Universidade Estadual do Sudoeste da Bahia
Departamento de Ciências Exatas e Tecnológicas

Licenciatura em Matemática

Maria Clara Brito dos Reis

Método de Newton Riemanniano para Diagonalização Conjunta de Matrizes Aplicado à Separação de Imagens

Vitória da Conquista
2025

Maria Clara Brito dos Reis

Método de Newton Riemanniano para Diagonalização Conjunta de Matrizes Aplicado à Separação de Imagens

Monografia apresentada ao Departamento de Ciências Exatas e Tecnológicas da Universidade Estadual do Sudoeste da Bahia - Campus Vitória da Conquista-BA, para obtenção do Título de Licenciada em Matemática, sob orientação do Prof. Marcio Antônio de Andrade Bortoloti.

Vitória da Conquista
2025

Maria Clara Brito dos Reis

Método de Newton Riemanniano para Diagonalização Conjunta de Matrizes Aplicado à Separação de Imagens

Monografia apresentada ao Colegiado do Curso de Matemática como requisito parcial para aprovação na disciplina Seminário de Pesquisa II do Curso de Licenciatura em Matemática.
Trabalho aprovado em 5 de dezembro de 2025.

BANCA EXAMINADORA

gov.br Documento assinado digitalmente
MARCIO ANTONIO DE ANDRADE BORTOLOTI
Data: 13/12/2025 09:09:55-0300
Verifique em https://validar.iti.gov.br

Prof. Marcio Antônio de Andrade Bortoloti - UESB
Orientador

gov.br Documento assinado digitalmente
FERNANDO DOS SANTOS SILVA
Data: 18/12/2025 19:34:26-0300
Verifique em https://validar.iti.gov.br

Prof. Fernando dos Santos Silva - UESB
Examinador

gov.br Documento assinado digitalmente
RICARDO FREIRE DA SILVA
Data: 18/12/2025 20:39:49-0300
Verifique em https://validar.iti.gov.br

Prof. Ricardo Freire da Silva - UESB
Examinador

Vitória da Conquista
2025

À minha família.

AGRADECIMENTOS

Dedico este trabalho à minha família, por todo cuidado, paciência, incentivo e suporte que me deram ao longo de toda a graduação. Em especial, à minha mãe, Katrine, pelo amor e carinho dedicados em meus momentos mais felizes e também nos mais difíceis. Você me inspira a buscar sempre o melhor de mim. Ao meu padrasto, Remilson, pelo cuidado, ensinamentos e presença constante ao longo desses anos. Ao meu irmão, João Lucas, por ser minha âncora e meu alívio diário. Espero poder te inspirar, um dia, a trilhar o caminho dos estudos com a mesma paixão que me move. Ao meu tio Júnior, pela disponibilidade e boa vontade em sempre me ajudar quando precisei. À minha avó Maria Lúcia, pelo amor que transborda e acolhe em todos os momentos. Vocês são o melhor de mim, e é a vocês que dedico este trabalho, com todo o meu amor e gratidão.

Agradeço profundamente ao meu orientador, professor Marcio Antônio de Andrade Bortoloti, pela generosidade em compartilhar seu tempo, conhecimento e paciência durante o desenvolvimento não apenas deste trabalho, mas também ao longo de todo o período de iniciação científica. Minha formação e minhas perspectivas não seriam as mesmas sem sua orientação. Deixo aqui meus sinceros agradecimentos.

Aos meus amigos Alison, Franciele, Luane, Malu e Wéllington, obrigada por cada risada, conversa e apoio ao longo dessa jornada. O carinho e a amizade de vocês tornaram o caminho mais leve e cheio de boas lembranças. Esta conquista é compartilhada com cada um de vocês.

À equipe do Programa de Educação Tutorial Institucional em Matemática (PETIMAT), expresso minha profunda gratidão por todo o apoio e estímulo recebidos ao longo desses anos. Foi no PETIMAT que desenvolvi meu interesse por matemática aplicada, encontrei inspiração para expandir meus estudos e me aproximei de temas que se tornaram centrais neste trabalho, especialmente otimização, área que passei a explorar com entusiasmo graças ao incentivo e às oportunidades proporcionadas nesse programa. Espero que ao longo do tempo, ele possa transformar outros estudantes, como me transformou.

Ao Grupo de Estudos de Matemática Pura e Aplicada (GEMPA), expresso minha sincera gratidão por ter sido um espaço de aprendizado, troca e crescimento. As experiências vividas no grupo me proporcionaram amadurecimento acadêmico e pessoal, além de fortalecerem meu interesse pela pesquisa e pela docência. A convivência com colegas e professores fez do GEMPA uma parte essencial da minha trajetória, deixando lembranças e ensinamentos que levarei comigo.

Agradeço aos professores Ricardo Freire da Silva e Fernando dos Santos Silva, membros da banca examinadora, pelas críticas e sugestões que contribuíram para o aprimoramento deste trabalho.

Por fim, obrigada a todos que, de alguma forma, fizeram parte deste percurso.

RESUMO

Neste estudo, comparamos o desempenho do método de Newton Riemanniano clássico com sua versão amortecida, equipada com a busca linear de Armijo, na resolução do problema de diagonalização conjunta de matrizes sobre a variedade de Stiefel. A formulação proposta insere-se no contexto da otimização Riemanniana, em que as direções de Newton são determinadas a partir do gradiente e da Hessiana Riemannianos. Para viabilizar o cálculo dessas direções, adotou-se uma estratégia de vetorização para a obtenção da direção de Newton. Os experimentos numéricos envolveram a minimização de uma função objetivo associada à diagonalização conjunta de matrizes simétricas, considerando diferentes pontos iniciais para a geração da sequência, e utilizando como métrica principal de comparação o número de iterações necessárias para a convergência. Os resultados mostraram que o método de Newton Riemanniano amortecido apresenta potencial para resolver problemas quando o ponto inicial está mais distante da solução, além de promover uma redução mais estável da norma do gradiente. Paralelamente, o método de Newton clássico tende a convergir mais rapidamente quando o ponto inicial é favorável. Na aplicação ao problema de separação de imagens sobrepostas, cuja formulação corresponde à diagonalização conjunta envolvendo matrizes que contêm informações das sobreposições, o método amortecido apresentou desempenho superior, alcançando soluções satisfatórias com menor número de iterações em comparação ao método clássico, embora ambos tenham retornado soluções satisfatórias.

Palavras-chave: Método de Newton Riemanniano; Busca de Armijo; Diagonalização conjunta de matrizes.

ABSTRACT

In this study, we compare the performance of the classical Riemannian Newton method with its damped variant, equipped with the Armijo line search, in solving the joint diagonalization problem on the Stiefel manifold. The proposed formulation lies within the framework of Riemannian optimization, where the Newton directions are computed based on the Riemannian gradient and Hessian. To make the computation of these directions feasible, a vectorization strategy was employed for obtaining the Newton direction. The numerical experiments involved minimizing an objective function associated with the joint diagonalization of symmetric matrices, considering different initial points for generating the iterative sequence, and using the number of iterations required for convergence as the main performance metric. The results showed that the damped Riemannian Newton method exhibits greater potential for solving problems when the initial guess is far from the solution, as well as providing a more stable reduction of the gradient norm. In contrast, the classical Newton method tends to converge faster when the initial guess is favorable. When applied to the image separation problem, formulated as a joint diagonalization involving matrices that encode information from the superimposed images, the damped method achieved superior performance, reaching satisfactory solutions with fewer iterations compared to the classical approach, although both returned satisfactory solutions.

Keywords: Riemannian Newton method; Armijo line search; Joint matrix diagonalization.

Conteúdo

Introdução **6**

1 Introdução à Variedades **9**
1.1 Definições preliminares 9
1.2 Variedade Diferenciável 10
1.3 Espaço tangente 15
1.4 Variedade Riemanniana 17

2 Método de Newton Riemanniano **19**
2.1 Método de Newton Euclidiano 19
2.2 Otimização em Variedades 26
2.2.1 Condições de otimalidade em Variedades 28
2.3 Método de Newton Riemanniano 30
2.3.1 Gradiente e Hessiana Riemannianos 30

3 Problema de Diagonalização conjunta de matrizes **38**
3.1 Introdução 38
3.2 Estratégia de vetorização 48
3.2.1 Produto de Kronecker e operadores vec e veck 50
3.2.2 Equação de Newton 55

4 Método de Newton Riemanniano Amortecido **61**
4.1 Busca de Armijo 61
4.2 Experimentos numéricos 64

5 Separação de imagens sobrepostas **70**
5.1 Análise de componente independente 70
5.2 Experimentos numéricos 74

Introdução

A ideia de otimizar está presente de forma natural em diferentes aspectos da vida cotidiana e em diversas áreas do conhecimento. Em termos gerais, otimizar significa buscar a melhor escolha entre várias possibilidades, seja na tomada de decisões, na administração de recursos limitados ou na busca de soluções mais eficientes. Cada um desses problemas quase sempre envolve o mesmo objetivo: *maximizar* ou *minimizar* uma certa função [1]. Em poucas palavras, podemos dizer que Otimização refere-se ao processo de determinar pontos de mínimo ou máximo de funções em um certo conjunto. Assim, dado um $\Omega \subseteq \mathbb{R}^n$ e uma função objetivo $f : \mathbb{R}^n \to \mathbb{R}$, o problema de minimizar f pode ser escrito como

$$\text{minimizar } f(x), \quad x \in \Omega.$$

Isto é, buscamos um ponto $x_* \in \Omega$ tal que $f(x_*) \leqslant f(x), \quad$ para todo $x \in \Omega$. Para atingir esse objetivo, geramos uma sequência (x_k) que deve convergir para a solução do problema x_*.

Uma das principais dificuldades nesse processo está nas restrições impostas pelo conjunto Ω, que podem tornar a geração da sequência de pontos computacionalmente custosa. Entretanto, quando Ω for uma *variedade diferenciável*, o problema restrito pode ser reformulado como um problema irrestrito sobre a própria variedade. Nesse contexto, a função objetivo é considerada diretamente sobre Ω, ou seja, $f : \Omega \to \mathbb{R}$, o que facilita a busca por soluções do problema minimizar f.

Neste trabalho, consideramos um problema de otimização cuja restrição é a *variedade de Stiefel*, denotada por $\mathrm{St}(p, n)$ e definida como

$$\mathrm{St}(p, n) = \{\, Y \in \mathbb{R}^{n \times p} \;;\; Y^T Y = I_p \,\},$$

onde I_p é a matriz identidade de ordem p e $p \leqslant n$. O problema de otimização, explorado neste trabalho, consiste em diagonalizar simultaneamente um conjunto de matrizes simétricas. Mais especificamente, dado um conjunto de matrizes simétricas $A_l \in \mathbb{R}^{n \times n}$, para $l = 1, \ldots, N$, buscamos encontrar uma matriz $Y \in \mathrm{St}(p, n)$ tal que

$$Y^T A_l Y \approx D_l, \quad l = 1, \ldots, N,$$

onde D_l é uma matriz diagonal. O objetivo é tornar cada matriz $Y^T A_l Y$ o mais próxima possível de D_l.

Para atingir tal objetivo, adotamos o Método de Newton no contexto de otimização em variedades diferenciáveis, munidas de uma *métrica Riemanniana*. Nesse caso, o gradiente e a Hessiana Riemannianas são utilizados para determinar uma direção de descida na atualização dos pontos da sequência. Especificamente, a direção de Newton é obtida como solução do sistema

$$\operatorname{Hess} f(Y)[\xi] = -\operatorname{grad} f(Y),$$

onde $\operatorname{grad} f(Y)$ e $\operatorname{Hess} f(Y)$ representam, respectivamente, o gradiente e a Hessiana da função objetivo em $Y \in \mathrm{St}(p, n)$.

Em particular, adotamos a estratégia de vetorização proposta por [12] para o cálculo da direção de Newton. No referido trabalho, [12] minimizou a mesma função objetivo utilizada neste estudo aplicando o método de Newton clássico, com sua estratégia de vetorização.

Neste estudo, propomos uma comparação entre o método adotado em [12] e o Método de Newton Riemanniano amortecido, na busca da solução do problema de diagonalização conjunta. Para isso, equipamos o método de Newton clássico com a busca de Armijo, muito popular entre as buscas lineares usadas em métodos de otimização.

A fim de fundamentar a discussão do método de Newton Riemanniano a ser conduzida ao longo deste trabalho e fornecer a base teórica necessária para o estudo do problema de diagonalização conjunta, o Capítulo 1 apresenta os conceitos fundamentais sobre variedades. São abordadas variedades topológicas e diferenciáveis, seguindo para variedades Riemannianas, que constituem o foco deste trabalho. Elementos essenciais, como cartas, espaço tangente e retrações, também são introduzidos, servindo de suporte para os desenvolvimentos e resultados apresentados nos capítulos seguintes.

No Capítulo 2, dedicamo-nos a apresentar o Método de Newton, inicialmente no contexto euclidiano, onde discutimos suas propriedades fundamentais, como convergência local e taxa de convergência. A partir desse entendimento, estendemos o método para o contexto das variedades Riemannianas, introduzindo o Método de Newton Riemanniano. Nessa versão, as direções de Newton são determinadas a partir do gradiente e da Hessiana Riemannianos, e os novos pontos são obtidos por meio de retrações, o que garante que os pontos da sequência gerada pelo método permaneçam na variedade.

No terceiro capítulo, apresentamos formalmente o problema de diagonalização conjunta como um problema de otimização em variedades. Desenvolvemos o cálculo do gradiente e da Hessiana Riemannianos, necessários para a determinação da direção de Newton associada ao problema. Durante esse processo, surgem limitações que dificultam o cálculo de tal direção; estas dificuldades são apontadas ao longo do capítulo e motivam o uso da estratégia de vetorização proposta em [12], a qual é reproduzida e detalhada neste

trabalho. Com base nessa formulação, apresentamos finalmente o algoritmo completo do método de Newton Riemanniano clássico proposto em [12].

Como mencionado anteriormente, o objetivo central deste trabalho é comparar o método de Newton Riemanniano clássico com sua versão equipada com a busca de Armijo, à qual nos referimos como método de Newton amortecido, aplicados à minimização da mesma função objetivo. No Capítulo 4, apresentamos a busca de Armijo, bem como a prova de convergência do método amortecido. Por fim, descrevemos e analisamos os resultados obtidos nos experimentos numéricos, destacando as diferenças de desempenho entre os dois métodos comparados.

No quinto e último capítulo, aplicamos os conceitos desenvolvidos nos capítulos anteriores a um problema de separação de imagens sobrepostas, interpretado como um caso particular de diagonalização conjunta de matrizes na variedade de Stiefel. O objetivo foi avaliar o desempenho do Método de Newton Riemanniano clássico e de sua versão equipada com a busca de Armijo, verificando sua capacidade de recuperar imagens originais a partir de sobreposições dessas imagens. Os resultados obtidos demonstram que ambos os métodos produzem soluções satisfatórias, mas o método amortecido apresenta vantagem em termos de número de iterações.

Capítulo 1

Introdução à Variedades

Com o objetivo de subsidiar a teoria apresentada nos capítulos seguintes, iniciamos com uma breve introdução à noção de Variedades Diferenciáveis, bem como aos conceitos essenciais que servirão de base para os resultados posteriores.

1.1 Definições preliminares

Intuitivamente, uma variedade é uma generalização de curvas e superfícies para dimensões mais altas, na qual é possível, em cada ponto, aproximar localmente uma vizinhança por um aberto do $\mathbb{R}^n$, veja [15, p.47]. Para introduzir formalmente esse conceito, partiremos da definição de variedade topológica. Antes, porém, recordemos algumas noções fundamentais de topologia de conjuntos, conforme apresentadas em [8].

Definição 1. Uma *topologia* num conjunto $\mathcal{M}$ é uma coleção τ de subconjuntos $\mathcal{M}$, chamados os *subconjuntos abertos* (segundo a topologia τ) satisfazendo as seguintes condições:

1. $\varnothing \in \tau$ e $\mathcal{M} \in \tau$;
2. $\bigcup_{i\in I} \mathcal{U}_i \in \tau$ para toda família $\{\mathcal{U}_i\}_{i\in I} \subseteq \tau$;
3. $\mathcal{U}_1 \cap \cdots \cap \mathcal{U}_n \in \tau$ para quaisquer $\mathcal{U}_1, \ldots, \mathcal{U}_n \in \tau$.

O par $(\mathcal{M}, \tau)$ é chamado de *espaço topológico*. Usualmente, denotaremos o espaço topológico $(\mathcal{M}, \tau)$, apenas por $\mathcal{M}$.

De forma análoga ao conceito de base de espaços vetoriais, também é possível definir uma base em topologia. Uma base para o espaço topológico $\mathcal{M}$ é uma coleção $\mathcal{B}$ de subconjuntos abertos de $\mathcal{M}$ tal que todo aberto $\mathcal{U} \subset \mathcal{M}$ exprime-se como reunião $\mathcal{U} = \cup B_\lambda$ de conjuntos $B_\lambda \in \mathcal{B}$. Em outras palavras, dados um aberto $\mathcal{U}$ em $\mathcal{M}$ e um ponto $Y \in \mathcal{U}$ existe um aberto $B \in \mathcal{B}$ com $Y \in B \subset \mathcal{U}$. Nosso interesse recai sobre bases enumeráveis, a qual definimos a seguir.

Definição 2. Um espaço topológico $\mathcal{M}$ possui uma *base enumerável* $\mathcal{B} = \{B_n; n \in \mathbb{N}\}$ para a topologia τ, se para todo $\mathcal{U} \in \tau$ e $Y \in \mathcal{U}$, existe $B_n \in \mathcal{B}$ tal que $Y \in B_n \subset \mathcal{U}$.

Para introduzir formalmente o conceito de variedade topológica, é necessário recorrer ainda às duas definições que se seguem.

Definição 3. Um espaço topológico $\mathcal{M}$ é dito *Hausdorff* se, para quaisquer dois pontos distintos $Y, Z \in \mathcal{M}$, existem vizinhanças abertas $\mathcal{U}$ de Y e $\mathcal{V}$ de Z tais que $\mathcal{U} \cap \mathcal{V} = \varnothing$. Em outras palavras, quaisquer dois pontos distintos podem ser separados por vizinhanças disjuntas.

Exemplo 1. Considere o conjunto $X = \{a, b\}$ com a topologia trivial $\tau = \{\varnothing, X\}$.
Note que as únicas vizinhanças abertas de a e b são X e $\varnothing$. Portanto, não existem duas vizinhanças abertas disjuntas de a e b. Assim, X não satisfaz a condição de Hausdorff. Portanto, (X, τ) não é um espaço Hausdorff.

Definição 4. Um espaço topológico $\mathcal{M}$ é *localmente euclidiano* se, para todo ponto $Y \in \mathcal{M}$, existe uma vizinhança $\mathcal{U} \subset \mathcal{M}$ de Y tal que, a função $\varphi : \mathcal{U} \to \varphi(\mathcal{U})$, onde $\varphi(\mathcal{U})$ é um aberto do $\mathbb{R}^m$, é um homeomorfismo. O par $(\mathcal{U}, \varphi)$ é chamado de carta de $\mathcal{M}$, enquanto os elementos $\varphi(Y)$ são chamados de coordenadas de Y na carta $(\mathcal{U}, \varphi)$.

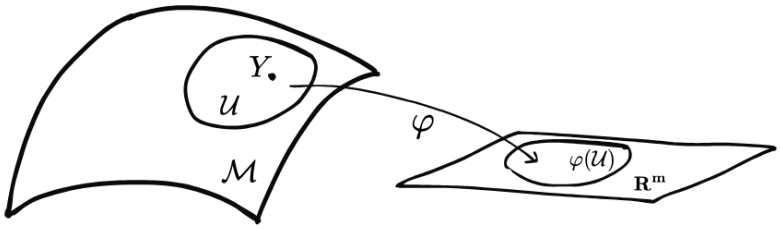

Figura 1.1: Carta $(\mathcal{U}, \varphi)$ de uma variedade $\mathcal{M}$.

Em poucas palavras, uma *variedade topológica* $\mathcal{M}$ é um espaço topológico que satisfaz três propriedades: é Hausdorff, tem base enumerável e é localmente euclidiano. Em particular, a última propriedade é especialmente importante para os objetivos que serão apresentados adiante.

Observação 1. Dizemos que a variedade topológica $\mathcal{M}$ é de dimensão m, se para toda carta $(\mathcal{U}, \varphi)$ de $\mathcal{M}$, onde $\varphi : \mathcal{U} \to \varphi(\mathcal{U})$, tem-se $\varphi(\mathcal{U}) \subset \mathbb{R}^m$.

