode só se aproximar da direção correta do vértice ótimo. Em outras palavras, basta que o método dê um passo suficientemente grande na direção indicada pelo gradiente — que já aponta para fora do simplexo no vértice correto — para que a projeção o leve diretamente a esse vértice. Assim, o algoritmo não precisa convergir de forma gradual como em regiões interiores: uma vez que o movimento extrapola o simplexo na direção certa, a projeção realiza o "salto" final até a solução exata.

Abaixo, as Figura C.3 e Figura C.4 mostram as diferenças de valores (em escala log) entre as **médias** dos pontos gerados e o valor ótimo da função em dimensões 100 e 10.000, o que está mais relacionado com o resultado 8.1, assim como os limites superiores para as diferenças do GD e do MD (Bound GD e Bound MD). A ideia é comparar o resultado obtido pelos métodos na última iteração (com $k = T = 200$) e o limite superior previsto pelo resultado $2RL\sqrt{\frac{2}{\nu T}}$, onde, lembrando, R e L são os mesmos do Capítulo 8.

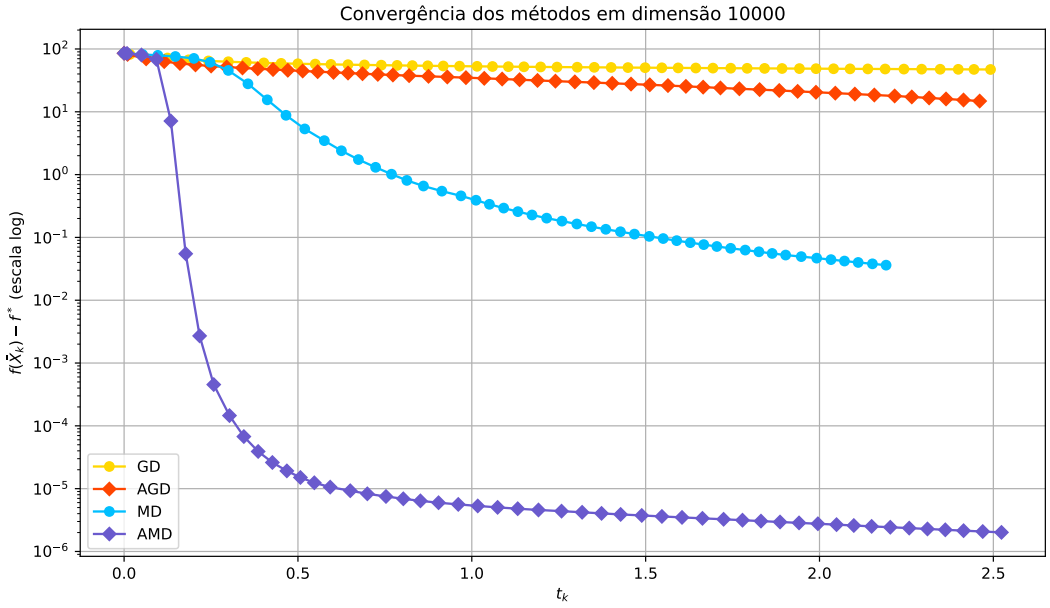

Figura C.2: Comparação das convergências dos métodos GD, AGD, MD e AMD para um problema em dimensão 10000.