1.2 Variedade Diferenciável

O interesse ao se definir o homeomorfismo φ, está em estabelecer um meio de "traduzir" propriedades locais de uma vizinhança de $\mathcal{M}$, para o contexto familiar de um aberto do $\mathbb{R}^m$, viabilizando a análise de objetos definidos na variedade, a partir de suas imagens.

Nosso objetivo é realizar cálculo em variedades. Para isso, utilizaremos o fato de que toda variedade topológica é localmente euclidiana. Assim, ao considerarmos uma função $F : \mathcal{M} \to \mathcal{N}$, onde $\mathcal{M}$ e $\mathcal{N}$ são variedades topológicas, podemos investigar, por exemplo, a diferenciabilidade de F por meio de cartas. Mais precisamente, tomamos os homeomorfismos $\varphi : \mathcal{U}_Y \to \varphi(\mathcal{U}_Y) \subset \mathbb{R}^m$, onde $\mathcal{U}_Y$ é um aberto contido em $\mathcal{M}$ e $\psi : \mathcal{U}_{F(Y)} \to \varphi(\mathcal{U}_{F(Y)}) \subset \mathbb{R}^n$, em que $\mathcal{U}_{F(Y)}$ é um aberto contido e, $\mathcal{N}$, e analisamos a função composta $\psi \circ F \circ \varphi^{-1}$. Como essa composição é uma função entre abertos do $\mathbb{R}^m$ e $\mathbb{R}^n$, podemos aplicar as ferramentas usuais do Cálculo Diferencial para verificar sua diferenciabilidade.

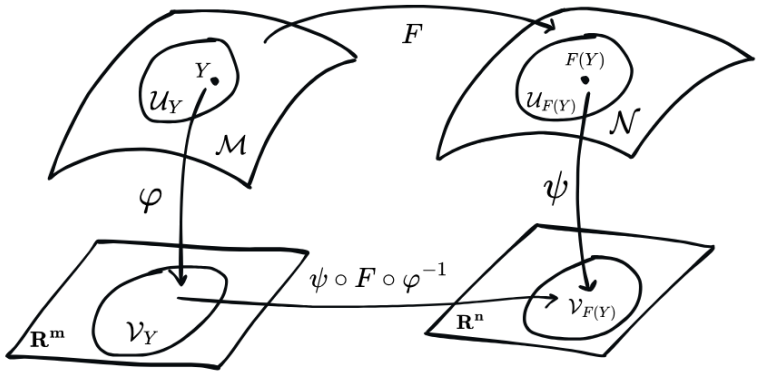

Figura 1.2: Verificação da diferenciabilidade de F por meio da composição $\psi \circ F \circ \varphi^{-1}$.

À primeira vista, essa abordagem parece adequada. No entanto, existe uma dificuldade sutil: a conclusão sobre a diferenciabilidade de F pode depender da escolha das cartas. É possível que, com uma certa escolha de cartas, a função composta não seja diferenciável, enquanto com outras, seja. Essa inconsistência indica que a conclusão da diferenciabilidade de F, não pode depender de cartas arbitrárias. Esse problema motiva a seleção de um conjunto adequado de cartas.

Definição 5. Sejam $(\mathcal{U}_1, \varphi)$ e $(\mathcal{U}_2, \psi)$ cartas de $\mathcal{M}$, onde $\varphi : \mathcal{U}_1 \to \varphi(\mathcal{U}_1) \subset \mathbb{R}^m$ e $\psi : \mathcal{U}_2 \to \varphi(\mathcal{U}_2) \subset \mathbb{R}^m$. Dizemos que $(\mathcal{U}_1, \varphi)$ e $(\mathcal{U}_2, \psi)$ são C^∞- *compatíveis* se as funções $\psi \circ \varphi^{-1} : \varphi(\mathcal{U}_1 \cap \mathcal{U}_2) \to \psi(\mathcal{U}_1 \cap \mathcal{U}_2)$ e $\varphi \circ \psi^{-1} : \psi(\mathcal{U}_1 \cap \mathcal{U}_2) \to \varphi(\mathcal{U}_1 \cap \mathcal{U}_2)$ são de classe C^∞.

Definição 6. Seja $\mathcal{M}$ uma variedade topológica. O conjunto $\mathcal{A} = \{(\mathcal{U}_\alpha, \varphi_\alpha)\}$ é dito *atlas*, se todas as cartas $(\mathcal{U}_\alpha, \varphi_\alpha)$ de $\mathcal{M}$, são C^∞- compatíveis e $\{\mathcal{U}_\alpha\}$ é uma cobertura de $\mathcal{M}$.

Definição 7. Um atlas $\mathcal{A}$ é dito *maximal* se, dada uma carta $(\mathcal{U}, \varphi)$ de uma variedade topológica $\mathcal{M}$, compatível com $(\mathcal{U}_\alpha, \varphi_\alpha)$, para todo $(\mathcal{U}_\alpha, \varphi_\alpha) \in \mathcal{A}$, temos que $(\mathcal{U}, \varphi) \in \mathcal{A}$.

O conceito de atlas maximal é fundamental, pois garante que a estrutura diferenciável da variedade não dependa da escolha particular das cartas. Em outras palavras, todas

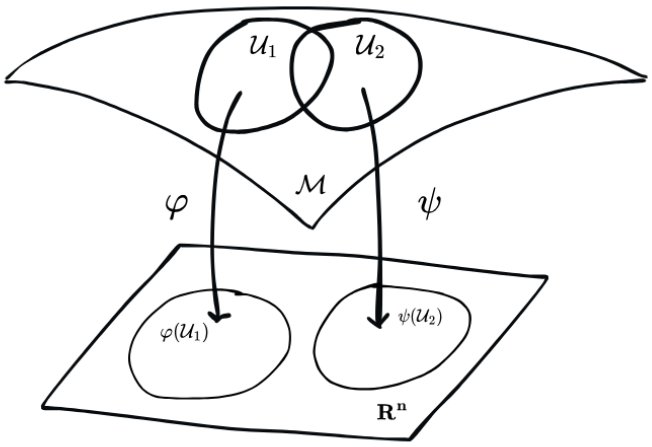

Figura 1.3: Compatibilidade entre duas cartas de $\mathcal{M}$.

as cartas compatíveis já estão incluídas no atlas. Com essa noção estabelecida, podemos agora apresentar a definição de variedade diferenciável, a qual pode ser encontrada em [15, p.53].

Definição 8. Uma *variedade diferenciável* é um par $(\mathcal{M}, \mathcal{A})$, onde $\mathcal{M}$ é uma variedade topológica e $\mathcal{A}$ é um atlas maximal de $\mathcal{M}$. O atlas maximal $\mathcal{A}$ também é chamado de estrutura diferenciável em $\mathcal{M}$. Por simplicidade, é comum referir-se à variedade diferenciável apenas por $\mathcal{M}$, omitindo a menção ao atlas.

Exemplo 2. O espaço euclidiano $\mathbb{R}^{n\times p}$, conjunto das matrizes reais de ordem $n \times p$ munido da topologia usual, é uma variedade diferenciável.
Tomando a norma de Frobenius, dada por

$$||A||_F := \left(\sum_{i=1}^{n}\sum_{j=1}^{p} a_{ij}^2\right)^{1/2}, \tag{1.1}$$

podemos definir a distância entre duas matrizes $A, B \in \mathbb{R}^{n\times p}$ como $d(A, B) := ||A - B||_F$. Com isso, se $A \neq B$, tomando $r = \frac{1}{2}d(A, B) > 0$, temos que $B_r(A) = \{X \in \mathbb{R}^{n\times p}; d(A, X) < r\}$ e $B_r(B)$ são disjuntas. Logo, $\mathbb{R}^{n\times p}$ é Hausdorff.
Além disso, dado um aberto $\mathcal{U}$ tal que $A \in \mathcal{U}$, existe $r > 0$ com $B_r(A) \subset \mathcal{U}$. Escolhendo $q \in \mathbb{Q}_+ = \{x \in \mathbb{Q}; x > 0\}$ e $C \in \mathbb{Q}^{n\times p} = \{X = (x_{ij}); x_{ij} \in \mathbb{Q}, 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p\}$, tais que $0 < q < r$ e $||A - C||_F < r - q$, temos que $B_q(C) \subset B_r(A) \subset U$. Logo, $B = \{B_q(C); C \in \mathbb{Q}^{n\times p}, q \in \mathbb{Q}_+\}$ é uma base para $\mathbb{R}^{n\times p}$. Como B é enumerável, concluímos que $\mathbb{R}^{n\times p}$ tem base enumerável.
Note que $\mathbb{R}^{n\times p}$ é localmente euclidiano de dimensão np. De fato, considerando a aplicação

$$\text{vec} : \mathbb{R}^{n\times p} \to \mathbb{R}^{np}, \qquad \text{vec}\big(a_{ij}\big) = (a_{11}, a_{12}, \ldots, a_{1p}, a_{21}, \ldots, a_{np}), \tag{1.2}$$

que é linear e bijetora, temos um homeomorfismo entre $\mathbb{R}^{n\times p}$ e $\mathbb{R}^{np}$. Portanto, cada ponto de $\mathbb{R}^{n\times p}$ possui vizinhanças abertas homeomorfas a abertos de $\mathbb{R}^{np}$.
Para mostrar que $\mathbb{R}^{n\times p}$ é uma variedade diferenciável, consideremos a aplicação dada por (1.2). Tal aplicação é linear e bijetora, logo um isomorfismo de espaços vetoriais reais de dimensão np. Como aplicações lineares entre espaços de dimensão finita são contínuas e vec é invertível, a inversa vec^{-1} é linear e portanto, contínua. Logo, vec é um homeomorfismo entre $\mathbb{R}^{n\times p}$ (com a norma de Frobenius) e $\mathbb{R}^{np}$ (com a norma euclidiana). Assim, obtemos um atlas $\mathcal{A} = \{(\mathbb{R}^{n\times p}, \text{vec})\}$, uma vez que vec é diferenciável, pois é linear. Note que, conforme definimos o homeomorfismo em (1.2), obtemos a seguinte a carta $(\mathbb{R}^{n\times p}, \text{vec})$ para a variedade topológica $\mathbb{R}^{n\times p}$. Como essa carta é única, não existem compatibilidade entre diferentes cartas a serem verificadas. Portanto, a regularidade necessária para a estrutura diferenciável é satisfeita trivialmente. Assim, $\mathcal{A}$ define uma estrutura diferenciável de classe C^∞ em $\mathbb{R}^{n\times p}$.
Concluímos que $\mathbb{R}^{n\times p}$ é uma variedade diferenciável de dimensão np. □

Definição 9. Seja $\mathcal{M}$ uma variedade diferenciável e seja $\mathcal{X} \subset \mathcal{M}$ um subconjunto não vazio. O conjunto $\mathcal{X}$ é dito subvariedade de dimensão k de $\mathcal{M}$, se para cada $p \in \mathcal{X}$, existir uma carta $(\mathcal{U}, \varphi)$ de $\mathcal{M}$ tal que

$$\varphi(\mathcal{U} \cap \mathcal{X}) = \varphi(\mathcal{U}) \cap (\mathbb{R}^k \times \{0\}),$$

onde $0 \in \mathbb{R}^{\dim\mathcal{M}\text{-k}}$. A restrição dessas cartas a $\mathcal{U} \cap \mathcal{X}$ formam um atlas em $\mathcal{X}$, que induz uma estrutura diferenciável em $\mathcal{X}$, com dimensão menor ou igual que a de $\mathcal{M}$.

Observação 2. Se $\mathcal{X}$ é uma subvariedade de $\mathcal{M}$, então ele próprio é, por definição, uma variedade diferenciável.

Exemplo 3. A *variedade de Stiefel*, denotada por $\text{St}(p, n)$, é definida como

$$\text{St}(p, n) = \{Y \in \mathbb{R}^{n\times p}; Y^T Y = I_p\}.$$

Trata-se de um subconjunto de $\mathbb{R}^{n\times p}$ que, conforme será provado na Proposição 3, é uma subvariedade de $\mathbb{R}^{n\times p}$.

Definição 10. Seja $\mathcal{M}$ uma variedade diferenciável, $Y \in \mathcal{M}$ e $f : \mathcal{M} \to \mathbb{R}$. A função f é C^∞ ou *suave no ponto* $Y \in \mathcal{M}$, se existe uma carta $\varphi : \mathcal{U} \to \mathbb{R}^m$, onde $\mathcal{U}$ é uma vizinhança de Y, tal que $f \circ \varphi^{-1} : \varphi(\mathcal{U}) \to \mathbb{R}$ é C^∞ em $\varphi(Y)$. A função f é dita *suave* se é suave em todo ponto $Y \in \mathcal{M}$.

Observação 3. A suavidade de f no ponto Y independe da escolha da carta utilizada.
De fato, seja $Y \in \mathcal{U}$ e $(\mathcal{U}, \varphi)$ uma carta tal que $f \circ \varphi^{-1}$ é suave em $\varphi(Y)$. Suponha agora que $(\mathcal{V}, \psi)$ seja outra carta, com $Y \in \mathcal{V}$ e $\mathcal{U} \cap \mathcal{V} \neq \varnothing$. Então, na região $\psi(\mathcal{U} \cap \mathcal{V})$, temos

que

$$f \circ \psi^{-1} = (f \circ \varphi^{-1}) \circ (\varphi \circ \psi^{-1}).$$

Como $f \circ \varphi^{-1}$ é suave em $\varphi(Y)$ e $\varphi \circ \psi^{-1}$ é um difeomorfismo suave entre abertos de $\mathbb{R}^m$, a composição acima é suave em $\psi(Y)$. Logo, a suavidade de f é bem definida, ou seja, não depende da carta escolhida.

Observação 4. Seja $f : \mathcal{M} \to \mathbb{R}$ uma função suave em $Y \in \mathcal{M}$. Por definição, isso significa que, para uma carta $(\mathcal{U}, \varphi)$ com $Y \in \mathcal{U}$, a função $f \circ \varphi^{-1}$ é de classe C^∞ em $\varphi(Y)$, sendo, em particular, contínua nesse ponto. Como φ também é contínua (pois é um homeomorfismo), segue que $f = (f \circ \varphi^{-1}) \circ \varphi$ é contínua em Y. Portanto, toda função suave é, em particular, contínua. Assim, ao trabalharmos com funções suaves, não há perda de generalidade em supor que tais funções são contínuas.

Definição 11. Sejam $\mathcal{M}$ e $\mathcal{N}$ variedades diferenciáveis de dimensões m e n, respectivamente. Dizemos que um mapa $F : \mathcal{M} \to \mathcal{N}$ é C^∞ ou *suave em um ponto* $Y \in \mathcal{M}$ se existirem cartas $(\mathcal{U}, \varphi)$ de $\mathcal{M}$, com $Y \in \mathcal{U}$, e $(\mathcal{V}, \psi)$ de $\mathcal{N}$, com $F(Y) \in \mathcal{V}$, tais que a composição

$$\psi \circ F \circ \varphi^{-1} : \varphi(\mathcal{U} \cap F^{-1}(\mathcal{V})) \subset \mathbb{R}^m \to \mathbb{R}^n$$

seja de classe C^∞ no ponto $\varphi(Y)$. Dizemos que F é *suave* se for suave em todo ponto $Y \in \mathcal{M}$.

Exemplo 4. Seja $S^1 = \{(x, y) \in \mathbb{R}^2; x^2 + y^2 = 1\}$. Considere as variedades diferenciáveis $\mathcal{M} = S^1 \subset \mathbb{R}^2$ e $\mathcal{N} = \mathbb{R}$. Definimos o mapa $F : S^1 \to \mathbb{R}$, dado por $F(x, y) = x$. Mostremos que F é suave em um ponto arbitrário; para isso, verificaremos a suavidade em $Y = (1, 0)$.
Seja $\mathcal{U} = S^1 \backslash \{(0, 1)\}$. A aplicação $\varphi : \mathcal{U} \to \mathbb{R}$, dada por $\varphi(x, y) = \frac{x}{1-y}$ é uma carta contendo o ponto Y, pois $\varphi(1, 0) = 1$. Para a variedade $\mathcal{N}$, tomamos a carta $(\mathcal{V}, \psi)$, onde $\mathcal{V} = \mathbb{R}$ e $\psi(t) = t$, para todo $t \in \mathbb{R}$. Note que

$$\varphi^{-1}(t) = \left(\frac{2t}{1 + t^2}, \frac{t^2 - 1}{1 + t^2} \right).$$

Assim, a composição $\psi \circ F \circ \varphi^{-1} : \mathbb{R} \to \mathbb{R}$ é dada por

$$(\psi \circ F \circ \varphi^{-1})(t) = F\left(\frac{2t}{1 + t^2}, \frac{t^2 - 1}{1 + t^2} \right) = \frac{2t}{1 + t^2}.$$

A função $\psi \circ F \circ \varphi^{-1}$ é suave em toda a reta, pois é uma função racional cujo denominador é estritamente positivo. Portanto, $\psi \circ F \circ \varphi^{-1}$ é de classe C^∞ no ponto $\varphi(Y) = 1$. Concluímos que F é suave no ponto Y.

1.3 Espaço tangente

Uma vez estabelecido o conceito de diferenciabilidade para funções definidas em variedades diferenciáveis, o passo seguinte consiste em analisar como se define e se calcula a sua derivada.

No $\mathbb{R}^n$, a derivada de uma função consiste na sua linearização em um determinado ponto. De forma análoga, ao considerarmos um mapa suave definido entre variedades diferenciáveis, também desejamos obter sua linearização. No entanto, para que isso seja possível, precisamos dispor de um espaço vetorial associado a cada ponto do domínio e da imagem da função, de modo a podermos definir uma aplicação linear entre eles. Para isso, construiremos um mapa cujo domínio seja o espaço tangente à variedade no ponto $Y \in \mathcal{M}$, e cujo contradomínio seja o espaço tangente à imagem no ponto Y.

Definição 12. Seja $\mathfrak{F}(\mathcal{M})$ o conjunto das funções suaves $f : \mathcal{M} \to \mathbb{R}$. Uma *derivação* de $\mathfrak{F}(\mathcal{M})$ em $Y \in \mathcal{M}$ é uma aplicação $D : \mathfrak{F}(\mathcal{M}) \to \mathbb{R}$ tal que D é linear, isto é,

$$D(af + bg) = aDf + bDg, \quad \text{onde} \quad a, b \in \mathbb{R}, \quad \text{e}$$

$$D(fg) = (Df)g(Y) + f(Y)Dg, \quad \forall f, g \in \mathfrak{F}(\mathcal{M}).$$

Definição 13. Um *vetor tangente* em um ponto Y em uma variedade diferenciável $\mathcal{M}$ é uma derivação em Y.

O conjunto de todos os vetores tangentes a $\mathcal{M}$ em um ponto Y formam um espaço vetorial $T_Y\mathcal{M}$, chamado de *espaço tangente*. A união disjunta dos espaços tangentes $T_Y\mathcal{M}$ para todos os pontos $Y \in \mathcal{M}$, formam o *fibrado tangente* de $\mathcal{M}$, usualmente denotado por $T\mathcal{M}$.

Definição 14. Uma *curva suave* em uma variedade diferenciável $\mathcal{M}$ é um mapa suave $c : (-\epsilon, \epsilon) \to \mathcal{M}$ tal que $c(0) = Y$, para algum ponto fixado $Y \in \mathcal{M}$.

A cada curva suave c em $\mathcal{M}$ com $c(0) = Y$, pode-se associar uma aplicação linear $c'(0) : \mathfrak{F}(\mathcal{M}) \to \mathbb{R}$, definida, para toda função $f \in \mathfrak{F}(\mathcal{M})$, por

$$c'(0)(f) = \frac{d}{dt}(f \circ c)(t)\bigg|_{t=0}.$$

A aplicação $c'(0)$, assim definida, é uma derivação no ponto Y, isto é, uma aplicação linear que satisfaz $c'(0)(fg) = c'(0)(f) \cdot g(Y) + f(Y) \cdot c'(0)(g), \quad \forall f, g \in \mathfrak{F}(\mathcal{M})$. Dessa forma, podemos definir o espaço tangente à variedade $\mathcal{M}$ no ponto Y, como o conjunto de todas as derivadas $c'(0)$ de curvas suaves $c : (-\varepsilon, \varepsilon) \to \mathcal{M}$ tais que $c(0) = Y$, ou seja, $T_Y\mathcal{M} = \{c'(0) \mid c \in C^\infty((-\varepsilon, \varepsilon), \mathcal{M}),\ c(0) = Y\}$.

Observação 5. Um campo vetorial $\xi : \mathcal{M} \to T_Y\mathcal{M}$ é um mapa suave que atribui, a cada ponto $Y \in \mathcal{M}$ um vetor tangente $\xi \in T_Y\mathcal{M}$. O conjunto de todos os campos vetoriais suaves em $\mathcal{M}$ é denotado por $\mathfrak{X}(\mathcal{M})$.

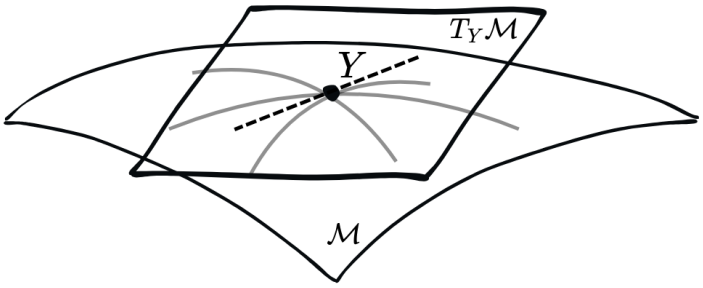

Figura 1.4: Espaço tangente $T_Y\mathcal{M}$ à variedade $\mathcal{M}$ no ponto Y. Cada vetor tangente corresponde à derivada de uma curva suave em $\mathcal{M}$ que passa por Y.

Em um espaço euclidiano $\mathcal{E}$, mover-se na direção de um vetor é uma operação simples: dado um ponto $x \in \mathcal{E}$ e um vetor v, o ponto $x + tv \in \mathcal{E}$, para todo $t \in \mathbb{R}$. Em uma variedade, essa operação não está necessariamente bem definida, já que somar um ponto a um vetor tangente nem sempre resulta em um ponto na variedade. Para formalizar deslocamentos ao longo de direções tangentes dentro da variedade, introduz-se o conceito de *retração*, que associa a cada vetor tangente um ponto da variedade.

Definição 15. Uma *retração* R, em uma variedade $\mathcal{M}$, é um mapa suave

$$\mathrm{R} : T\mathcal{M} \longrightarrow \mathcal{M}, \quad (Y, \xi) \longmapsto \mathrm{R}_Y(\xi),$$

tal que cada curva $c(t) = \mathrm{R}_Y(t\xi)$ satisfaz $c(0) = Y$ e $c'(0) = \xi$.

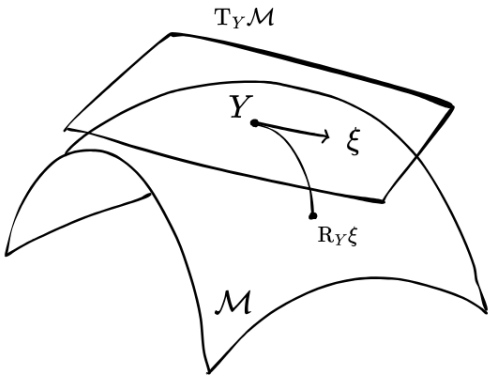

Figura 1.5: Retração R em uma variedade $\mathcal{M}$, mostrando como um vetor tangente $\xi \in T_Y\mathcal{M}$ é mapeado para um ponto $\mathrm{R}_Y\,\xi \in \mathcal{M}$.

Exemplo 5. Dada uma matriz $X \in \mathrm{St}(p, n)$ e uma direção tangente $\xi \in T_X\mathrm{St}(p, n)$, define-se a aplicação

$$\mathrm{R}_X(\xi) := \mathrm{qf}(X + \xi),$$

onde $\mathrm{qf}(A)$ denota o fator Q da decomposição QR de A, isto é, a decomposição $A = QR$ com $Q \in \mathrm{St}(p, n)$ e R triangular superior com diagonal estritamente positiva.
Essa aplicação satisfaz as propriedades que caracterizam uma retração. Para mais detalhes e uma verificação completa dessas propriedades, veja [20].

Existem diversas escolhas possíveis de retrações em variedades diferenciáveis. A escolha de cada uma delas pode ser mais adequada para diferentes objetivos.

1.4 Variedade Riemanniana

Para analisar variações de uma função $f : \mathcal{M} \to \mathbb{R}$ em diferentes direções sobre uma variedade, é essencial dispor a esta variedade, uma métrica. Para isso, atribui-se a cada espaço tangente $T_Y\mathcal{M}$ um produto interno $\langle \cdot, \cdot \rangle_Y$, como definimos a seguir.

Definição 16. Um *produto interno* no espaço tangente $T_Y\mathcal{M}$ é uma aplicação bilinear, simétrica e definida positiva

$$\langle \cdot, \cdot \rangle_Y : T_Y\mathcal{M} \times T_Y\mathcal{M} \to \mathbb{R}.$$

Ele induz uma norma para vetores tangentes dada por $\|u\|_Y = \sqrt{\langle u, u \rangle_Y}$. Equipar uma *métrica* em $\mathcal{M}$ consiste em atribuir, a cada ponto $Y \in \mathcal{M}$, um produto interno $\langle \cdot, \cdot \rangle_Y$ no espaço tangente $T_Y\mathcal{M}$.

Em particular, consideraremos métricas que variam suavemente ao longo da variedade. Para formalizar essa noção, apresentamos a seguinte definição.

Definição 17. Uma *variedade Riemanniana* é uma variedade diferenciável $\mathcal{M}$ equipada com uma *métrica Riemanniana* $\langle \cdot, \cdot \rangle$, isto é, um produto interno definido em cada espaço tangente $T_Y\mathcal{M}$ tal que, para quaisquer campos vetoriais diferenciáveis η e ξ definidos em uma vizinhança $\mathcal{V} \subset \mathcal{M}$, a função $\langle \eta_Y, \xi_Y \rangle$, é diferenciável em $\mathcal{V}$.

Observação 6. Se uma variedade $\mathcal{M}$ é munida de uma métrica Riemanniana, é natural esperar que variedades construídas a partir de $\mathcal{M}$ possam herdar, de maneira adequada, uma métrica Riemanniana induzida pela métrica de $\mathcal{M}$.

Definição 18. Seja $\overline{\mathcal{M}}$ uma variedade Riemannina e $\mathcal{M}$ uma subvariedade de $\overline{\mathcal{M}}$. O *complemento ortogonal* de $T_Y\mathcal{M}$ em $T_Y\overline{\mathcal{M}}$, denotado por $(T_Y\mathcal{M})^\perp$, é o conjunto

$$(T_Y\mathcal{M})^\perp = \{\xi \in T_Y\overline{\mathcal{M}}; \langle \xi, \eta \rangle_Y = 0 \ \text{ para todo } \ \eta \in T_Y\mathcal{M}\}. \tag{1.3}$$

Observação 7. Usando uma métrica Riemanniana de $\overline{\mathcal{M}}$, cada espaço tangente $T_Y\overline{\mathcal{M}}$ pode ser decomposto como a soma direta de $T_Y\mathcal{M}$ e de seu complemento ortogonal $(T_Y\mathcal{M})^\perp$.

Assim, todo vetor $\xi \in T_Y\overline{\mathcal{M}}$, com $Y \in \mathcal{M}$, admite uma decomposição única da forma

$$\xi = \mathrm{P}_Y\,\xi + \mathrm{P}_Y^{\perp}\,\xi, \tag{1.4}$$

onde $\mathrm{P}_Y\,\xi \in T_Y\mathcal{M}$, denota a projeção ortogonal de ξ em $T_Y\mathcal{M}$ e $\mathrm{P}_Y^{\perp}\,\xi \in (T_Y\mathcal{M})^{\perp}$, é a projeção ortogonal de ξ em $(T_Y\mathcal{M})^{\perp}$.

Associado a cada ponto $Y \in \mathrm{St}(p,n)$, da variedade de Stiefel, tem-se o espaço tangente dado por

$$T_Y\,\mathrm{St}(p,n) = \{\xi \in \mathbb{R}^{n\times p}; \xi^T Y + Y^T\xi = 0\}. \tag{1.5}$$

Equivalentemente, pode-se escrever

$$T_Y\,\mathrm{St}(p,n) = \left\{ YB + Y_\perp C \;\middle|\; B \in \mathrm{S}_{\mathrm{skew}}(p),\ C \in \mathbb{R}^{(n-p)\times p} \right\}, \tag{1.6}$$

onde $Y_\perp$ é uma matriz arbitrária $n \times (n-p)$ que satisfaz $Y^TY_\perp = 0$ e $Y_\perp^TY_\perp = I_{n-p}$, e $\mathrm{S}_{\mathrm{skew}}(p)$ denota o conjunto das matrizes $p \times p$ antissimétricas. Para mais detalhes, consulte [12] ou [20, p.42]. Além disso, como $\mathrm{St}(p,n)$ é uma subvariedade de $\mathbb{R}^{n\times p}$, ela pode ser dotada da métrica Riemanniana

$$\langle \xi_1, \xi_2 \rangle_Y := \mathrm{tr}\left(\xi_1^T\xi_2\right), \qquad \xi_1, \xi_2 \in T_Y\,\mathrm{St}(p,n), \tag{1.7}$$

a qual é induzida pelo produto interno em $\mathbb{R}^{n\times p}$, veja [20]. Sob esta métrica, a projeção ortogonal P_Y de uma matriz W no espaço tangente $T_Y\,\mathrm{St}(p,n)$ é dada por

$$\mathrm{P}_Y(W) = W - Y\ \mathrm{sym}(Y^TW), \qquad Y \in \mathrm{St}(p,n),\ \ W \in \mathbb{R}^{n\times p}, \tag{1.8}$$

onde $\mathrm{sym}(A) = \frac{1}{2}(A + A^T)$ denota a parte simétrica de A. Ou ainda,

$$\mathrm{P}_Y(W) = (I - YY^T)W + Y\,\mathrm{skew}(Y^TW), \tag{1.9}$$

onde $\mathrm{skew}(A) = \frac{1}{2}(A - A')$. Para mais detalhes sobre a estrutura, métrica e projeções associadas à variedade de Stiefel, bem como suas propriedades geométricas, consulte [20, p.48].

Capítulo 2

Método de Newton Riemanniano

Neste capítulo, apresentamos o Método de Newton, iniciando com sua formulação no espaço euclidiano e, posteriormente, sua versão para variedades Riemannianas. Abordam-se conceitos fundamentais, como gradiente e Hessiana Riemannianos, a atualização dos pontos da sequência gerada pelo método e suas propriedades matemáticas.

2.1 Método de Newton Euclidiano

Otimizar, de modo geral, consiste em determinar os pontos em que uma função atinge valores mínimos ou máximos, caracterizando as soluções que proporcionam o melhor resultado dentro do conjunto de possibilidades. Para formalizar essa ideia, consideremos uma função $f : \mathbb{R}^n \to \mathbb{R}$. O problema de encontrar x_* tal que f assuma seu menor valor nesse ponto pode ser descrito como

$$\text{minimizar } f(x),\ x \in \mathbb{R}^n. \tag{2.1}$$

Nosso objetivo é, portanto, identificar o minimizador da função, isto é, o ponto no qual o valor de f é menor do que em qualquer outro ponto do domínio. Esse conceito é expresso formalmente pela definição a seguir.

Definição 19. Dizemos que o ponto $x_* \in \mathbb{R}^n$ é *minimizador global* de $f : \mathbb{R}^n \to \mathbb{R}$ se

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

Com o intuito de solucionar problemas dessa natureza, busca-se construir uma sequência de pontos (x_k), onde $x_k \in \mathbb{R}^n$, de modo que essa sequência convirja para o minimizador x_* de f. De modo geral, essa sequência é gerada da seguinte maneira: toma-se um ponto inicial $x_0 \in \mathbb{R}^n$ e, com o uso de diferentes técnicas, obtêm-se aproximações sucessivas $x_1, x_2, x_3, \ldots$, cada uma representando uma melhoria em relação à anterior. Isto é, a cada novo ponto gerado, espera-se que este esteja cada vez mais próximo de x_*.

Em particular, para funções $f : \mathbb{R}^n \to \mathbb{R}$, a atualização de cada ponto da sequência é feita a partir do ponto anterior e uma direção de descida, conceito que será definido a seguir.

Definição 20. Dizemos que $\xi \in \mathbb{R}^n \backslash \{0\}$ é uma *direção de descida* da função $f : \mathbb{R}^n \to \mathbb{R}$ no ponto $x \in \mathbb{R}^n$, se existe $\epsilon > 0$ tal que

$$f(x + t\xi) < f(x) \ \forall t \in (0, \epsilon).$$

Observação 8. Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função diferenciável no ponto $x \in \mathbb{R}^n$. Se $\xi \in \mathbb{R}^n$ é uma direção de descida de f, tem-se $\langle f'(x), \xi \rangle < 0$. Para mais detalhes, veja [2].

A atualização de pontos da sequência (x_k) é feita do seguinte modo:

$$x_{k+1} = x_k + \xi_k, \tag{2.2}$$

onde ξ_k é uma direção de descida de f.

A escolha da direção de descida, usada na atualização de pontos da sequência, distingue os métodos da Otimização. Cada método apresenta características próprias, com vantagens e limitações que o tornam mais adequado para determinados tipos de problemas. Dentre esses métodos, o Método de Newton distingue-se pela sua convergência excepcionalmente rápida ao minimizador, decorrente do aproveitamento de informações de segunda ordem durante o processo iterativo.

Tradicionamente, o método de Newton é empregado como uma ferramenta para a resolução do problema de determinação de zeros de um campo vetorial $F : \mathbb{R}^n \to \mathbb{R}^n$, [2]. Em particular, quando o campo vetorial em questão é o gradiente de uma função duas vezes diferenciável, $f : \mathbb{R}^n \to \mathbb{R}$, o problema se reduz à encontrar os pontos onde esse gradiente se anula, ou seja, os pontos críticos de f.

Uma das principais características do Método de Newton é o uso da Hessiana da função para determinar a direção de descida da sequência gerada. Especificamente, a direção ξ_k em um ponto x_k é obtida a partir da solução da equação linear

$$\operatorname{Hess} f(x_k)\xi_k = -\operatorname{grad} f(x_k). \tag{2.3}$$

O uso dessa informação permite, em geral, uma convergência significativamente mais rápida, desde que o ponto inicial x_0 esteja em uma vizinhança suficientemente próxima do minimizador x_*. Por essa razão, dizemos que o Método de Newton apresenta convergência local, ou seja, a convergência da sequência (x_k) para a solução x_*, depende da escolha adequada do ponto inicial [2].

A seguir, apresenta-se o algoritmo do Método de Newton, que descreve o procedimento para a geração da sequência (x_k).

Algoritmo 1: Método de Newton

1 Escolha um ponto inicial $x_0 \in \mathbb{R}^n$ e $k = 0$.
2 Calcule $\xi_k = -\operatorname{Hess}(f(x_k))^{-1} \operatorname{grad} f(x_k)$.
3 Defina $x_{k+1} = x_k + \xi_k$.
4 Tome $k := k + 1$ e retorne ao passo 2.

A velocidade com que a sequência gerada pelo Algoritmo 1 converge para a solução é caracterizada, de forma mais precisa, por meio de sua *taxa de convergência*. Diz-se que uma sequência (x_k) converge *superlinearmente* para x_* se

$$\lim_{k\to\infty} \frac{||x_{k+1} - x_*||}{||x_k - x_*||} = 0. \tag{2.4}$$

Além disso, a convergência é considerada *quadrática* se existir uma constante $C > 0$ tal que

$$\lim_{k\to\infty} \frac{||x_{k+1} - x_*||}{||x_k - x_*||^2} \leqslant C. \tag{2.5}$$

O uso de informações de segunda ordem para a obtenção da direção ξ_k, definida como a solução da equação (2.3), confere ao método de Newton uma taxa de convergência superlinear, podendo alcançar convergência quadrática, desde que o gradiente da função objetivo seja Lipschitz em uma certa vizinhança de x_*. Essa característica torna o método extremamente eficiente no que diz respeito à rapidez com que a sequência gerada pelo Algoritmo 1 se aproxima da solução. Entretanto, conforme já discutido, o método de Newton apresenta convergência local, isto é, sua eficiência depende da escolha apropriada do ponto inicial.

Essas propriedades são formalizadas pelo teorema de convergência local do Método de Newton. Para demonstrar esse resultado, utilizamos o lema apresentado a seguir, cuja prova foi retirada de [10, p.74].

Lema 1. Seja $||\cdot||$ uma noma do $\mathbb{R}^{n\times n}$ tal que $||I|| = 1$, onde I é a matriz identidade de ordem n e $E \in \mathbb{R}^{n\times n}$. Se $||E|| < 1$, então $(I - E)^{-1}$ existe e

$$||(I - E)^{-1}|| \leqslant \frac{1}{1 - ||E||}.$$

Além disso, se $A \in \mathbb{R}^{n\times n}$ é invertível e $||A^{-1}(B - A)|| < 1$, então $B \in \mathbb{R}^{n\times n}$ é invertível e

$$||B^{-1}|| \leqslant \frac{||A^{-1}||}{1 - ||A^{-1}(B - A)||}.$$

Demonstração: Suponha que $I - E$ não seja invertível. Por consequência $(I - E)x = 0$, para algum $x \neq 0$, o que implica que $x = Ex$. Segue que, $||x|| = ||Ex||$. Logo, $||E|| \geqslant 1$, uma contradição. Portanto, $I - E$ é invertível, isto é $(I - E)^{-1}$ existe. Considere agora a

identidade

$$\left(\sum_{k=0}^{N} E^k\right)(I-E) = I - E^{N+1}. \tag{2.6}$$

Como $0 < ||E|| < 1$, temos que $\lim_{k\to\infty} E^k = 0$, pois $||E^k|| \leqslant ||E||^k$. Passando o limite em (2.6), quando $N \to \infty$, obtemos

$$\left(\lim_{N\to\infty}\sum_{k=0}^{N} E^k\right)(I-E) = I.$$

Multiplicando à direita por $(I-E)^{-1}$, temos

$$(I-E)^{-1} = \sum_{k=0}^{\infty} E^k.$$

Note que a série $\sum_{k=0}^{\infty} E^k$ é convergente. Como a norma é contínua, podemos escrever

$$||(I-E)^{-1}|| = \left|\left|\sum_{k=0}^{\infty} E^k\right|\right|.$$

Segue que

$$\left|\left|\sum_{k=0}^{\infty} E^k\right|\right| \leqslant \sum_{k=0}^{\infty} ||E^k|| \leqslant \sum_{k=0}^{\infty} ||E||^k.$$

Além disso,

$$\sum_{k=0}^{\infty} ||E||^k = \frac{1}{1-||E||},$$

pois $\sum_{k=0}^{\infty} ||E||^k$ é uma série geométrica. Deste modo, obtemos

$$||(I-E)^{-1}|| \leqslant \sum_{k=0}^{\infty} ||E||^k = \frac{1}{1-||E||}.$$

Para a segunda afirmação, tome $B - A = F$ e $E = FA^{-1}$. Assim,

$$F = FA^{-1}A = EA \Rightarrow A + F = A + EA = (I+E)A.$$

Como $||E|| < 1$, a matriz $I + E$ é não singular e

$$||(I-E)^{-1}|| \leqslant \frac{1}{1-||A^{-1}F||}, \tag{2.7}$$

conforme já mostrado. Logo, $(A+F)^{-1} = A^{-1}(I+F)^{-1}$ é invertível. Temos ainda que

$$\begin{aligned}(A+F)^{-1} - A^{-1} &= (I - A^{-1}(A+F)) \cdot (A+F)^{-1} \\ &= (I - I - A^{-1}F) \cdot (A+F)^{-1}.\end{aligned}$$

Escrevendo $I = AA^{-1}$ e agrupando termos para colocar A^{-1} em evidência, obtemos

$$\begin{aligned}(A+F)^{-1} - A^{-1} &= (AA^{-1} - AA^{-1} - A^{-1}F) \cdot (A+F)^{-1} \\ &= A^{-1}(A - (A+F)) \cdot (A+F)^{-1}.\end{aligned}$$

Substituindo $A+F$ por $(I+E)A$, temos

$$\begin{aligned}((I+E)A)^{-1} - A^{-1} &= A^{-1}(A - (I+E)A) \cdot ((I+E)A)^{-1} \\ &= A^{-1}A(I - (I+E)) \cdot A^{-1}(I+E)^{-1} \\ &= -EA^{-1}(I+E)^{-1} \\ &= -FA^{-1}A^{-1}(I+E)^{-1}.\end{aligned}$$

Segue que

$$\begin{aligned}||((I+E)A)^{-1} - A^{-1}|| &= || - FA^{-1}A^{-1}(I+E)^{-1}|| \\ &\leqslant || - F||\ ||A^{-1}||^2||(I+E)^{-1}||.\end{aligned}$$

Usando (2.7), obtemos finalmente

$$||(A+F)^{-1} - A^{-1}|| \leqslant \frac{||F||||A^{-1}||^2}{1 - ||A^{-1}F||}.$$

□

O lema demonstrado é também conhecido como Teorema sobre Perturbações Pequenas de uma Matriz Não Singular (ver [2, p.24]). Em nosso contexto, esse resultado é fundamental para assegurar a convergência da sequência gerada pelo Algoritmo 1, pois garante que, em uma vizinhança adequada de x_*, a Hessiana permanece inversível e seu inverso é limitado. Dessa forma, a sequência está bem definida nessa vizinhança.