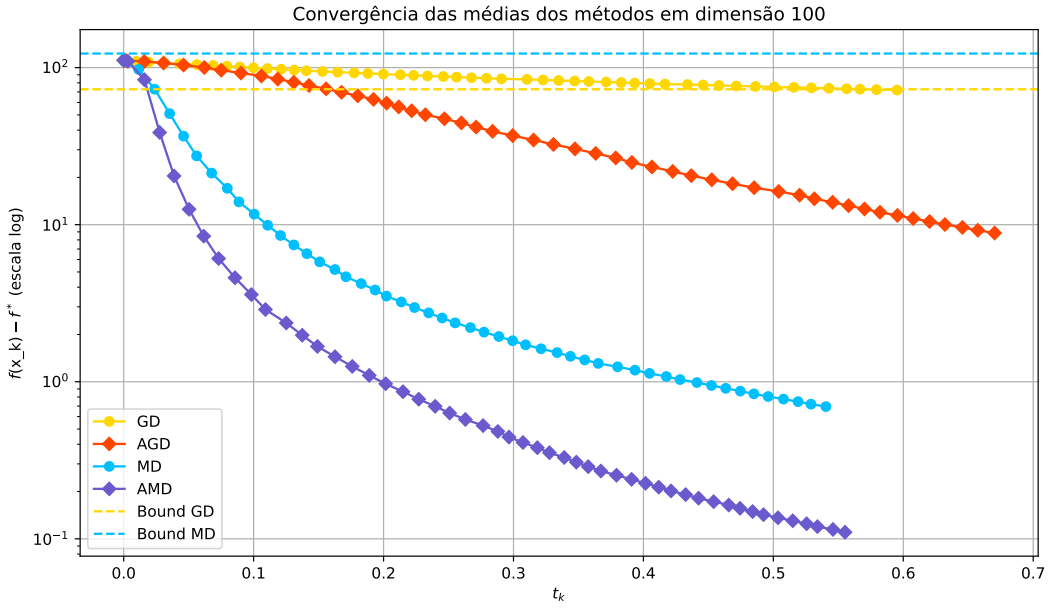

Figura C.3: Comparação das convergências das médias dos métodos GD, AGD, MD e AMD para um problema em dimensão 100.

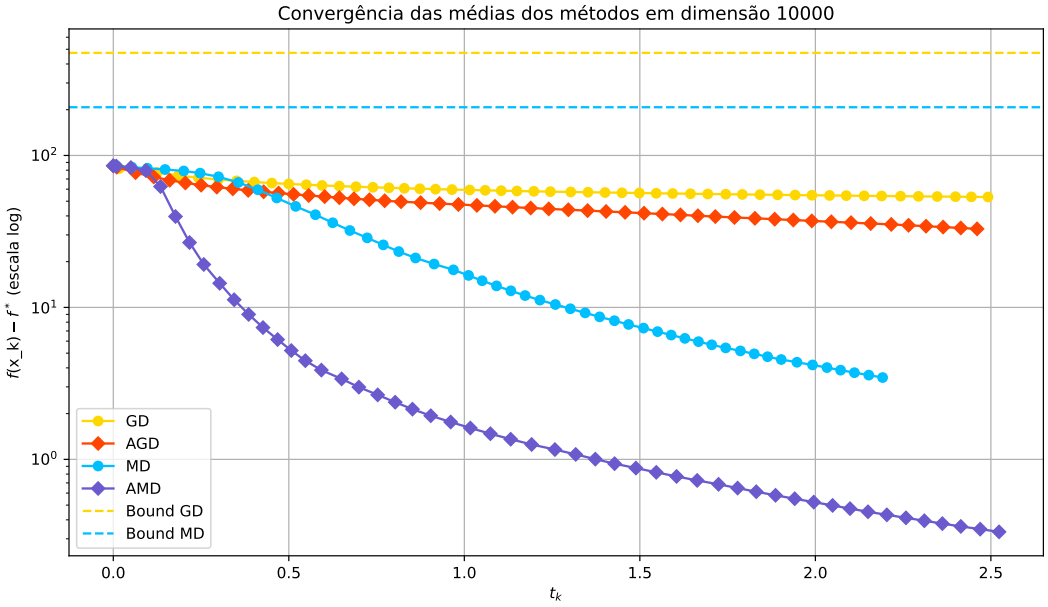

Figura C.4: Comparação das convergências das médias dos métodos GD, AGD, MD e AMD para um problema em dimensão 100.

Os limites no caso de dimensão 10.000 ficaram grandes, até maiores do que a diferença inicial, o que acaba fazendo com que o resultado pareça trivial. Isso ocorre porque o parâmetro R cresce com a dimensão e os parâmetros L_2 e L_∞ são (ainda que indiretamente) gerados aleatoriamente. Aumentando o número de iterações, podemos diminuir esses limites, mas isso tem um custo computacional.

Apêndice D

Taxas de convergência empíricas

Neste capítulo apresentamos as taxas de convergência encontradas empiricamente para cada método, baseadas no experimento numérico domCapítulo 8. Para cada método, apresentamos a curva de convergência multiplicada por fatores como k ou k^2, onde k é a iteração de cada ponto. Dessa forma, consideramos que um método tem taxa $O(1/k)$, por exemplo, quando sua curva multiplicada por k é limitada e mais ou menos constante (não apresentando crescimento ou decrescimento estável).