A seguir, apresenta-se a prova da convergência local do Método de Newton, evidenciando as condições sob as quais a sequência (x_k) atinge taxas superlinear e quadrática. A demonstração aqui apresentada, pode ser consultada em [2, p.104].

Teorema 1. Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função duas vezes diferenciável numa vizinhança do ponto $x_* \in \mathbb{R}^n$, com segunda derivada contínua neste ponto. Seja x_* uma solução do problema de minimizar f, que satisfaz a condição suficiente de segunda ordem. Então para qualquer ponto inicial $x_0 \in \mathbb{R}^n$ suficientemente próximo a x_*, o Algoritimo 1, gera uma sequência (x_k) bem definida que converge a x_*, com taxa superlinear. Se o gradiente

de f é Lipschitz numa vizinhança U de x_*, isto é,

$$|| \operatorname{grad} f(x) - \operatorname{grad} f(y)|| \leqslant L||x - y||, \quad \forall x, y \in U,$$

então a taxa de convergência de (x_k) é quadrática.

Demonstração: Pelo Lema 1, existe uma vizinhança U do ponto x_* e um número $M > 0$ tais que

$$\det(\operatorname{Hess} f(x)) \neq 0, \quad || \operatorname{Hess} f(x)^{-1}|| \leqslant M \quad \forall x \in U. \tag{2.8}$$

Além disso, utilizando (2.8) e $x_{k+1} = x_k - \operatorname{Hess} f(x)^{-1} \operatorname{grad} f(x)$, podemos tomar uma vizinhança U de x_* tal que, se $x_k \in U$, então

$$||x_{k+1} - x_*|| = ||x^k - x_* - \operatorname{Hess} f(x_k))^{-1} \operatorname{grad} f(x_k)||.$$

Colocando $\operatorname{Hess} f(x_k))^{-1}$ em evidência e sabendo que $\operatorname{grad} f(x_*) = 0$, obtemos

$$\begin{aligned} ||x_{k+1} - x_*|| &= || \operatorname{Hess} f(x_k))^{-1}(\operatorname{Hess} f(x_k)(x_k - x_*)) - \operatorname{grad} f(x_k))|| \\ &= ||(\operatorname{Hess} f(x_k))^{-1}(\operatorname{Hess} f(x_k)(x_k - x_*) - \operatorname{grad} f(x_k) + \operatorname{grad} f(x_*))||. \end{aligned}$$

Usando a desigualdade de Schwarz, temos

$$||x_{k+1} - x_*|| \leqslant ||(\operatorname{Hess} f(x_k))^{-1}||\ || \operatorname{grad} f(x_k) - \operatorname{grad} f(x_*) - \operatorname{Hess} f(x_k)(x_k - x_*)||.$$

Note que

$$\operatorname{grad} f(x_k) - \operatorname{grad} f(x_*) = \int_0^1 \operatorname{Hess} f(tx_k + (1-t)x_*)(x_k - x_*)dt.$$

Assim, podemos escrever

$$||x_{k+1} - x_*|| \leqslant ||(\operatorname{Hess} f(x_k))^{-1}||\ ||(\int_0^1 \operatorname{Hess} f(tx_k + (1-t)x_*)(x_k - x_*))dt - \operatorname{Hess} f(x_k)(x_k - x_*)||.$$

Segue que

$$\begin{aligned} ||x_{k+1} - x_*|| &\leqslant ||(\operatorname{Hess} f(x_k))^{-1}||\ || \int_0^1 (\operatorname{Hess} f(tx_k + (1-t)x_*) - \operatorname{Hess} f(x_k))(x_k - x_*)dt|| \\ &\leqslant ||(\operatorname{Hess} f(x_k))^{-1}|| \int_0^1 || \operatorname{Hess} f(tx_k + (1-t)x_*) - \operatorname{Hess} f(x_k)||dt\ ||(x_k - x_*)|| \\ &\leqslant ||(\operatorname{Hess} f(x_k))^{-1}|| \sup_{t \in [0,1]} || \operatorname{Hess} f(tx_k + (1-t)x_*) - \operatorname{Hess} f(x_k)||\ ||(x_k - x_*)||. \end{aligned}$$

Como $||\operatorname{Hess} f(x)^{-1}|| \leqslant M$, obtemos

$$||x_{k+1} - x_*|| \leqslant M \ \sup_{t\in[0,1]} \ ||\operatorname{Hess} f(tx_k + (1-t)x_*) - \operatorname{Hess} f(x_k)|| \ ||(x_k - x_*)||. \quad (2.9)$$

Sabemos, por hipótese, que $\operatorname{Hess} f(x_k)$ é contínua. Isto é, para todo $\epsilon > 0$, existe $\delta > 0$ tal que

$$x_k \in B(x_*, \delta) \Rightarrow ||\operatorname{Hess} f(x_k) - \operatorname{Hess} f(x_*)|| < \epsilon.$$

Como $x_k \in B(x_*, \delta)$, então $\|x_k - x_*\| < \delta$. Para $t \in [0,1]$, vale

$$\|tx_k + (1-t)x_* - x_*\| = t\|x_k - x_*\| < \delta,$$

isto é, $tx_k + (1-t)x_* \in B(x_*, \delta)$. Note ainda que

$$\begin{aligned} ||\operatorname{Hess} f(tx_k + (1-t)x_*) - \operatorname{Hess} f(x_*)|| =& ||\operatorname{Hess} f(tx_k + (1-t)x_*) - \operatorname{Hess} f(x_*) \\ &+ \operatorname{Hess} f(x_*) - \operatorname{Hess} f(x_*)||. \end{aligned}$$

Usando a desigualdade triangular, obtemos

$$\begin{aligned} ||\operatorname{Hess} f(tx_k + (1-t)x_*) - \operatorname{Hess} f(x_*)|| \leqslant& ||\operatorname{Hess} f(tx_k + (1-t)x_*) - \operatorname{Hess} f(x_k)|| \\ &+ ||\operatorname{Hess} f(x_k) - \operatorname{Hess} f(x_*)|| \\ <& \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned}$$

Logo, de (2.9), se $\sup_{t\in[0,1]} \ ||\operatorname{Hess} f(tx_k + (1-t)x_*) - \operatorname{Hess} f(x_k)|| \to 0$ $(k \to \infty)$, tem-se que $x_k \to x_*$, uma vez que $||x_k - x_*||$ é limitado para todo $k \to \infty$.
Tomando $q_k = M \ \sup_{t\in[0,1]} \ ||\operatorname{Hess} f(tx_k + (1-t)x_*) - \operatorname{Hess} f(x_k)||$, da desigualdade (2.9), obtemos

$$\frac{||x_{k+1} - x_*||}{||x_k - x_*||} \leqslant q_k.$$

Segue imediatamente que $\lim_{k\to\infty} q_k = 0$. Concluímos que, para todo x_0 suficientemente próximo a x_*, o Algoritmo 1 gera uma sequência (x_k) bem definida que converge para x_*. Ademais, a taxa de convergência é superlinear.
Se a matriz hessiana de f é Lipschitz em U com módulo $L > 0$, obtemos

$$||\operatorname{Hess} f(tx_k + (1-t)x_*) - \operatorname{Hess} f(x_k)|| \leqslant L||(tx_k + (1-t)x_*) - x_k||.$$

Note que

$$\begin{aligned} ||(tx_k + (1-t)\bar{x}) - x_k|| &= ||(tx_k + (1-t)\bar{x}) - tx_k - (1-t)x_k|| \\ &= (1-t)||\bar{x} - x_k||, \quad t \in [0,1]. \end{aligned}$$

Segue que

$$|| \operatorname{Hess} f(tx_k + (1-t)x_*) - \operatorname{Hess} f(x_k)|| \leqslant L(1-t)||x_* - x_k||.$$

Substituindo em (2.9), obtemos

$$||x_{k+1} - x_*|| \leqslant M \ \sup_{t\in[0,1]} \ L(1-t)||x_* - x_k||^2.$$

Note que o valor máximo de $L(1-t)$ ocorre quando $t = 0$. Logo,

$$\sup_{t\in[0,1]} (L(1-t)) = L.$$

Portanto

$$||x_{k+1} - x_*|| \leqslant ML||x_k - x_*||^2.$$

Isto implica que a convergência é quadrática. □

2.2 Otimização em Variedades

O problema de otimização apresentado até aqui é classificado como irrestrito, ou seja, não está sujeito a quaisquer restrições sobre a variável $x \in \mathbb{R}^n$. No entanto, é frequente que problemas de otimização envolvam a minimização de uma função sob certas restrições. Nesses casos, o problema passa a ser classificado como restrito, e sua formulação geral pode ser expressa por

$$\text{minimizar } f(x) \text{ sujeito a } x \in \Omega, \tag{2.10}$$

onde $f : \mathbb{R}^n \to \mathbb{R}$ é uma função diferenciável e $\Omega \subset \mathbb{R}^n$ é não vazio.

Para encontrar soluções de problemas do tipo (2.10), torna-se necessário adaptar os métodos empregados em problemas irrestritos. O método do gradiente, o qual adota $\xi_k = -\operatorname{grad} f(x_k)$ como direção de descida é um exemplo disso. Em problemas restritos, nem todas as direções de descida, fornecidas pelo método do gradiente, garantem que o próximo ponto da sequência gerada permaneça dentro do conjunto Ω. Para contornar essa limitação, utiliza-se a projeção da direção de descida sobre o conjunto Ω, obtendo-se então uma direção viável. A Figura 2.1 ilustra esse processo.

No entanto, projetar um ponto x_k em um conjunto Ω é, em muitos casos, uma tarefa desafiadora, mesmo quando Ω possui uma estrutura relativamente simples. Essa dificuldade pode ser atribuída ao fato de que a projeção nem sempre possui uma expressão fechada e, mesmo quando existe, seu cálculo pode ser computacionalmente custoso.

Evidentemente, lidar com problemas irrestritos é, em geral, mais simples do que com problemas restritos. Convenientemente, é possível transformar um problema restrito em um problema irrestrito, desde que o conjunto Ω seja uma variedade diferenciável. Assim,

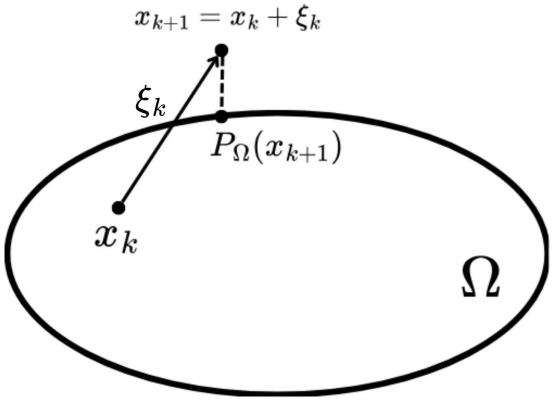

Figura 2.1: Projeção do ponto x_{k+1} no conjunto Ω.

se $\Omega = \mathcal{M}$, onde $\mathcal{M}$ é uma variedade diferenciável, então o problema (2.10) equivale a

$$\text{minimizar } f(Y),\ Y \in \mathcal{M}, \tag{2.11}$$

onde $f : \mathcal{M} \to \mathbb{R}$.
O procedimento para resolver o problema (2.11) mantém a mesma estratégia apresentada: parte-se de um ponto inicial $Y_0 \in \mathcal{M}$ e, a partir deste ponto, gera-se uma sequência Y_k que deve convergir para a solução (Y_*). A atualização de cada ponto é feita a partir do ponto atual Y_k e de uma direção ξ_k.
Como discutido no Capítulo 1, atualizar os pontos da sequência utilizando (2.2) não garante que eles permaneçam na variedade $\mathcal{M}$. Por isso, para o problema (2.11), cada ponto da sequência (Y_k) é atualizado por meio de uma retração:

$$Y_{k+1} = \mathrm{R}_{Y_k}(\xi_k), \tag{2.12}$$

onde R_{Y_k} é uma retração sobre $\mathcal{M}$ e $\xi_k \in T_{Y_k}\mathcal{M}$ é a direção de descida considerada.

Observação 9. Lembre-se de que a retração R_Y é construída de forma que, para qualquer vetor tangente $\xi \in T_Y\mathcal{M}$, o ponto $\mathrm{R}_Y(\xi) \in \mathcal{M}$. Essa característica é fundamental pois garante que, ao atualizar cada iterado seguindo uma direção tangente, todos os pontos gerados permaneçam no conjunto $\Omega = \mathcal{M}$. Em outras palavras, essa propriedade é a que permite tratar o problema originalmente restrito como um problema irrestrito sobre a variedade.

A seguir, apresentam-se noções fundamentais para a minimização de funções definidas em variedades.

2.2.1 Condições de otimalidade em Variedades

Definição 21. O *minimizador* de um mapa suave $f : \mathcal{M} \to \mathbb{R}$, é o ponto $Y_* \in \mathcal{M}$ tal que $f(Y_*) \leqslant f(Y)$, para todo $Y \in \mathcal{M}$. Em particular, dizemos que Y_* é minimizador local de f se existir um aberto $\mathcal{U} \subset \mathcal{M}$ contendo Y_*, tal que $f(Y_*) \leqslant f(Y)$, para todo $Y \in \mathcal{U}$.

Assim, resolver o problema (2.11), considerando f como a função objetivo, consiste em encontrar o minimizador $Y* \in \mathcal{M}$ de f.

Definição 22. Um ponto $Y \in \mathcal{M}$ é *ponto crítico* da função suave $f : \mathcal{M} \to \mathbb{R}$ se $(f \circ c)'(0) \geqslant 0$, para toda curva suave c em $\mathcal{M}$ tal que $c(0) = Y$.

Observação 10. Equivalentemente, poderíamos exigir que $(f \circ c)'(0) = 0$ na Definição 22, bastando considerar simultaneamente as parametrizações $t \mapsto c(t)$ e $t \mapsto c(-t)$ da curva c. Deste modo, Y é ponto crítico se, e somente se, $Df(Y) = 0$, ou seja, no caso $\mathcal{M} = \mathbb{R}^n$, essa definição coincide com a definição usual de ponto crítico.

Proposição 1. Todo minimizador local da função suave $f : \mathcal{M} \to \mathbb{R}$ é um ponto crítico de f.

A demonstração apresentada a seguir for retirada de [19, p. 56].

Demonstração: Seja Y_* um minimizador local de f. Sabemos que existe uma vizinhança $\mathcal{U} \subset \mathcal{M}$ de Y_* tal que $f(Y_*) \leqslant f(Y)$ para todo $Y \in \mathcal{U}$. Suponha que Y_* não seja ponto crítico de f. Isto é, existe uma curva suave $c : (-\epsilon, \epsilon) \to \mathcal{M}$, com $c(0) = Y_*$, tal que $(f \circ c)'(0) < 0$. Note que $(f \circ c)'$ é contínua, pois resulta da derivada da composição de funções suaves. Como f é suave em $\mathcal{M}$ (isto é, de classe C^∞) e c é uma curva suave em $\mathcal{M}$, a composição $f \circ c$ é uma função suave, e, portanto, sua derivada $(f \circ c)'$ é contínua. Logo, sabendo que $(f \circ c)' : (-\epsilon, \epsilon) \to \mathbb{R}$, é correto afirmar que existe $\delta > 0$ tal que $(f \circ c)'(t) < 0$, $\forall t \in [0, \delta]$.
Pelo Teorema Fundamental do Cálculo, temos que

$$f(c(t)) - f(c(0)) = \int_0^t (f \circ c)'(t)dt.$$

Rearranjando, obtemos

$$f(c(t)) = f(Y_*) + \int_0^t (f \circ c)'(t)dt,$$

pois $Y_* = c(0)$. Como $(f \circ c)'(t) < 0$, temos que $\int_0^t (f \circ c)'(t)dt < 0$, para todo $t \in (0, \delta]$, pois estamos somando valores negativos. Portanto,

$$f(c(t)) = f(Y_*) + \int_0^t (f \circ c)'(t)dt < f(Y_*).$$

No entanto, como c é contínua e $\mathcal{U}$ é um aberto contendo $Y_* = c(0)$, o conjunto $c^{-1}(\mathcal{U}) =$

$\{t \in (-\epsilon, \epsilon) | c(t) \in \mathcal{U}\}$ é um aberto em $(-\epsilon, \epsilon)$ que contém o ponto $t = 0$. Assim, existe $t \in [0, \delta]$ tal que $c(t) \in \mathcal{U}$.
Por outro lado, vimos que, para todo $t \in (0, \delta]$, $f(c(t)) < f(Y_*)$. No entanto, como $c(t) \in \mathcal{U}$, e $\mathcal{U}$ é uma vizinhança na qual $f(Y_*) \leqslant f(Y)$ para todo $Y \in U$, isso implica que $f(Y_*) \leqslant f(c(t))$, uma contradição. □

Definição 23. Um ponto $Y_* \in \mathcal{M}$ é *ponto crítico de segunda ordem* da função $f : \mathcal{M} \to \mathbb{R}$ se $(f \circ c)'(0) = 0$ e $(f \circ c)''(0) \geqslant 0$, para toda curva suave c em $\mathcal{M}$ tal que $c(0) = Y_*$.

Proposição 2. Todo minimizador local da função suave $f : \mathcal{M} \to \mathbb{R}$ é um ponto crítico de segunda ordem f.

Demonstração: Sabemos, pela Proposição 1 que se Y_* é minimizador local de f, então Y_* é ponto crítico de f. Suponha que Y_* não seja um ponto crítico de segunda ordem de f. Então, existe uma curva $c : (-\epsilon, \epsilon) \to \mathcal{M}$ com $c(0) = Y_*$, tal que $(f \circ c)'(0) = 0$ e $(f \circ c)''(0) < 0$.
Note que a função $(f \circ c)''$ é contínua, pois f e c são suaves. Portanto, a composição $f \circ c$ é uma função de classe C^∞ em um intervalo real, garantindo que todas as suas derivadas, incluindo $(f \circ c)''$, sejam contínuas.
Deste modo, existe $\delta > 0$ tal que $(f \circ c)''(t) < 0$ para todo $t \in [0, \delta]$. Usando a expansão de Taylor, para cada $\tau \in [0, \delta]$, existe um $t \in [0, \tau]$ tal que

$$f(c(t)) = f(c(0)) + \tau \cdot (f \circ c)'(0) + \frac{\tau^2}{2} \cdot (f \circ c)''(t).$$

Portanto, $f(Y_*) > f(Y)$ para todo $\tau \in (0, \delta]$, uma contradição. □

Teorema 2. Seja $f : \mathcal{M} \to \mathbb{R}$ uma função suave, onde $\mathcal{M}$ é uma variedade Riemanniana. O ponto Y_* é ponto crítico de segunda ordem de f se, e só se, $\operatorname{grad} f(x) = 0$ e $\operatorname{Hess} f(x)$ é semidefinida positiva, isto é $\langle v, \operatorname{Hess} f(Y_*)[v] \rangle \geqslant 0$, para todo $v \in T_{Y_*}\mathcal{M}$ não nulo.

Demonstração: Seja $c : (-\epsilon, \epsilon) \to \mathcal{M}$ uma curva suave de $\mathcal{M}$, com $c(0) = Y_*$, $v = c'(0)$ e $u = c''(0)$. Sabemos que

$$\begin{aligned}(f \circ c)'(0) &= Df(c(0))[c'(0)] = \langle \operatorname{grad} f(Y_*), v \rangle_{Y_*} \quad \text{e} \\ (f \circ c)''(0) &= \langle \operatorname{grad} f(Y_*), u \rangle_{Y_*} + \langle \operatorname{Hess} f(Y_*)[v], v \rangle_{Y_*}.\end{aligned}$$

Suponha que $\operatorname{grad} f(Y_*) = 0$ e $\langle v, \operatorname{Hess} f(Y_*)[v] \rangle \geqslant 0$. Temos que $\langle \operatorname{grad} f(Y_*), v \rangle_{Y_*} = 0 \Rightarrow (f \circ c)'(0) = 0$. Além disso, $(f \circ c)''(0) = \langle \operatorname{Hess} f(Y_*)[v], v \rangle_{Y_*}$. Segue que $(f \circ c)''(0) \geqslant 0$. Portanto, Y_* é um ponto crítico de segunda ordem de f.
Reciprocamente, suponha que Y_* seja um ponto crítico de segunda ordem de f. Temos que $(f \circ c)'(0) = 0 \Rightarrow \langle \operatorname{grad} f(Y_*), v \rangle_{Y_*} = 0 \Rightarrow \operatorname{grad} f(Y_*) = 0$.
Além disso, $(f \circ c)''(0) \geqslant 0 \Rightarrow \langle \operatorname{grad} f(Y_*), u \rangle_{Y_*} + \langle \operatorname{Hess} f(Y_*)[v], v \rangle_{Y_*} \geqslant 0$. Como $\langle \operatorname{grad} f(Y_*), u \rangle_{Y_*} = 0$, temos que $\langle \operatorname{Hess} f(Y_*)[v], v \rangle_{Y_*} \geqslant 0$. □

A Proposição 2, assim como o Teorema 2, foram retirados de [19, p.120].

2.3 Método de Newton Riemanniano

O Método de Newton, apresentado anteriormente no contexto euclidiano, também pode ser usado na busca pelo minimizador de funções definidas sobre variedades Riemannianas. Nesses problemas, a direção de descida considerada, pertence a um espaço tangente da variedade. Em particular, a direção de Newton ξ_k também é obtida como solução da equação linear

$$\operatorname{Hess} f(Y_k)[\xi_k] = -\operatorname{grad} f(Y_k). \tag{2.13}$$

No entando, nesta versão para variedades, a Hessiana e o gradiente considerados são, respectivamente, a *Hessiana riemanniana* e o *gradiente riemanniano*, os quais serão apresentados a seguir. Essa adaptação do Método de Newton leva ao que chamamos de Método de Newton Riemanniano.

2.3.1 Gradiente e Hessiana Riemannianos

Considere a função $F : \mathbb{R}^m \to \mathbb{R}^n$ e o vetor $v \in \mathbb{R}^n$. A derivada direcional de F no ponto $x \in \mathbb{R}^m$ na direção v é dada por

$$DF(x)[v] = \lim_{t \to 0} \frac{F(x + tv) - F(x)}{t}.$$

Essa definição não é apropriada para uma função f definida em uma variedade $\mathcal{M}$, pois o ponto $x + tv$ não está, em geral, bem definido em $\mathcal{M}$.
Para contornar essa dificuldade, utilizamos a ideia de curvas na variedade. Em vez de somar diretamente um vetor ao ponto x, consideramos uma curva suave $c : (-\epsilon, \epsilon) \to \mathbb{R}$ que passa pelo ponto Y e cujo vetor tangente em $t = 0$ corresponde à direção de interesse $v \in T_Y\mathcal{M}$. Dessa forma, a derivada de F ao longo de v é obtida derivando $F \circ c$ em relação a t no ponto $t = 0$. Essa estratégia leva naturalmente à noção de diferencial de uma aplicação entre variedades, definida a seguir.

Definição 24. Seja $F : \mathcal{M} \to \mathcal{N}$ uma aplicação diferenciável entre variedades e $Y \in \mathcal{M}$. O *diferencial* de F em Y, denotado por $DF(Y) : T_Y\mathcal{M} \to T_{F(Y)}\mathcal{N}$, é definido como a aplicação linear que associa a cada vetor tangente $v \in T_Y\mathcal{M}$ à derivada direcional de F ao longo de v, isto é,

$$DF(Y)[v] \;=\; \frac{d}{dt} F(c(t))\Big|_{t=0},$$

onde $c : (-\varepsilon, \varepsilon) \to \mathcal{M}$ é uma curva suave tal que $c(0) = Y$ e $c'(0) = v$.

Definição 25. Seja f uma função suave definida em uma variedade Riemanniana $\mathcal{M}$ e $T_Y\mathcal{M}$ o espaço tangente a $\mathcal{M}$ no ponto $Y \in \mathcal{M}$. O *gradiente Riemanniano* de f no ponto

Y, denotado por $\operatorname{grad} f(Y)$, é definido como o único elemento de $T_Y\mathcal{M}$ que satisfaz

$$\langle \operatorname{grad} f(Y), \xi \rangle_Y = Df(Y)[\xi], \quad \forall \xi \in T_Y\mathcal{M}. \tag{2.14}$$

O teorema a seguir, retirado de [20], é o resultado central que utilizaremos ao longo do trabalho no que se refere ao cálculo do gradiente em variedades.

Teorema 3. Seja $\bar{f}(x)$ uma função definida em uma variedade Riemanniana $\overline{\mathcal{M}}$ e f a restrição da função $\bar{f}$ a uma subvariedade $\mathcal{M}$ de $\overline{\mathcal{M}}$. O gradiente de f é igual a projeção ortogonal do gradiente $\bar{f}$ em $T_x\mathcal{M}$, isto é,

$$\operatorname{grad} f(x) = \mathrm{P}_x \operatorname{grad} \bar{f}(x). \tag{2.15}$$

Demonstração: Sabemos que $\operatorname{grad} \bar{f}(x) \in T_x\overline{\mathcal{M}}$. Logo,

$$\operatorname{grad} f(x) = \mathrm{P}_x \operatorname{grad} \bar{f}(x) + \mathrm{P}_x^{\perp} \operatorname{grad} f(x).$$

Isolando $\mathrm{P}_x \operatorname{grad} \bar{f}(x)$, obtemos $\mathrm{P}_x \operatorname{grad} \bar{f}(x) = \operatorname{grad} \bar{f}(x) - \mathrm{P}_x^{\perp} \operatorname{grad} \bar{f}(x)$. Segue que

$$\langle \mathrm{P}_x \operatorname{grad} \bar{f}(x), \xi \rangle = \langle \operatorname{grad} \bar{f}(x) - \mathrm{P}_x^{\perp} \operatorname{grad} \bar{f}(x), \xi \rangle,$$

onde $\xi \in T_x\mathcal{M}$. Usando a bilinearidade do produto interno, obtemos

$$\langle \mathrm{P}_x \operatorname{grad} \bar{f}(x), \xi \rangle = \langle \operatorname{grad} \bar{f}(x), \xi \rangle - \langle \mathrm{P}_x^{\perp} \operatorname{grad} \bar{f}(x), \xi \rangle.$$

Sabemos que $\langle \mathrm{P}_x^{\perp} \operatorname{grad} \bar{f}(x), \xi \rangle = 0$. Portanto,

$$\langle \mathrm{P}_x \operatorname{grad} \bar{f}(x), \xi \rangle = \langle \operatorname{grad} \bar{f}(x), \xi \rangle = D\bar{f}(x)[\xi] = Df(x)[\xi].$$

□

O Teorema 3 é particularmente útil porque permite reutilizar cálculos feitos no espaço euclidiano, que geralmente são mais simples, ao invés de computar diretamente o gradiente na subvariedade, o que pode ser mais trabalhoso.

Assim como o gradiente Riemanniano difere do gradiente euclidiano, a Hessiana Riemanniana apresenta características próprias, ajustando-se à geometria da variedade. A seguir, apresentam-se algumas definições e, posteriormente, o principal resultado em relação a hessiana que será usado nesse trabalho.

Definição 26. Seja $\mathfrak{X}(\mathcal{M})$ o conjunto dos campos vetoriais suaves definidos em $\mathcal{M}$. Uma *conexão afim* ∇ em uma variedade $\mathcal{M}$ é uma aplicação bilinear

$$\nabla : \mathfrak{X}(\mathcal{M}) \times \mathfrak{X}(\mathcal{M}) \longrightarrow \mathfrak{X}(\mathcal{M}), \quad (\eta, \xi) \mapsto \nabla_\eta \xi,$$

que satisfaz as seguintes propriedades:
i) $\nabla_{f\eta+g\mathcal{X}}\xi = f\nabla_\eta\xi + g\nabla_{\mathcal{X}}\xi$, para todo $\eta, \mathcal{X}, \xi \in \mathfrak{X}(\mathcal{M})$ e $f, g \in \mathfrak{F}(\mathcal{M})$;
ii) $\nabla_\eta(a\xi + b\zeta) = a\nabla_\eta\xi + b\nabla_\eta\zeta$, para todo $\eta, \xi, \zeta \in \mathfrak{X}(\mathcal{M})$ e $a, b \in \mathbb{R}$;
iii) $\nabla_\eta(f\xi) = (\eta f)\xi + f\nabla_\eta\xi$, para todo $\eta, \xi \in \mathfrak{X}(\mathcal{M})$ e $f \in \mathfrak{F}(\mathcal{M})$.

Uma conexão afim acrescenta à variedade informações que vão além de sua estrutura diferenciável, tornando possível uma análise mais detalhada de suas propriedades [20]. Em geral, uma variedade admite uma infinidade de conexões afins distintas, entretando, algumas delas possuem características específicas que as tornam especialmente adequadas para nosso estudo. Ao equipar a variedade com uma métrica Riemanniana, exigimos duas propriedades adicionais para que a conexão e a métrica interajam de maneira adequada: simetria e compatibilidade com a métrica Riemanniana. Com essas duas propriedades satisfeitas, a conexão afim passa a ser chamada de conexão Riemanniana;
Para definir a simetria de uma conexão afim, precisaremos do conceito colchete de Lie de dois campos vetoriais. Sejam ξ e ζ campos vetoriais da variedade $\mathcal{M}$, cujos domínios se intersectam em um conjunto aberto $\mathcal{U} \subset \mathcal{M}$.
O colchete de Lie $[\xi, \zeta]$ é a função de $\mathfrak{F}(\mathcal{U})$ em si mesma definida por

$$[\xi, \zeta]f = \xi(\zeta f) - \zeta(\xi f),$$

para todo $f \in \mathfrak{F}(\mathcal{U})$.

Definição 27. Seja $\mathcal{M}$ uma variedade Riemanniana. Se para todo $U, V, W \in \mathfrak{F}(\mathcal{M})$, a conexão afim ∇ satisfaz:
i) $[U, V]f = \big(\nabla_U V - \nabla_V U\big)f, \quad \forall f \in F(M)$;
ii) $U\langle V, W\rangle = \langle \nabla_U V, W\rangle + \langle V, \nabla_U W\rangle$.
Então ∇ é uma *conexão Riemanniana*.

Definição 28. Dada uma função $f : \mathcal{M} \to \mathbb{R}$, onde $\mathcal{M}$ é uma variedade Riemanniana, a *Hessiana Riemanniana* de f em um ponto $Y \in \mathcal{M}$ é a aplicação linear $\operatorname{Hess} f(Y) : T_Y\mathcal{M} \to T_Y\mathcal{M}$, definida por

$$\operatorname{Hess} f(Y)[\xi] = \nabla_\xi \operatorname{grad} f(Y),$$

para todo $\xi \in T_Y\mathcal{M}$, onde ∇ é uma conexão Riemanniana em $\mathcal{M}$.

Teorema 4. Sejam $\mathcal{M}$ uma subvariedade Riemanniana de um espaço Euclidiano, $f : \mathcal{M} \to \mathbb{R}$ uma função suave. Se G é um um campo vetorial suave definido em uma vizinhança de $\mathcal{M}$, tal que $G(Y) = \operatorname{grad} f(Y)$ para todo $Y \in \mathcal{M}$, então, para todo $Y \in \mathcal{M}$ e $\xi \in T_Y\mathcal{M}$, vale

$$\operatorname{Hess} f(Y)[\xi] \;=\; \mathrm{P}_Y \; DG(Y)[\xi],$$

onde P_Y denota a projeção ortogonal de $DG(Y)[\xi]$ no espaço tangente $T_Y\mathcal{M}$.

Neste estudo, é conveniente calcularmos a hessiana Riemanniana a partir do resultado apresentado no Teorema 4. Para a demonstração de tal resultado, confira [19, p.56].
Em geral, a direção usada no método de Newton, obtida como solução da equação (2.13), não é necessariamente uma direção de descida da função objetivo. De fato, temos

$$Df(Y_k)[\xi_k] = \langle \operatorname{grad} f(Y_k), \xi_k \rangle = -\langle \operatorname{grad} f(Y_k), (\operatorname{Hess} f(Y_k)^{-1} \operatorname{grad} f(Y_k) \rangle,$$

e não se pode garantir, em geral, que esse valor seja negativo. Uma condição suficiente para assegurar que ξ_k seja uma direção de descida é que o operador $\operatorname{Hess} f(Y_k)$ seja definido positivo, ou seja, $\langle \xi_k, \operatorname{Hess} f(Y_k)[\xi_k] \rangle > 0$ para todo $\xi_k \neq 0$.
A seguir, apresenta-se o algoritmo do Método de Newton Riemanniano.

Algoritmo 2: Método de Newton Riemanniano

1 Escolha uma retração R sobre $\mathcal{M}$, $f : \mathcal{M} \to \mathbb{R}$, uma conexão Riemanniana ∇ de $\mathcal{M}$, um ponto inicial $Y_0 \in \mathcal{M}$ e $k = 0$.
2 Calcule $\xi_k \in T_Y\mathcal{M}$ como a solução da equação de Newton

$$\operatorname{Hess} f(Y_k)\xi_k = -\operatorname{grad} f(Y_k),$$

onde $\operatorname{Hess} f(Y_k)\xi_k := \nabla_{\xi_k} \operatorname{grad} f$.
3 Defina $Y_k := \mathrm{R}_{Y_k}(\xi_k)$.
4 Tome $k := k + 1$ e retorne ao passo 2.

O Algoritmo 2, correspondente ao Método de Newton Riemanniano para funções reais definidas em variedades, pode ser interpretado como um caso particular do seguinte Algoritmo.

Algoritmo 3: Método de Newton Riemanniano para campos vetoriais

1 Escolha uma retração R de $\mathcal{M}$, um ponto inicial $Y_0 \in \mathcal{M}$, um campo vetorial suave ξ, uma conexão afim ∇ de $\mathcal{M}$ e $k = 0$.
2 Calcule $\eta_k \in T_Y\mathcal{M}$ como a solução da equação de Newton

$$J(Y_k)\eta_k = -\xi_{Y_k}, \tag{2.16}$$

onde $J(Y) : T_Y\mathcal{M} \to T_Y\mathcal{M}$ é dado por $J(Y_k)\xi_k := \nabla_{\xi_k}\xi$.
3 Defina $Y_k := \mathrm{R}_{Y_k}(\eta_k)$.
4 Tome $k := k + 1$ e retorne ao passo 2.

O Algoritmo 3 trata do problema geral de encontrar zeros de campos vetoriais em variedades. No caso particular em que esse campo é o gradiente Riemanniano de uma função $f : \mathcal{M} \to \mathbb{R}$, ou seja, tomando $\xi_{Y_k} = \operatorname{grad} f(Y_k)$, a formulação do Algoritmo 3 reduz-se naturalmente à do Algoritmo 2.
A característica de convergência local do Método de Newton é preservada no contexto de variedades diferenciáveis. Em outras palavras, se $f : \mathcal{M} \to \mathbb{R}$ é a função a ser minimi-

zada, onde $\mathcal{M}$ é uma variedade Riemanniana, o ponto inicial $Y_0 \in \mathcal{M}$ deve ser escolhido suficientemente próximo do minimizador Y_* de f. Nessas condições, o método preserva sua rápida convergência, assim como no caso euclidiano.
A seguir, apresentam-se as definições de convergência superlinear e quadrática adaptadas ao contexto de variedades, bem como o teorema de convergência local do Método de Newton Riemanniano, retirado de [20, p.114], que formaliza essas propriedades.

Definição 29. Seja $\mathcal{M}$ uma variedade, (Y_k) uma sequência em $\mathcal{M}$ que converge para $Y_* \in \mathcal{M}$ e $(\mathcal{U}, \varphi)$ uma carta de $\mathcal{M}$, com $Y \in \mathcal{U}$. Se

$$\lim_{k\to\infty} \frac{||\varphi(Y_{k+1}) - \varphi(Y_*)||}{||\varphi(Y_k) - \varphi(Y_*)||} = 0, \tag{2.17}$$

então a sequência (Y_k) converge *superlinearmente* para Y_*. Se existir $c \geqslant 0$ e $K \geqslant 0$ tais que, para todo $k \geqslant K$, tem-se

$$||\varphi(Y_{k+1}) - \varphi(Y_*)|| \leqslant c||\varphi(Y_k) - \varphi(Y_*)||^2, \tag{2.18}$$

então dizemos que a sequência (Y_k) converge *quadráticamente* para Y_*.

Observação 11. Uma sequência de pontos (Y_k) em uma variedade $\mathcal{M}$ é dita convergente se existe uma carta $(\mathcal{U}, \varphi)$ de $\mathcal{M}$, um ponto $Y_* \in \mathcal{M}$ e $K > 0$ tal que $Y_k \in \mathcal{U}$ para todo $k > K$ e a sequência $(\varphi(Y_k))$ converge para $\varphi(Y_*)$.

Para mais detalhes, consulte [20, p.63].

Observação 12. Sejam E e F dois espaços vetoriais normados de dimensão finita, e seja $F : E \to F$ p-vezes continuamente diferenciável em um conjunto convexo aberto $U \subseteq E$, com $x \in U$. Suponha que a diferencial de ordem p, $D^pF : E \to L^p(E; F)$, em que $L^p(E; F)$ denota o espaço das transformações p-lineares contínuas de $E \times \cdots \times E$ em F, seja Lipschitz contínua em x em U com constante de Lipschitz α (usando a norma induzida em $L^p(E; F)$). Então, para qualquer $x + h \in U$, vale

$$\left\| F(x+h) - F(x) - \frac{1}{1!}DF(x)[h] - \cdots - \frac{1}{p!}D^pF(x)[h, \ldots, h] \right\| \leqslant \frac{\alpha}{(p+1)!}\|h\|^{p+1}.$$

Em particular, para $p = 1$, isto é, F continuamente diferenciável com diferencial Lipschitz contínua, temos

$$\|F(x+h) - F(x) - DF(x)[h]\| \leqslant \frac{\alpha}{2}\|h\|^2.$$

Para mais detalhes a respeito da observação 12, consulte [20, p.198].

Teorema 5 (Convergência Local do Método de Newton em Variedades). Sob as mesmas hipóteses e notações do Algoritmo 3, assumimos que existe $Y_* \in \mathcal{M}$ tal que $\xi_{Y_*} = 0$ e $\mathrm{J}(Y_*)^{-1}$ existe. Então, existe um aberto $\mathcal{U}$, onde $Y_* \in \mathcal{U}$, tal que, para todo $Y_0 \in \mathcal{U}$ o Algoritmo 3 gera uma sequência (Y_k) que converge superlinearmente para Y_*.

Demonstração: Seja $(\mathcal{U}, \varphi)$ uma carta de $\mathcal{M}$, com $Y_* \in \mathcal{U}$ e $\varphi : \mathcal{U} \to \varphi(\mathcal{U}) \subset \mathbb{R}^m$. Para mostrar que a sequência (Y_k) converge para Y_*, é suficiente mostrar que a sequência $(\varphi(Y_k))$ converge para $\varphi(Y_*)$, conforme aponta a *Observação* 11. Primeiramente, observamos que $D\varphi(Y_k)[\xi] : T_{Y_k}\mathcal{M} \to \mathbb{R}^m$. Assim, cada vetor tangente $\xi_{Y_k} \in T_{Y_k}\mathcal{M}$ passa a ser atribuído a um vetor no $\mathbb{R}^n$. De forma análoga, a Jacobiana $J(Y_k) : T_{Y_k}\mathcal{M} \to T_{Y_k}\mathcal{M}$ pode ser representada em coordenadas por $\hat{J}(\hat{Y}_k) := (D\varphi(Y_k)) \circ J(Y_k) \circ D\varphi(Y_k))^{-1}$, que é uma aplicação de $\mathbb{R}^m$ em $\mathbb{R}^m$. Note ainda que $\hat{J}(\hat{Y}_k)$ é linear, pois D φ é uma derivação. Para simplificar, adotaremos as seguintes notações: $\hat{Y}_k = \varphi(Y_k)$, $\hat{\xi}_{\hat{Y}_k} = D\varphi(Y_k)[\xi]$ e $\hat{\mathrm{R}}_{\hat{Y}}\hat{\zeta} = \varphi(\mathrm{R}_Y\,\zeta)$. A atualização de pontos da sequência gerada pelo Algoritmo 3 é então dada por

$$\hat{Y}_{k+1} = \hat{\mathrm{R}}_{\hat{Y}_k}(\hat{\eta}_k) = \hat{\mathrm{R}}_{\hat{Y}_k}(-\hat{\mathrm{J}}(\hat{\mathrm{Y}}_\mathrm{k})^{-1}\hat{\xi}_{\hat{Y}_k}), \tag{2.19}$$

enquanto o método de Newton euclidiano aplicado a função $\hat{\xi} : \mathbb{R}^m \to \mathbb{R}^m$ é dado por

$$\hat{Y}_{k+1} = \hat{Y}_k + (-D\hat{\xi}(\hat{Y}_k)^{-1}\hat{\xi}_{\hat{Y}_k}). \tag{2.20}$$

A estratégia nesta prova é mostrar que (2.19) está suficientemente próximo de (2.20), para que o resultado de convergência superlinear do método de Newton apresentado no Teorema 1 seja preservado.