Seguindo a ordem de apresentação dos métodos, comecemos pelo GD. Aqui, mantivemos a convergência do método sem multiplicá-la por nenhum fator, observando em Figura D.1 que esta parece ser linear.

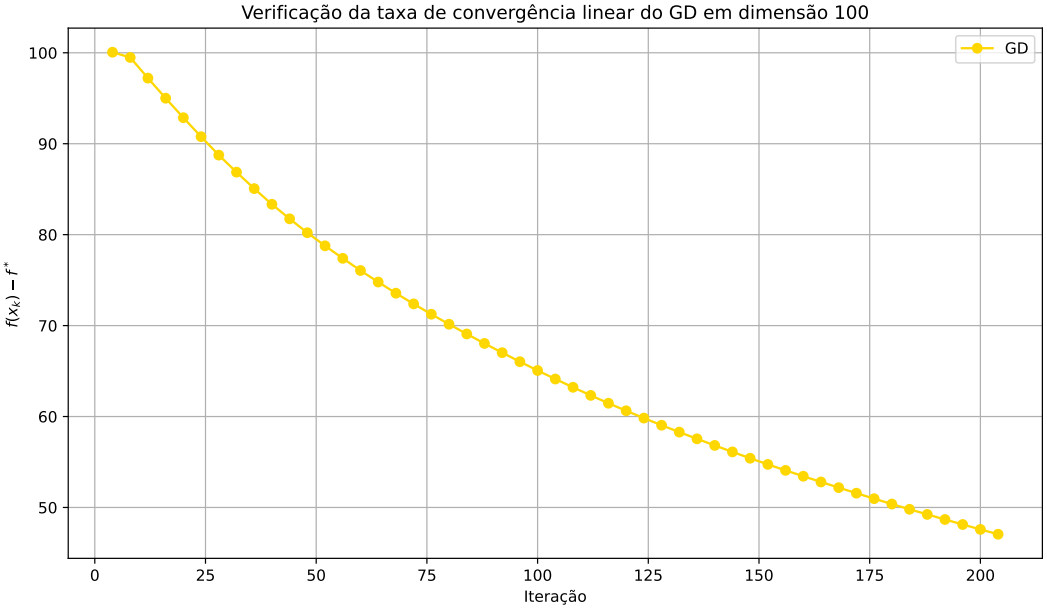

Figura D.1: Curva mostrando aparente taxa de convergência $O(1/k)$ para o GD.

Multiplicamos o AGD por k^2, com sua taxa teórica $O(1/k^2)$ em mente. O resultado se encontra na Figura D.2. Aqui existe uma oscilação no começo, e depois a curva fica igual a zero, de forma que a taxa de convergência do método não fique muito clara, já que não é assintótica.

Começamos multiplicando a convergência do MD por k, mas percebemos que havia um decrescimento. Multiplicando por k^2, obtemos a curva em Figura D.3 que, da mesma maneira que o AGD, apresenta uma oscilação no começo, mas logo estabiliza. Desta maneira, a taxa empírica para o MD neste caso foi de $O(1/K^2)$. Isso já podia ser esperado, tendo observado o desempenho dos AGD e MD no Capítulo 8.

Por fim, como foi com o MD, começamos multiplicando a taxa do AMD por um fator, mas pudemos aumentá-lo após observar um decréscimo na curva resultante. A Figura D.4 mostra a curva obtida ao