Sabemos que ξ é um campo vetorial suave. Assim, é correto afirmar que $J(Y) = \nabla_\eta\xi$, onde $\eta \in T_Y\mathcal{M}$, também é suave. Pela suavidade de φ e $D\varphi$, $\hat{J}$ também é suave, pois a composição de aplicações suaves é suave. Logo, existem $r_J > 0$ e $\gamma_J > 0$ tais que

$$||\hat{J}(\hat{Y}) - \hat{J}(\hat{Z})|| \leqslant \gamma_J||\hat{Y} - \hat{Z}||,$$

para todo $\hat{Y}, \hat{Z} \in \mathrm{B}_{r_J}(\hat{Y}_*) = \{\hat{Y} \in \mathbb{R}^n; ||\hat{Y} - \hat{Y}_*|| < r_J\}$.

Seja $\beta := ||\hat{J}(\hat{Y}_*)^{-1}||$ e

$$\epsilon = \min\left\{r_J, \frac{1}{2\beta\gamma_J}\right\}.$$

Tome $\hat{Y}_k \in \mathrm{B}_\epsilon(\hat{\mathrm{Y}}_*)$ e obtenha

$$||\hat{J}(\hat{Y}_k) - \hat{J}(\hat{Y}_*)|| \leqslant \gamma_J||\hat{Y}_k - \hat{Y}_*||.$$

Note que

$$||\hat{J}(\hat{Y}_*)^{-1}(\hat{J}(\hat{Y}_k) - \hat{J}(\hat{Y}_*))|| \leqslant ||\hat{J}(\hat{Y}_*)||^{-1}||(\hat{J}(\hat{Y}_k) - \hat{J}(\hat{Y}_*))||.$$

Assim, obtemos

$$\begin{aligned} ||\hat{J}(\hat{Y}_*)^{-1}(\hat{J}(\hat{Y}_k) - \hat{J}(\hat{Y}_*))|| &\leqslant \beta\gamma_J||\hat{Y}_k - \hat{Y}_*|| \\ &\leqslant \beta\gamma_J\epsilon \leqslant \frac{1}{2}. \end{aligned}$$

Segue do Lema 1 que $\hat{J}(\hat{Y}_k)$ é não singular e

$$||\hat{J}(\hat{Y}_k)^{-1}|| \leqslant \frac{||\hat{J}(\hat{Y}_k)^{-1}||}{1 - ||\hat{J}(\hat{Y}_*)^{-1}(\hat{J}(\hat{Y}_k) - \hat{J}(\hat{Y}_*))||} \leqslant 2||\hat{J}(\hat{Y}_*)^{-1}|| \leqslant 2\beta. \tag{2.21}$$

Por conseguinte, para todo $\hat{Y}_k \in \mathrm{B}_\epsilon(\hat{Y}_*)$, a direção de Newton $\hat{\xi}_k := \hat{\mathrm{J}}(\hat{Y}_k)^{-1}\hat{\xi}_{\hat{Y}_k}$ está bem definida.
Sabemos que $\hat{\eta_k} = \hat{J}(\hat{Y}_k)^{-1}\hat{\xi}_{\hat{Y}_k}$. Como $\hat{\xi}$ é suave e $\hat{\xi}_{\hat{Y}_*} = 0$, existe $\gamma > 0$ tal que

$$||\hat{\xi}_{\hat{Y}_k} - \hat{\xi}_{\hat{Y}_*}|| = ||\hat{\xi}_{\hat{Y}_k}|| \leqslant \gamma||\hat{Y}_k - \hat{Y}_*||. \tag{2.22}$$

Usando (2.21), (2.22), para $\hat{Y}_k$ suficientemente próximo de $\hat{Y}_k$, é válido que

$$||\hat{\eta_k}|| = ||\hat{J}(\hat{Y}_k)^{-1}\hat{\xi}_{\hat{Y}_k}|| \leqslant ||\hat{J}(\hat{Y}_k)^{-1}||\ ||\hat{\xi}_{\hat{Y}_k}|| \leqslant 2\beta\gamma||\hat{Y}_k - \hat{Y}_*|| =: c||\hat{Y}_k - \hat{Y}_*||. \tag{2.23}$$

Note que $\hat{\mathrm{R}}_{\hat{Y}_k} : \mathbb{R}^m \to \mathbb{R}^m$. Além disso, $\hat{\mathrm{R}}_{\hat{Y}_k}$ é suave pois R_Y e φ são, por definição, suaves. Assim, usando a expansão de Taylor de segunda ordem para $\hat{\mathrm{R}}_{\hat{Y}_k}(\hat{\eta}_k)$ (ver [20, p.198]), obtemos

$$\hat{\mathrm{R}}_{\hat{Y}_k}(\hat{\eta}_k) = \hat{\mathrm{R}}_{\hat{Y}_k}(0) + D\hat{\mathrm{R}}_{\hat{Y}_k}(0)[\hat{\eta}_k] + \frac{1}{2}D^2\hat{\mathrm{R}}_{\hat{Y}_k}(0)[\hat{\eta}_k, \hat{\eta}_k] + r(\hat{\eta}_k).$$

Por definição, sabe-se que $\hat{\mathrm{R}}_{\hat{Y}_k}(0) = \hat{Y}_k$. Logo,

$$\hat{\mathrm{R}}_{\hat{Y}_k}(\hat{\eta}_k) = \hat{Y}_k + \hat{\eta}_k + \frac{1}{2}D^2\hat{\mathrm{R}}_{\hat{Y}_k}(0)[\hat{\eta}_k, \hat{\eta}_k] \Rightarrow \hat{\mathrm{R}}_{\hat{Y}_k}(\hat{\eta}_k) - (\hat{Y}_k + \hat{\eta}_k) = \frac{1}{2}D^2\hat{\mathrm{R}}_{\hat{Y}_k}(0)[\hat{\eta}_k, \hat{\eta}_k].$$

Segue que

$$\|\hat{\mathrm{R}}_{\hat{Y}_k}(\hat{\eta}_k) - (\hat{Y}_k + \hat{\eta}_k)\| \leqslant \frac{1}{2}\|D^2\hat{\mathrm{R}}_{\hat{Y}_k}(0)\|\|\hat{\eta}_k\|^2.$$

Devido à continuidade de $D^2\hat{\mathrm{R}}_{\hat{Y}_k}$ em uma vizinhança de 0, existe uma constante $K > 0$ tal que $\|D^2\hat{\mathrm{R}}_{\hat{Y}_k}(0)\| \leqslant K$. Definindo $c = K/2$, obtemos

$$\|\hat{\mathrm{R}}_{\hat{Y}_k}(\hat{\eta}_k) - (\hat{Y}_k + \hat{\eta}_k)\| \leqslant c\|\hat{\eta}_k\|^2. \tag{2.24}$$

Usando (2.24) e (2.23), temos

$$||\,\hat{\mathrm{R}}_{\hat{Y}_k}\,\hat{\eta}_k - (\hat{Y}_k + \hat{\eta}_k)|| \leqslant \gamma_R||\hat{Y}_k - \hat{Y}_*||^2,$$

para algum $\gamma_R > 0$. Defina $\hat{\Gamma}_{\hat{Y},\hat{\xi}}$ por $\hat{\Gamma}_{\hat{Y},\hat{\xi}}\hat{\zeta} = \hat{J}(\hat{Y})\hat{\zeta} - D\hat{\xi}(\hat{Y})[\hat{\zeta}]$. Note que $\hat{\Gamma}_{\hat{Y},\hat{\xi}}$ é um operador linear. Novamente, pela suavidade de $\hat{\Gamma}_{\hat{Y},\hat{\xi}}$, existem r_Γ e $\gamma_\Gamma > 0$ tais que

$$||\hat{\Gamma}_{\hat{Y},\hat{\xi}} - \hat{\Gamma}_{\hat{Z},\hat{\xi}}|| \leqslant \gamma_\Gamma||\hat{Y} - \hat{Z}||,$$

para todo $\hat{Y}, \hat{Z} \in B_{r_\Gamma}(\hat{Y}_*)$. Em particular,

$$||\hat{\Gamma}_{\hat{Y},\hat{\xi}}|| \leqslant \gamma_\Gamma ||\hat{Y}_k - \hat{Y}_*||,$$

para todo $\hat{Y}_k \in B_\epsilon(\hat{Y}_*)$. Ademais, para todo $\hat{Y}, \hat{Z} \in B_{\min\{r_J, r_\Gamma\}}(\hat{Y}_*)$, temos

$$\begin{aligned}\|D\hat{\xi}(\hat{Y}) - D\hat{\xi}(\hat{Z}) - (\hat{\Gamma}_{\hat{Y},\hat{\xi}} - \hat{\Gamma}_{\hat{Z},\hat{\xi}})\| &\leqslant \|D\hat{\xi}(\hat{Y}) + \hat{\Gamma}_{\hat{Y},\hat{\xi}} - (D\hat{\xi}(\hat{Z}) + \hat{\Gamma}_{\hat{Z},\hat{\xi}})\| \\ &= \|\hat{J}(\hat{Y}) - \hat{J}(\hat{Z})\| \leqslant \gamma_J \|\hat{Y} - \hat{Z}\|.\end{aligned}$$

De (2.19), obtemos

$$\hat{Y}_{k+1} - \hat{Y}_* = \hat{R}_{\hat{Y}_k}(-\hat{J}(\hat{Y}_k)^{-1}\hat{\xi}_{\hat{Y}_k}) - \hat{Y}_*.$$

Segue que

$$\begin{aligned}\|\hat{Y}_{k+1} - \hat{Y}_*\| &\leqslant \|\hat{Y}_k - \hat{J}(\hat{Y}_k)^{-1}\hat{\xi}_{\hat{Y}_k} - \hat{Y}_*\| + \gamma_R \|\hat{Y}_k - \hat{Y}_*\|^2 \\ &\leqslant \|\hat{J}(\hat{Y}_k)^{-1}(\hat{\xi}_{\hat{Y}_*} - \hat{\xi}_{\hat{Y}_k} - \hat{J}(\hat{Y}_k)(\hat{Y}_* - \hat{Y}_k))\| + \|\hat{J}(\hat{Y}_k)^{-1}\hat{\Gamma}_{\hat{Y}_k,\hat{\xi}}(\hat{Y}_* - \hat{Y}_k)\| \\ &\quad + \gamma_R \|\hat{Y}_k - \hat{Y}_*\|^2 \\ &\leqslant 2\beta\frac{1}{2}(\gamma_J + \gamma_\Gamma)\|\hat{Y}_k - \hat{Y}_*\|^2 + 2\beta\gamma_\Gamma\|\hat{Y}_k - \hat{Y}_*\|^2 + \gamma_R\|\hat{Y}_k - \hat{Y}_*\|^2,\end{aligned}$$

sempre que $\|\hat{Y}_k - \hat{Y}_*\| \leqslant \min\{\epsilon, r_\Gamma, r_R\}$, onde usamos a Observação 12. □

Capítulo 3

Problema de Diagonalização conjunta de matrizes

Neste capítulo, apresentamos o problema de diagonalização conjunta de um conjunto de matrizes simétricas, formulado como a minimização de uma função objetivo na variedade de Stiefel. Para resolver esse problema, reescrevemos a equação de Newton por meio de uma estratégia de vetorização que permitiu representar a Hessiana da função objetivo como uma transformação linear. Essa reformulação conduz a um sistema linear cuja solução determina a direção de Newton, dando origem à versão vetorizada do método empregada na resolução do problema.

3.1 Introdução

Se uma matriz $A \in \mathbb{R}^{n\times n}$ possui n autovalores distintos, existe uma matriz invertível $P \in \mathbb{R}^{n\times n}$ tal que

$$P^{-1}AP = D,$$

onde D é uma matriz diagonal cujos elementos diagonais são exatamente os autovalores de A. No caso particular em que A é simétrica, existe de uma matriz ortogonal $Q \in \mathbb{R}^{n\times n}$ satisfazendo

$$Q^T AQ = D.$$

Assim, diagonalizar uma matriz A significa encontrar uma matriz Y (invertível ou ortogonal, dependendo do contexto) tal que a matriz transformada $Y^T AY$ seja diagonal. Em outras palavras, buscamos uma matriz Y que faça com que todos os elementos fora da diagonal seja igual a zero, [5].

O problema de interesse neste trabalho consiste em diagonalizar simultaneamente um conjunto de matrizes simétricas $A_1, \ldots, A_K \in \mathbb{R}^{n\times n}$. Deseja-se encontrar uma matriz $Y \in \mathrm{St}(p, n)$ tal que

$$Y^T A_l Y = D_l, \quad l = 1, \ldots, K, \tag{3.1}$$

onde cada D_l é uma matriz diagonal.

Quando as matrizes A_l comutam duas a duas, isto é, $A_iA_j = A_jA_i$ para todo i, j, existe uma matriz Y que as diagonaliza simultaneamente [22]. No entanto, a exigência de comutatividade é bastante restritiva e nem sempre é verificada em problemas reais. Assim, quando as matrizes não comutam, busca-se uma diagonalização aproximada, isto é, procura-se uma matriz Y que torne cada matriz Y^TA_lY o mais diagonal possível, tornando ao máximo os elementos fora da diagonal próximos de zero.

Essa ideia conduz naturalmente à formulação do problema como uma tarefa de otimização. Diversos métodos para diagonalização conjunta aproximada foram propostos na literatura, distinguindo-se pela escolha da função objetivo, pelas estratégias de otimização empregadas e pelas diferentes condições impostas ao problema [22]. A formulação adotada neste trabalho segue [12] e consiste em um problema de otimização restrito à variedade de Stiefel, cuja solução fornece uma matriz $Y_* \in \mathrm{St}(p, n)$ que diagonaliza, de maneira aproximada, o conjunto de matrizes simétricas A_l.

Seja $\mathrm{St}(p, n) = \{\, Y \in \mathbb{R}^{n\times p}; Y^TY = I_p \,\}$, onde I_p é a matriz identidade de ordem p e $p \leqslant n$. Consideremos a função $g : \mathbb{R}^{n\times p} \to \mathbb{R}$, definida por

$$g(Y) = \sum_{l=1}^{N} \left\|\mathrm{off}\left(Y^TA_lY\right)\right\|_F^2, \qquad Y \in \mathrm{St}(p, n), \tag{3.2}$$

em que $A_l \in \mathbb{R}^{n\times n}$, $l = 1, \ldots, N$, são matrizes simétricas, $\mathrm{off}(X)$ denota a parte fora da diagonal da matriz, e $\|\cdot\|_F$ é a norma de Frobenius. Para mais detalhes, consulte [12]. Note que, quanto maiores forem os elementos fora da diagonal das matrizes Y^TA_lY, maior é o valor de $g(Y)$. Logo, resolver o problema de diagonalização conjunta de matrizes consiste em minimizar a função g.

Em vez de avaliar diretamente os termos fora da diagonal, pode-se considerar apenas os elementos diagonais das matrizes Y^TA_lY. Isso conduz outra formulação do mesmo problema, na qual se busca uma matriz $Y \in \mathrm{St(p, n)}$ que maximize a função $\bar{g} : \mathbb{R}^{n\times p} \to \mathbb{R}$, dada por

$$\bar{g}(Y) = \sum_{l=1}^{N} || \,\mathrm{diag}\left(Y^TA_lY\right) ||_F^2,$$

em que $A_l \in \mathbb{R}^{n\times n}$, com $l = 1, \ldots, N$ são matrizes simétricas, diag(X) denota a matriz diagonal cujos elementos coincidem com a diagonal principal de X e $|| \cdot ||_F$ é a norma de Frobenius. Equivalentemente, desejamos minimizar $\bar{f} : \mathbb{R}^{n\times p} \to \mathbb{R}$, dada por

$$\bar{f}(Y) = -\sum_{l=1}^{N} || \,\mathrm{diag}\left(Y^TA_lY\right) ||_F^2, \quad Y \in \mathrm{St}(p, n). \tag{3.3}$$

Observe que o problema definido em (3.3) constitui um problema de otimização restrito. Uma vez que $\mathrm{St}(p, n)$ é uma variedade diferenciável, é possível formular um problema de

otimização irrestrito equivalente. Para estabelecer formalmente esta equivalência, recorremos ao seguinte resultado.

Observação 13. Seja $F : \mathcal{M} \to \mathcal{N}$ um mapa suave entre variedades diferenciáveis e seja $Z \in \mathcal{N}$ tal que $DF(Y) : T_Y\mathcal{M} \to T_Z\mathcal{N}$ é sobrejetora para todo $Y \in F^{-1}(Z)$. Então o conjunto $F^{-1}(Z) = \{Y \in \mathcal{M} \mid F(Y) = Z\}$ é uma subvariedade diferenciável $\mathcal{M}$. Para mais detalhes, veja [20, p.26].

Proposição 3. O conjunto $\mathrm{St}(p, n) = \{Y \in \mathbb{R}^{n\times p}; Y^TY = I_p\}$ é uma subvariedade diferenciável de $\mathbb{R}^{n\times p}$.

Demonstração: Considere a função suave

$$F : \mathbb{R}^{n\times p} \to S_{\mathrm{sym}}(p), \quad \text{dada por} \quad F(X) = X^TX - I_p,$$

onde $\mathrm{S_sym}$ denota o conjunto de todas as matrizes reais $p \times p$ simétricas.
Note que, para todo $X \in \mathrm{St}(p, n)$ temos $X = F^{-1}(0)$, pois X satisfaz $X^TX - I_p = 0$, logo $F(X) = 0$. Além disso, a derivada de F em X, na direção $Z \in \mathbb{R}^{n\times p}$, pode ser obtida observando que

$$\begin{aligned} F(X + Z) &= (X + Z)^T(X + Z) - I_p \\ &= X^TX - I_p + (X^TZ + Z^TX) + Z^TZ \\ &= X^TZ + Z^TX + Z^TZ. \end{aligned}$$

Segue que

$$DF(X)[Z] = X^TZ + Z^TX.$$

Precisamos mostrar que $DF(X)$ é sobrejetora para todo $X \in \mathrm{St}(p, n)$. Em outras palavras, queremos provar que, dado qualquer $W \in \mathrm{S_sym}(p)$, existe Z tal que $X^TZ + Z^TX = W$. Escolhendo $Z = \frac{1}{2}XW$, temos

$$\begin{aligned} DF(X)[Z] &= X^T\left(\frac{1}{2}XW\right) + \left(\frac{1}{2}XW\right)^T X \\ &= \frac{1}{2}\left(X^TXW\right) + \frac{1}{2}\left(W^TX^TX\right) \\ &= \frac{1}{2}W + \frac{1}{2}W = W, \end{aligned}$$

onde utilizamos $X^TX = I_p$ e $W^T = W$.
Portanto, $DF(X)$ é sobrejetora para todo $X \in \mathrm{St}(p, n)$ e, pela Observação 13, segue-se que $\mathrm{St}(p, n)$ é uma subvariedade diferenciável de $\mathbb{R}^{n\times p}$. □
Para mais detalhes a respeito do resultado apresentado na Proposição 3, consulte [20, p.26].

A partir da estrutura diferenciável de $\mathrm{St}(p,n)$, é possível reformular o problema (3.3) como um problema de otimização irrestrito. Em particular, consideremos a minimização da função $f : \mathrm{St}(p,n) \to \mathbb{R}$ dada por

$$f(Y) = -\sum_{l=1}^{N} || \operatorname{diag}\left(Y^T A_l Y\right) ||_F^2. \tag{3.4}$$

O problema definido em (3.4) é conhecido como o problema de diagonalização conjunta sobre a variedade de $\mathrm{St}(p,n)$. Este problema surge em diferentes áreas, incluindo a análise de sinais multivariados, que será discutido com mais detalhes adiante.
Nosso objetivo é determinar a solução do problema irrestrito (3.4) utilizando o Método de Newton Riemanniano. Para isso, é necessário calcular a direção de Newton definida pela equação (2.13), fazendo uso de resultados discutidos anteriormente.
Em particular, pelo Teorema 3, o gradiente Riemanniano da função $f : \mathrm{St}(p,n) \to \mathbb{R}$ é dado por

$$\operatorname{grad} f(Y) = \mathrm{P}_{\mathrm{Y}}(\operatorname{grad} \bar{f}), \tag{3.5}$$

onde P_{Y} é a projeção ortogonal sobre o espaço tangente da variedade Stiefel em Y, e $\operatorname{grad} \bar{f}$ é o gradiente euclidiano da função $\bar{f} : \mathbb{R}^{n\times p} \to \mathbb{R}$, conforme definido em (3.3).
Além disso, sabemos que a Hessiana Riemanniana de f no ponto Y aplicada a um elemento do espaço tangente

$$\operatorname{Hess} f(Y)[\xi] = \mathrm{P}_{\mathrm{Y}}(D(\operatorname{grad} f(Y))[\xi], \tag{3.6}$$

onde $D(\operatorname{grad} f(Y))[\xi]$ é a derivada direcional do gradiente Riemanniano na direção ξ.
Notemos que, a partir de (1.8) e (3.5), obtemos

$$\operatorname{grad} f(Y) = \operatorname{grad} \bar{f}(Y) - Y \operatorname{sym}(Y^T \operatorname{grad} \bar{f}(Y)).$$

Como D é um operador linear, obtemos

$$D(\operatorname{grad} f(Y))[\xi] = D(\operatorname{grad} \bar{f}(Y))[\xi] - D(Y \operatorname{sym}(Y^T \operatorname{grad} \bar{f}(Y)))[\xi].$$

Aplicando a regra do produto, segue que

$$D(\operatorname{grad} f(Y))[\xi] = D(\operatorname{grad} \bar{f}(Y))[\xi] - \xi \operatorname{sym}(Y^T \operatorname{grad} \bar{f}(Y)) - D(\operatorname{sym}(Y^T \operatorname{grad} \bar{f}(Y)))[\xi].$$

Seja $M = D(\operatorname{sym}(Y^T \operatorname{grad} \bar{f}(Y)))[\xi]$, podemos escrever

$$D(\operatorname{grad} f(Y))[\xi] = D(\operatorname{grad} \bar{f}(Y))[\xi] - \xi \operatorname{sym}(Y^T \operatorname{grad} \bar{f}(Y)) - YM. \tag{3.7}$$

Como P_Y é linear, de (3.6) e (3.7), temos

$$\operatorname{Hess} f(Y)[\xi] = \mathrm{P}_Y(D(\operatorname{grad} \bar{f}(Y))[\xi] - \xi \operatorname{sym}(Y^T \operatorname{grad} \bar{f}(Y)) - \mathrm{P}_Y(YM).$$

Além disso, sabendo que $Y \in \mathrm{St}(p,n)$ e M é simétrica, é válido que

$$\mathrm{P}_Y(YM) = YM - Y \operatorname{sym}(Y^T YM) = YM - Y \operatorname{sym}(M) = 0.$$

Portanto, concluímos que a Hessiana Riemanniana pode ser calculada como

$$\operatorname{Hess} f(Y)[\xi] = \mathrm{P}_Y(D(\operatorname{grad} \bar{f}(Y))[\xi] - \xi \operatorname{sym}(Y^T \operatorname{grad} \bar{f}(Y)). \tag{3.8}$$

Proposição 4. O gradiente euclidiano de $\bar{f}$ é dado por

$$\operatorname{grad} \bar{f}(Y) = -4 \sum_{l=1}^{N} A_l Y \operatorname{diag}(Y^T A_l Y), \tag{3.9}$$

onde $A_l, l = 1, \ldots, N$ são simétricas.

Demonstração: Seja $g_l : \mathbb{R}^{n \times p} \to \mathbb{R}$ uma função tal que

$$g_l(Y) = \| \operatorname{diag}(Y^T A_l Y) \|_F^2,$$

e $Y = [y_1, \ldots, y_p] \in \mathbb{R}^{n \times p}$, com $y_i \in \mathbb{R}^n$ denotando a i-ésima coluna de Y. O produto $Y^T A_l Y \in \mathbb{R}^{p \times p}$ tem entradas dadas por

$$(Y^T A_l Y)_{ij} = y_i^T A_l y_j, \quad i, j = 1, \ldots, p.$$

Em particular, a diagonal de $Y^T A_l Y$ é

$$\operatorname{diag}(Y^T A_l Y) = \begin{bmatrix} y_1^T A_l y_1 \\ y_2^T A_l y_2 \\ \vdots \\ y_p^T A_l y_p \end{bmatrix} \in \mathbb{R}^p.$$

Segue que

$$\| \operatorname{diag}(Y^T A_l Y) \|_F^2 = \sum_{i=1}^{p} (y_i^T A_l y_i)^2.$$

Além disso, para $\Delta \in \mathbb{R}^{n \times p}$ $(\Delta \neq 0)$, onde $\Delta = [\delta_1, \delta_2, \ldots, \delta_p]$ $(\delta_i \in \mathbb{R}^n, i = 1, \ldots, p)$, temos

$$g_l(Y + \Delta) = \sum_{i=1}^{p} \left((y_i + \delta_i)^T A_l (y_i + \delta_i) \right)^2.$$

Notemos que,

$$\begin{aligned} g_l(Y+\Delta) &= \sum_{i=1}^{p} \left((y_i+\delta_i)^T A_l (y_i+\delta_i)\right)^2 \\ &= \sum_{i=1}^{p} (y_i^T A_l y_i + y_i^T A_l \delta_i + \delta_i^T A_l y_i + \delta_i^T A_l \delta_i)^2 \\ &= \sum_{i=1}^{p} (y_i^T A_l y_i + 2 y_i^T A_l \delta_i + \delta_i^T A_l \delta_i)^2, \\ &= \sum_{i=1}^{p} ((y_i^T A_l y_i)^2 + 4(y_i^T A_l y_i)(y_i^T A_l \delta_i) + r_i(\delta_i)), \end{aligned}$$

onde $r_i(\delta_i) = 4(y_i^T A_l \delta_i)^2 + 2(y_i^T A_l y_i)(\delta_i^T A_l \delta_i) + 4(y_i^T A_l \delta_i)(\delta_i^T A_l \delta_i) + (\delta_i^T A_l \delta_i)^2$. Usando a propriedade $|u^T A v| \leqslant \|u\| \, \|A\| \, \|v\|$, obtemos

$$|y_i^T A_l \delta_i| \leqslant \|y_i\| \, \|A_l\| \, \|\delta_i\|, \qquad |\delta_i^T A_l \delta_i| \leqslant \|A_l\| \, \|\delta_i\|^2.$$

Assim, existe uma constante $C > 0$, tal que

$$|r_i(\delta_i)| \leqslant C \|\delta_i\|^2.$$

Tomando $r(\Delta) = \sum_{i=1}^{p} r_i(\delta_i)$ e usando $\|\Delta\|^2 = \sum_{i=1}^{p} \|\delta_i\|^2$, podemos escrever

$$|r(\Delta)| \leqslant C \|\Delta\|^2.$$

Dividindo ambos os lados por $\|\Delta\|$, temos

$$\frac{|r(\Delta)|}{\|\Delta\|} \leqslant C \, \|\Delta\|.$$

Tomando o limite quando $\Delta \to 0$, segue do lado direito que

$$\lim_{\Delta \to 0} C \, \|\Delta\| = 0,$$

e, pelo Teorema do Confronto, concluímos que

$$\lim_{\Delta \to 0} \frac{|r(\Delta)|}{\|\Delta\|} = 0.$$

Assim

$$\begin{aligned} g_l(Y+\Delta) &= \sum_{i=1}^{p} \Big((y_i^T A_l y_i)^2 + 4(y_i^T A_l y_i)(y_i^T A_l \delta_i) \Big) + r(\Delta) \\ &= \sum_{i=1}^{p} (y_i^T A_l y_i)^2 + \sum_{i=1}^{p} 4(y_i^T A_l y_i)(y_i^T A_l \delta_i) + r(\Delta) \end{aligned}$$

Logo,

$$\langle \operatorname{grad} g_l(Y), \Delta \rangle = \sum_{i=1}^{p} 4(y_i^T A_l y_i) y_i^T A_l \delta_i.$$

Note que,

$$\sum_{i=1}^{p} 4(y_i^T A_l y_i)\, y_i^T A_l \delta_i = \sum_{i=1}^{p} \big\langle 4(y_i^T A_l y_i) A_l y_i, \delta_i \big\rangle = \sum_{i=1}^{p} \big\langle 2(A_l + A_l^T) y_i\, (y_i^T A_l y_i), \delta_i \big\rangle.$$

Deste modo, podemos concluir que

$$\operatorname{grad} g_l(Y) = 2(A_l + A_l^T)\, Y\ \operatorname{diag}(Y^T A_l Y).$$

Como $\bar{f}(Y) = -\sum_{l=1}^{N} g_l(Y)$, temos

$$\operatorname{grad} \bar{f}(Y) = -\sum_{l=1}^{N} \operatorname{grad} g_l(Y) = -2 \sum_{l=1}^{N} (A_l + A_l^T) Y \operatorname{diag}(d^{(l)}).$$

Finalmente, usando a simetria de A_l, obtemos

$$\operatorname{grad} \bar{f}(Y) = -4 \sum_{l=1}^{N} A_l Y \operatorname{diag}(Y^T A_l Y). \tag{3.10}$$

□

A Proposição 4 corresponde a um resultado previamente estabelecido em [12].
O cálculo do gradiente euclidiano da função $\bar{f}$ é importante, pois serve como ponto de partida para obter o gradiente Riemanniano, por meio da projeção sobre o espaço tangente da variedade. Para determinar a Hessiana Riemanniana, no entanto, é necessário também determinar $D(\operatorname{grad} \bar{f})(Y)[\xi]$, onde $Y \in \mathcal{M}$ e $\xi \in T_Y \mathcal{M}$.
Note que, ao substituirmos $Y + t\xi$, onde $t \in \mathbb{R}$ e $\xi \in T_Y \operatorname{St}(p, n)$, em (3.9), obtemos

$$\operatorname{grad} \bar{f}(Y + t\xi) = -4 \sum_{l=1}^{N} A_l (Y + t\xi) \operatorname{diag}\big((Y + t\xi)^T A_l (Y + t\xi) \big). \tag{3.11}$$

Usando a fórmula de Taylor, temos

$$(Y+t\xi)^T A_l(Y+t\xi) = Y^T A_l Y + t(Y^T A_l\xi + \xi^T A_l Y) + o(t),$$

onde $\lim_{t\to 0^+} o(t)/t = 0$. Assim, podemos escrever

$$\operatorname{diag}\big((Y+t\xi)^T A_l(Y+t\xi)\big) = \operatorname{diag}(Y^T A_l Y) + t\,\operatorname{diag}(Y^T A_l\xi + \xi^T A_l Y) + o(t).$$

Substituindo em (3.11), obtemos

$$\begin{aligned}\operatorname{grad}\bar{f}(Y+t\xi) = -4\sum_{l=1}^{N}\Big[A_l Y\operatorname{diag}(Y^T A_l Y) + t\big(A_l\xi\operatorname{diag}(Y^T A_l Y)\\ + A_l Y\operatorname{diag}(Y^T A_l\xi + \xi^T A_l Y)\big)\Big] + o(t).\end{aligned} \tag{3.12}$$

Por definição, temos que

$$D(\operatorname{grad}\bar{f})(Y)[\xi] = \lim_{t\to 0}\frac{\operatorname{grad}\bar{f}(Y+t\xi) - \operatorname{grad}\bar{f}(Y)}{t}. \tag{3.13}$$

Substituindo (3.12) e (3.9) em (3.13), obtemos

$$D(\operatorname{grad}\bar{f})(Y)[\xi] = -4\sum_{l=1}^{N}\big(A_l\xi\operatorname{diag}(Y^T A_l Y) + A_l Y\operatorname{diag}(Y^T A_l\xi + \xi^T A_l Y)\big).$$

Como A_l é simétrica, então $\xi^T A_l Y = (Y^T A_l\xi)^T$. Segue que

$$\operatorname{diag}(Y^T A_l\xi + \xi^T A_l Y) = 2\operatorname{diag}(Y^T A_l\xi).$$

Portanto, a derivada direcional do gradiente na direção ξ é

$$D(\operatorname{grad}\bar{f})(Y)[\xi] = -4\sum_{l=1}^{N}\Big(A_l\xi\operatorname{diag}(Y^T A_l Y) + 2A_l Y\operatorname{diag}(Y^T A_l\xi)\Big). \tag{3.14}$$

Assim, usando (3.8), (3.9) e (3.14), obtemos

$$\begin{aligned}\operatorname{Hess} f(Y)[\xi] = -4\sum_{l=1}^{N}\mathrm{P}_Y(A_l\xi\operatorname{diag}(Y^T A_l Y) + 2A_l Y\operatorname{diag}(Y^T A_l\xi) -\\ \xi\operatorname{sym}(Y^T A_l Y\operatorname{diag}(Y^T A_l Y))).\end{aligned}$$

Logo, a direção de Newton para o problema (3.4) é obtida como a solução $\xi \in T_Y\operatorname{St}(p,n)$

da equação

$$
\begin{aligned}
&-4\sum_{l=1}^{N}\mathrm{P}_Y\left(A_l\xi\,\mathrm{diag}(Y^TA_lY)+2A_lY\,\mathrm{diag}(Y^TA_l\xi)-\xi\,\mathrm{sym}(Y^TA_lY\,\mathrm{diag}(Y^TA_lY))\right)\\
&=4\sum_{l=1}^{N}\mathrm{P}_Y\left(A_lY\,\mathrm{diag}(Y^TA_lY)\right).
\end{aligned}
\tag{3.15}
$$

Usando o fato de que P_Y é linear, obtemos

$$
\begin{aligned}
&-4\sum_{l=1}^{N}\mathrm{P}_Y\left(A_l\xi\,\mathrm{diag}(Y^TA_lY)\right)+2\,\mathrm{P}_Y\left(A_lY\,\mathrm{diag}(Y^TA_l\xi)\right)\\
&-\mathrm{P}_Y\left(\xi\,\mathrm{sym}(Y^TA_lY\,\mathrm{diag}(Y^TA_lY))\right)=4\sum_{l=1}^{N}\mathrm{P}_Y\left(A_lY\,\mathrm{diag}(Y^TA_lY)\right).
\end{aligned}
\tag{3.16}
$$

Para a primeira projeção do lado direito da equação (3.16), usando (1.8), teremos

$$
\begin{aligned}
\mathrm{P}_Y\left(A_l\xi\,\mathrm{diag}(Y^TA_lY)\right)&=A_l\xi\,\mathrm{diag}(Y^TA_lY)-Y\,\mathrm{sym}(Y^TA_l\xi\,\mathrm{diag}(Y^TA_lY))\\
&=A_l\xi\,\mathrm{diag}(Y^TA_lY)-Y\frac{1}{2}(Y^TA_l)\xi(\mathrm{diag}(Y^TA_lY))\\
&+Y\frac{1}{2}(\mathrm{diag}(Y^TA_lY))^T\xi^TA_lY.
\end{aligned}
$$

Para a segunda projeção, usaremos as seguintes observações:

Observação 14. Denotaremos por $E_{ij}^{(p\times q)}$ a matriz de dimensão $p\times q$ cujo elemento na posição (i,j) é igual a 1, enquanto todos os demais elementos são iguais a 0.

Exemplo 6. Seja $p=q=3$ e $(i,j)=(2,1)$, temos

$$
E_{21}^{(3\times 3)}=\begin{bmatrix}0&0&0\\1&0&0\\0&0&0\end{bmatrix}.
$$

Observação 15. Sejam $Y\in\mathbb{R}^{n\times p}$, $A_l\in\mathbb{R}^{n\times n}$ simétrica e $\xi\in T_YSt(p,n)$. Então, a matriz $\mathrm{diag}(Y^TA_l\xi)$ pode ser reescrita como

$$
\mathrm{diag}(Y^TA_l\xi)=\sum_{i=1}^{p}(y_i^TA_l\xi)\,E_{ii},
$$

onde $y_i\in\mathbb{R}^n$ denota a i-ésima coluna de Y.

Note que

$$Y^T A_l \xi = \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_p^T \end{bmatrix} A_l \xi = \begin{bmatrix} y_1^T A_l \xi \\ y_2^T A_l \xi \\ \vdots \\ y_p^T A_l \xi \end{bmatrix}.$$

Por definição,

$$\operatorname{diag}(Y^T A_l \xi) = \begin{bmatrix} y_1^T A_l \xi & 0 & \cdots & 0 \\ 0 & y_2^T A_l \xi & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & y_p^T A_l \xi \end{bmatrix}.$$

Seja E_{ii} a matriz $p \times p$ com 1 na posição (i, i) e 0 em todas as outras entradas. Notemos que,

$$\operatorname{diag}(Y^T A_l \xi) = \begin{bmatrix} y_1^T A_l \xi \cdot 1 & 0 & \cdots & 0 \\ 0 & y_2^T A_l \xi \cdot 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & y_p^T A_l \xi \cdot 1 \end{bmatrix} = \begin{bmatrix} y_1^T A_l \xi & 0 & \cdots & 0 \\ 0 & y_2^T A_l \xi & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & y_p^T A_l \xi \end{bmatrix} E_{ii}.$$

Assim, obtemos

$$\operatorname{diag}(Y^T A_l \xi) = \sum_{i=1}^{p} (y_i^T A_l \xi) E_{ii}.$$

□

Sabe-se que

$$\begin{aligned} \mathrm{P}_Y \left(A_l Y \operatorname{diag}(Y^T A_l \xi) \right) &= (A_l Y \operatorname{diag}(Y^T A_l \xi)) - Y \operatorname{sym}(Y^T (A_l Y \operatorname{diag}(Y^T A_l \xi)) \\ &= A_l Y \operatorname{diag}(Y^T A_l \xi) - \frac{1}{2} Y Y^T A_l Y \operatorname{diag}(Y^T A_l \xi) \\ &\quad - \frac{1}{2} Y (\operatorname{diag}(Y^T A_l \xi))^T Y^T A_l Y. \end{aligned}$$

Usando a *Observação 15*, tem-se

$$\begin{aligned} \mathrm{P}_Y \left(A_l Y \operatorname{diag}(Y^T A_l \xi) \right) &= A_l Y \sum_{i=1}^{p} (y_i^T A_l \xi) E_{ii} - \frac{1}{2} Y Y^T A_l Y \sum_{i=1}^{p} (y_i^T A_l \xi) E_{ii} \\ &\quad - \frac{1}{2} Y E_{ii}^T \left(\sum_{i=1}^{p} (y_i^T A_l \xi) \right)^T Y^T A_l Y. \end{aligned}$$

Para a terceira projeção do lado esquerdo da equação(3.16), temos

$$\begin{aligned}\mathrm{P}_Y\left(\xi \operatorname{sym}\left(Y^TA_lY \operatorname{diag}\left(Y^TA_lY\right)\right)\right) &= \xi \operatorname{sym}\left(Y^TA_lY \operatorname{diag}\left(Y^TA_lY\right)\right)\\ &\quad - Y \operatorname{sym}\left(Y^T\xi \operatorname{sym}\left(Y^TA_lY \operatorname{diag}\left(Y^TA_lY\right)\right)\right)\\ &= \xi \operatorname{sym}\left(Y^TA_lY \operatorname{diag}\left(Y^TA_lY\right)\right)\\ &\quad - \tfrac{1}{2}YY^T\xi \operatorname{sym}\left(Y^TA_lY \operatorname{diag}\left(Y^TA_lY\right)\right)\\ &\quad - \tfrac{1}{2}Y \operatorname{sym}\left(Y^TA_lY \operatorname{diag}\left(Y^TA_lY\right)\right)^T\xi^TY.\end{aligned}$$

Para o lado direiro da equação (3.16), temos

$$\mathrm{P}_Y(A_lY \operatorname{diag}(Y^TA_lY)) = A_lY \operatorname{diag}(Y^TA_lY) - Y \operatorname{sym}(A_lY \operatorname{diag}(Y^TA_lY)).$$

Note que, ao substituirmos as projeções em (3.16), obtemos para ξ, uma equação linear da forma

$$\sum_i A_i\,\xi\,B_i + \sum_j C_j\,\xi^T\,D_j = E,$$

onde A_i, B_i, C_j, D_j e E são matrizes conhecidas. Esta é uma generalização da *equação de Sylvester* e, embora linear em ξ, sua resolução é em geral difícil, pois gera alto custo computacional. Para mais detalhes, veja [21, p.111].