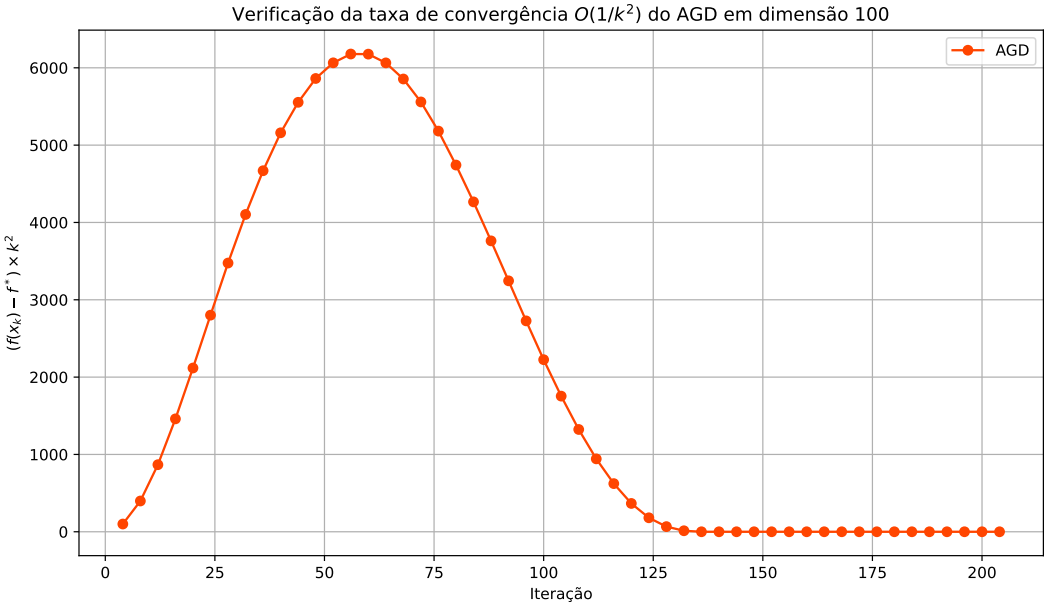

Figura D.2: Curva mostrando aparente taxa de convergência $O(1/k^2)$ para o AGD.

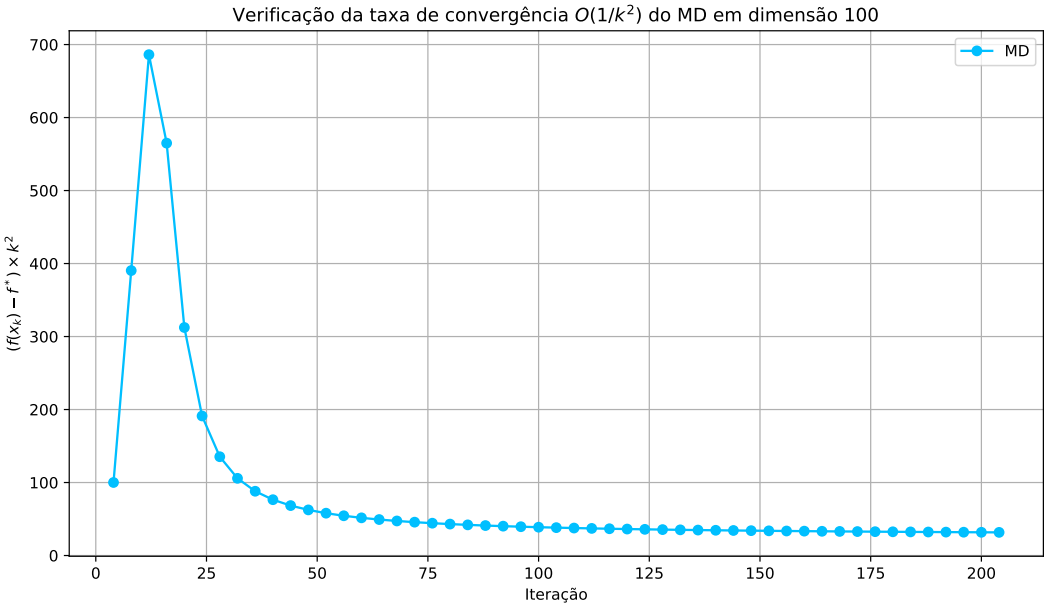

Figura D.3: Curva mostrando aparente taxa de convergência $O(1/k^2)$ para o MD.

multiplicarmos a convergência do método por k^4 e a Figura D.5 mostra a mesma curva para iterações maiores ou iguais a 25 (para melhor visualização). A curva apresenta uma oscilação no começo, depois um período de crescimento e, por fim, uma estabilização. A taxa empírica do AMD fica, portanto, entre $O(1/k^3)$ e $O(1/k^4)$.

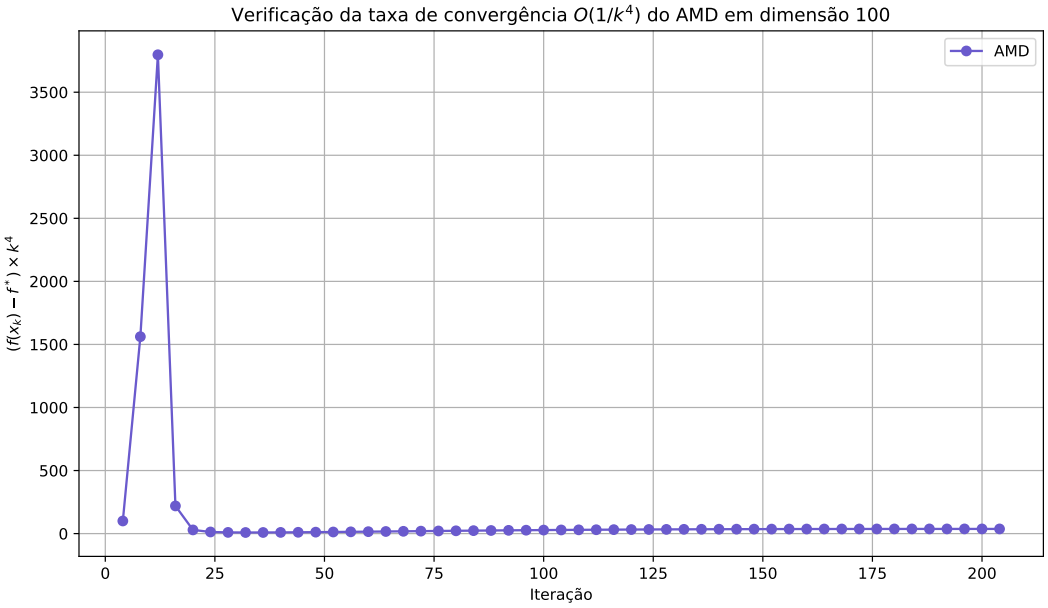

Figura D.4: Curva mostrando aparente taxa de convergência $O(1/k^4)$ para o AMD.

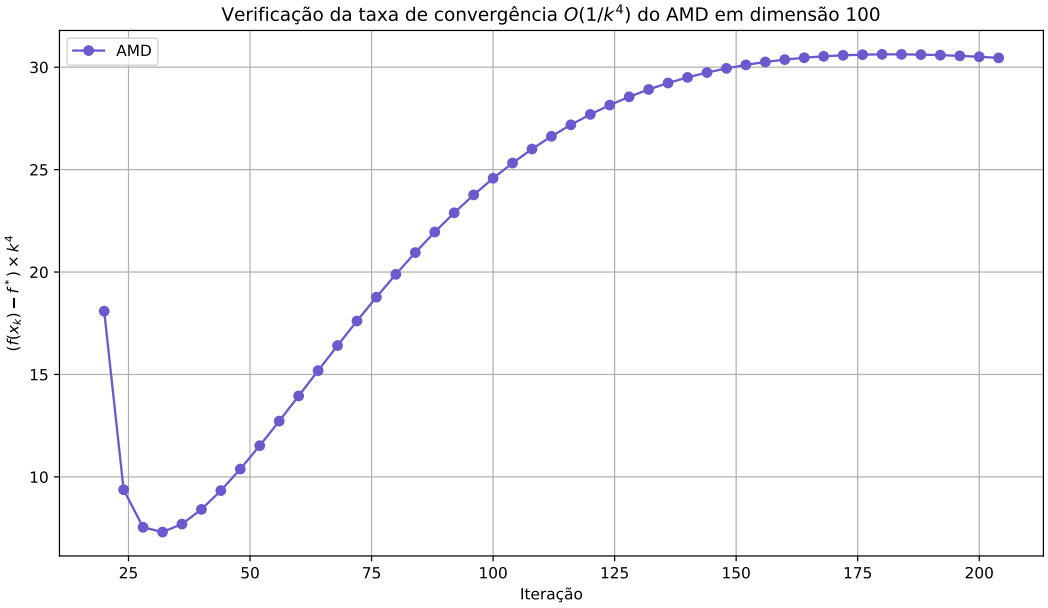

Figura D.5: Curva mostrando aparente taxa de convergência $O(1/k^4)$ para o AMD a partir da iteração 25.