3.2 Estratégia de vetorização

Com o objetivo de contornar as dificuldades associadas ao cálculo da direção de Newton para o problema (3.4), H. Sato em [12] propôs uma estratégia que consiste em obter a matriz que representa $\operatorname{Hess} f(Y)$ como uma transformação linear em $\operatorname{St}(p,n)$, para um Y fixado, de modo que possamos reescrever (3.15) como uma equação linear do tipo $Ax = b$. Para tanto, é conveniente utilizar a seguinte propriedade: todo vetor tangente $\xi \in T_Y \operatorname{St}(p,n)$ pode ser decomposto como

$$\xi = YB + Y_\perp C, \tag{3.17}$$

onde $B \in \mathrm{S}_{\mathrm{skew}}(p)$ e $C \in \mathbb{R}^{(n-p)\times p}$, veja [12]. Em particular, como $\operatorname{Hess} f(Y)[\xi] \in T_Y \operatorname{St}(p,n)$, podemos escrever

$$\operatorname{Hess} f(Y)[\xi] = YB_H + Y_\perp C_H, \tag{3.18}$$

onde $B_H \in \mathrm{S}_{\mathrm{skew}}(p)$ e $C_H \in \mathbb{R}^{(n-p)\times p}$.

Proposição 5. Seja $Y \in \mathrm{St}(p,n)$, com $p < n$ e $Y_\perp \in \mathrm{St}(n-p,n)$ satisfazendo $Y^TY_\perp = 0$ e $Y^TY = I_p$. Se $\xi \in T_Y\,\mathrm{St}(p,n)$, em que $\xi = YB + Y_\perp C$, então $\mathrm{Hess}\, f(Y)[\xi] = YB_H + Y_\perp C_H$, onde

$$\begin{aligned} B_H &= -4\sum_{l=1}^{N} \mathrm{skew}((Z_lB + Z^\perp C)\,\mathrm{diag}(Z_l) + 2Z_l\,\mathrm{diag}(Z_lB + Z_l^\perp C) \\ &\quad - B\,\mathrm{sym}(Z_l\,\mathrm{diag}(Z_l))), \\ \text{e } C_H &= -4\sum_{l=1}^{N}(((Z_l^\perp)^TB + Z^{\perp\perp}C)\,\mathrm{diag}(Z_l) + 2(Z_l^\perp)^T\,\mathrm{diag}(Z_lB + Z_l^\perp C) \\ &\quad - C\,\mathrm{sym}(\mathrm{Z_l}\,\mathrm{diag}(\mathrm{Z_l}))), \end{aligned}$$

onde $Z_l = Y^TA_lY$, $Z_l^\perp = Y^TA_lY_\perp$ e $Z_l^{\perp\perp} = Y_\perp^TA_lY_\perp$.

Demonstração: Note que, dado $W \in \mathbb{R}^{n\times p}$, é válido que

$$\begin{aligned} Y^TP_Y(W) &= Y^T[(I - YY^T)W + Y\,\mathrm{skew}(Y^TW)] \\ &= Y^TW - Y^TW + \mathrm{skew}(Y^TW) \\ &= \mathrm{skew}(Y^TW). \end{aligned}$$

Além disso,

$$\begin{aligned} Y_\perp^TP_Y(W) &= Y_\perp^T[W - Y\,\mathrm{sym}(Y^TW)] \\ &= Y_\perp^T(W) - Y_\perp^TY\,\mathrm{sym}(Y^TW) \\ &= Y_\perp^T(W), \end{aligned}$$

para todo $W \in \mathbb{R}^{n\times p}$. Multiplicando Y^T em ambos os lados da equação (3.18) e usando as relações $Y^TY = I_p$ e $Y^TY_\perp = 0$, temos

$$\begin{aligned} B_H &= Y^T\ \mathrm{Hess}\, f(Y)[\xi] \\ &= -4\sum_{l=1}^{N}\mathrm{skew}\Bigg(Y^T\Big(A_l\xi\ \mathrm{diag}\big(Y^TA_lY\big) + 2A_lY\,\mathrm{diag}(Y^TA_l\xi) \\ &\quad - \xi\ \mathrm{sym}\Big(Y^TA_lY\ \mathrm{diag}\big(Y^TA_lY\big)\Big)\Big)\Bigg). \end{aligned}$$

Similarmente, ao multiplicarmos $Y_\perp^T$ na equação (3.18), obtemos

$$\begin{aligned} C_H &= Y_\perp^T \operatorname{Hess} f(Y)[\xi] \\ &= -4\sum_{l=1}^{N} Y_\perp^T \Big(A_l\xi \operatorname{diag}\big(Y^T A_l Y\big) + 2A_l Y \operatorname{diag}\big(Y^T A_l \xi\big) \\ &\quad - \xi \operatorname{sym}\Big(Y^T A_l Y \operatorname{diag}\big(Y^T A_l Y\big)\Big)\Big). \end{aligned}$$

□

A Proposição 5 foi retirada de [12].

3.2.1 Produto de Kronecker e operadores vec e veck

Nesta seção, introduziremos alguns operadores que serão fundamentais para o desenvolvimento das etapas subsequentes.

Definição 30. Sejam $A \in \mathbb{R}^{m\times n}$ e $B \in \mathbb{R}^{p\times q}$. O *produto de Kronecker* $A \otimes B \in \mathbb{R}^{mp\times nq}$ é definido como

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix}.$$

Definição 31. Seja $W = (w_{ij}) \in \mathbb{R}^{m\times n}$. Definimos o operador vec como

$$\operatorname{vec}(W) = \begin{pmatrix} w_{11} \\ w_{21} \\ \vdots \\ w_{m1} \\ w_{12} \\ \vdots \\ w_{mn} \end{pmatrix} \in \mathbb{R}^{mn}.$$

Definição 32. Seja $S = (s_{ij}) \in \mathbb{R}^{n\times n}$ uma matriz antissimétrica, isto é, $S^T = -S$.

Definimos o operador veck como

$$\text{veck}(S) = \begin{pmatrix} s_{21} \\ s_{31} \\ \vdots \\ s_{n1} \\ s_{32} \\ \vdots \\ s_{n2} \\ \vdots \\ s_{n,n-1} \end{pmatrix} \in \mathbb{R}^{\frac{n(n-1)}{2}}.$$

Exemplo 7. Seja

$$S = \begin{bmatrix} 0 & -s_{12} & -s_{13} & -s_{14} \\ s_{12} & 0 & -s_{23} & -s_{24} \\ s_{13} & s_{23} & 0 & -s_{34} \\ s_{14} & s_{24} & s_{34} & 0 \end{bmatrix} \in \mathbb{R}^{4\times 4},$$

uma matriz antissimétrica. Aplicando o operador veck, temos

$$\text{veck}(S) = \begin{bmatrix} s_{21} \\ s_{31} \\ s_{41} \\ s_{32} \\ s_{42} \\ s_{43} \end{bmatrix} = \begin{bmatrix} -s_{12} \\ -s_{13} \\ -s_{14} \\ -s_{23} \\ -s_{24} \\ -s_{34} \end{bmatrix} \in \mathbb{R}^6.$$

Proposição 6. Seja $U \in \mathbb{R}^{m\times p}$, $V \in \mathbb{R}^{p\times q}$ e $W \in \mathbb{R}^{q\times n}$. Então

$$\text{vec}(UVW) = (W^T \otimes U)\,\text{vec}(V). \tag{3.19}$$

Demonstração: Note que, dadas $A \in \mathbb{R}^{m\times p}$ e $X \in \mathbb{R}^{p\times q}$, então

$$\text{vec}(AX) = (I_q \otimes A)\ \text{vec}(X).$$

De fato, escrevendo $X = [x_1, x_2, \ldots, x_q]$, onde x_j $(j = 1, \ldots q)$ denota a j-ésima coluna de X, obtemos

$$AX = [Ax_1, Ax_2, \ldots, Ax_q].$$

Considere a matriz

$$I_q \otimes A = \begin{bmatrix} A & 0 & \cdots & 0 \\ 0 & A & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A \end{bmatrix} \in \mathbb{R}^{mq \times pq}.$$

Assim, podemos escrever

$$\text{vec}(AX) = \begin{bmatrix} Ax_1 \\ Ax_2 \\ \vdots \\ Ax_q \end{bmatrix} = (I_q \otimes A) \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{bmatrix}.$$

Portanto, $\text{vec}(AX) = (I_q \otimes A)\ \text{vec}(X)$. Agora, seja $B = \begin{bmatrix} b_1 & b_2 & \cdots & b_n \end{bmatrix} \in \mathbb{R}^{q \times n}$, onde $b_1, b_2, \cdots, b_n \in \mathbb{R}^q$. Então $XB = \begin{bmatrix} Xb_1 & Xb_2 & \cdots & Xb_n \end{bmatrix}$. Logo, $\text{vec}(XB) = \begin{bmatrix} Xb_1 \\ Xb_2 \\ \vdots \\ Xb_n \end{bmatrix}$.

Se escrevemos $B^T = \begin{bmatrix} b_1^T \\ b_2^T \\ \vdots \\ b_n^T \end{bmatrix}$, então $B^T \otimes I_p = \begin{bmatrix} b_1^T \otimes I_p \\ b_2^T \otimes I_p \\ \vdots \\ b_n^T \otimes I_p \end{bmatrix}$.

Multiplicando por vec(X), obtemos

$$(B^T \otimes I_p)\,\text{vec}(X) = \begin{bmatrix} Xb_1 \\ Xb_2 \\ \vdots \\ Xb_n \end{bmatrix}.$$

Portanto, $\text{vec}(XB) = (B^T \otimes I_p)\,\text{vec}(X)$. Assim, podemos escrever

$$\text{vec}(VW) = (W^T \otimes I_p)\,\text{vec}(V). \tag{3.20}$$

Além disso,

$$\text{vec}(U(VW)) = (I_n \otimes U)\ \text{vec}(VW). \tag{3.21}$$

Substituindo (3.20) em (3.21), obtemos $\text{vec}(UVW) = (I_n \otimes U)(W^T \otimes I_p)\,\text{vec}(V)$. Observe que,

$$(I_n \otimes U)(W^T \otimes I_p) = (W^T \otimes U).$$

Logo, $\text{vec}(UVW) = (W^T \otimes U)\,\text{vec}(V)$. □

A Proposição 6 segue do resultado apresentado em [9].

Observação 16. Para toda matriz $W \in \mathbb{R}^{n\times n}$, é válido que $T_n \operatorname{vec}(W) = \operatorname{vec}(W^T)$, onde

$$T_n = \sum_{i,j=1}^{n} E_{ij}^{(n\times n)} \otimes E_{ji}^{(n\times n)}.$$

Para mais detalhes, veja [9].

Exemplo 8. Para ilustrar a construção do operador T_n, consideremos $n = 2$. Nesse caso,

$$T_2 = \sum_{i,j=1}^{2} E_{ij}^{(2\times 2)} \otimes E_{ji}^{(2\times 2)}.$$

Segue que

$$T_2 = E_{11} \otimes E_{11} \ + \ E_{12} \otimes E_{21} \ + \ E_{21} \otimes E_{12} \ + \ E_{22} \otimes E_{22},$$

onde

$$E_{11} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad E_{12} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad E_{21} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad E_{22} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Assim, obtemos

$$T_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Proposição 7. Para qualquer $W \in \mathbb{R}^{n\times n}$,

$$\operatorname{vec}(\operatorname{sym}(W)) = \frac{1}{2}(I_{n^2} + T_n)\operatorname{vec}(W), \quad \text{e} \quad \operatorname{vec}(\operatorname{skew}(W)) = \frac{1}{2}(I_{n^2} - T_n)\operatorname{vec}(W).$$

Demonstração: Escrevemos $W = [w_1 \ w_2 \ \dots \ w_n]$, onde $w_j \in \mathbb{R}^n$ são as colunas de W. Então

$$\operatorname{vec}(W) = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}, \quad \operatorname{vec}(W^T) = T_n \operatorname{vec}(W).$$

Sabemos que

$$\operatorname{vec}(\operatorname{sym}(W)) = \operatorname{vec}\left(\frac{1}{2}(W + W^T)\right) = \frac{1}{2}\operatorname{vec}(W + W^T).$$

Usando a linearidade do operador vec, temos

$$\operatorname{vec}(\operatorname{sym}(W)) = \frac{1}{2}\big(\operatorname{vec}(W) + \operatorname{vec}(W^T)\big) = \frac{1}{2}\big(\operatorname{vec}(W) + T_n \operatorname{vec}(W)\big).$$

Portanto,

$$\operatorname{vec}(\operatorname{sym}(W)) = \frac{1}{2}(I_{n^2} + T_n)\operatorname{vec}(W).$$

Para a parte antissimétrica, lembraremos que

$$\operatorname{skew}(W) = \frac{1}{2}(W - W^T).$$

Aplicando o operador vec e usando sua linearidade, obtemos

$$\operatorname{vec}(\operatorname{skew}(W)) = \operatorname{vec}\Big(\tfrac{1}{2}(W - W^T)\Big) = \tfrac{1}{2}\big(\operatorname{vec}(W) - \operatorname{vec}(W^T)\big).$$

Como $\operatorname{vec}(W^T) = T_n \operatorname{vec}(W)$, segue que

$$\operatorname{vec}(\operatorname{skew}(W)) = \tfrac{1}{2}\big(\operatorname{vec}(W) - T_n \operatorname{vec}(W)\big) = \tfrac{1}{2}(I_{n^2} - T_n)\operatorname{vec}(W).$$

□

A Proposição 7 foi retirada de [9].

Observação 17. Seja

$$D_n = \sum_{n \geqslant i > j \geqslant 1} \left(E^{(n^2 \times n(n-1)/2)}_{n(j-1)+i,\, j(n-(j+1)/2)-n+i} - E^{(n^2 \times n(n-1)/2)}_{n(i-1)+j,\, j(n-(j+1)/2)-n+i} \right). \tag{3.22}$$

Além disso, vale $D_n^T D_n = I_{n(n-1)/2}$. Para mais detalhes, consulte, [12].
Para toda matriz $n \times n$ antisimétrica S, vale que

$$\operatorname{vec}(\mathrm{S}) = D_n \operatorname{veck}(\mathrm{S}) \quad \text{e} \quad \operatorname{veck}(S) = \frac{1}{2} D_n^T \operatorname{vec}(\mathrm{S})\,.$$

Para mais detalhes, consulte [9].

Exemplo 9. Seja $n = 3$, obtemos $D_3 \in \mathbb{R}^{9\times 3}$ dada por

$$D_3 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Proposição 8. Para todo $n \in \mathbb{N}$, é válido que $D_n^T = -D_n^T T_n$.

Demonstração: Sabemos que

$$\operatorname{vec}(S) = D_n \operatorname{veck}(S), \quad \text{e} \quad \operatorname{veck}(S) = \frac{1}{2} D_n^T \operatorname{vec}(S).$$

Aplicando D_n^T em ambos os da igualdade, obtemos

$$D_n^T \operatorname{vec}(S) = D_n^T D_n \operatorname{veck}(S) = 2 \operatorname{veck}(S).$$

Para qualquer matriz S antisimétrica,

$$\operatorname{vec}(S^T) = T_n \operatorname{vec}(S) = \operatorname{vec}(-S) = -\operatorname{vec}(S).$$

Multiplicando ambos os lados por D_n^T

$$D_n^T \operatorname{vec}(S^T) = D_n^T T_n \operatorname{vec}(S) = D_n^T(-\operatorname{vec}(S)) = -D_n^T \operatorname{vec}(S).$$

A última igualdade nos dá

$$D_n^T \operatorname{vec}(S) = -D_n^T T_n \operatorname{vec}(S),$$

Segue que

$$D_n^T = -D_n^T T_n.$$

□

O resultado estabelecido na Proposição 8 encontra-se originalmente em [9].

3.2.2 Equação de Newton

Sabe-se que a Hessiana Riemanniana, $\operatorname{Hess} f(Y)$, é um operador que recebe um vetor $\xi \in T_Y\mathcal{M}$ e retorna outro vetor $\operatorname{Hess} f(Y)[\xi] \in T_Y\mathcal{M}$.
Seguindo a estratégia proposta em [12], interpretaremos a Hessiana de f no ponto $Y \in \operatorname{St}(p,n)$ como uma transformação linear definida sobre o espaço tangente $T_Y \operatorname{St}(p,n)$. Para isso, representamos vetores $\xi \in T_Y\mathcal{M}$ segundo a decomposição $\xi = YB + Y_\perp C$, onde $Y \in \mathcal{M}$, $B \in \mathrm{S}_{\mathrm{skew}}(p)$ e $C \in \mathbb{R}^{(n-p)\times p}$, como já mostrado em (3.17).
Note que $\operatorname{veck}(B) \in \mathbb{R}^{p(p-1)/2}$ e $\operatorname{vec}(C) \in \mathbb{R}^{p(n-p)}$. Definindo $K := \frac{p(p-1)}{2} + p(n-p)$, temos

$$\begin{bmatrix} \operatorname{veck}(B) \\ \operatorname{vec}(C) \end{bmatrix} \in \mathbb{R}^K, \qquad \begin{bmatrix} \operatorname{veck}(B_H) \\ \operatorname{vec}(C_H) \end{bmatrix} \in \mathbb{R}^K,$$

onde $B_H \in \mathrm{S}_{\mathrm{skew}}(p)$ e $C_H \in \mathbb{R}^{(n-p)\times p}$ são os coeficientes que satisfaz (3.18). Dessa forma,

podemos definir o operador

$$\begin{aligned} H : \mathbb{R}^K &\longrightarrow \mathbb{R}^K \\ \begin{bmatrix} \operatorname{veck}(B) \\ \operatorname{vec}(C) \end{bmatrix} &\longmapsto \begin{bmatrix} \operatorname{veck}(B_H) \\ \operatorname{vec}(C_H) \end{bmatrix}. \end{aligned}$$

A fim de mostrar que o operador H é linear, consideremos

$$x_1 = \begin{bmatrix} \operatorname{veck}(B_1) \\ \operatorname{vec}(C_1) \end{bmatrix}, \qquad x_2 = \begin{bmatrix} \operatorname{veck}(B_2) \\ \operatorname{vec}(C_2) \end{bmatrix} \in \mathbb{R}^K,$$

onde $B_1, B_2 \in \mathrm{S}_{\mathrm{skew}}(p)$ e $C_1, C_2 \in \mathbb{R}^{(n-p)\times p}$, e tomemos $\alpha, \beta \in \mathbb{R}$. Para cada par (B_i, C_i), onde $i = 1, 2$, definimos o vetor tangente

$$\xi_i = YB_i + Y_\perp C_i \in T_Y\mathcal{M}.$$

Pela linearidade da Hessiana, segue que

$$\operatorname{Hess} f(Y)[\alpha\xi_1 + \beta\xi_2] = \alpha \operatorname{Hess} f(Y)[\xi_1] + \beta \operatorname{Hess} f(Y)[\xi_2].$$

Multiplicando ambos os lados por Y^T e $Y_\perp^T$, e usando as relações $Y^TY = I_p$ e $Y^TY_\perp = 0$, obtemos $B_H = \alpha B_{H_1} + \beta B_{H_2}$ e $C_H = \alpha C_{H_1} + \beta C_{H_2}$, onde

$$B_{H_i} = Y^T \operatorname{Hess} f(Y)[\xi_i] \quad \text{e} \qquad C_{H_i} = Y_\perp^T \operatorname{Hess} f(Y)[\xi_i].$$

Como os operadores veck e vec são lineares, segue que

$$\begin{aligned} H(\alpha x_1 + \beta x_2) &= \begin{bmatrix} \operatorname{veck}(B_H) \\ \operatorname{vec}(C_H) \end{bmatrix} \\ &= \begin{bmatrix} \operatorname{veck}(\alpha B_{H_1} + \beta B_{H_2}) \\ \operatorname{vec}(\alpha C_{H_1} + \beta C_{H_2}) \end{bmatrix} \\ &= \alpha \begin{bmatrix} \operatorname{veck}(B_{H_1}) \\ \operatorname{vec}(C_{H_1}) \end{bmatrix} + \beta \begin{bmatrix} \operatorname{veck}(B_{H_2}) \\ \operatorname{vec}(C_{H_2}) \end{bmatrix} \\ &= \alpha H(x_1) + \beta H(x_2). \end{aligned}$$

Portanto, o operador H é linear.
Nosso objetivo, portanto, é determinar a matriz de representação desse operador H, que permitirá reescrever a equação de Newton como uma equação da forma $Ax = b$.

Proposição 9. Seja $K = \frac{p(p-1)}{2} + p(n-p)$. Seja H uma transformação linear em $\mathbb{R}^K$ tal

que

$$H \begin{bmatrix} \operatorname{veck}(B) \\ \operatorname{vec}(C) \end{bmatrix} = \begin{bmatrix} \operatorname{veck}(B_H) \\ \operatorname{vec}(C_H) \end{bmatrix}, \tag{3.23}$$

em que $B \in \mathrm{S}_{\mathrm{skew}}(p)$, $C \in \mathbb{R}^{(n-p)\times p}$. Então, a matriz de representação H_A de H é dada por:

$$H_A = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix},$$

onde

$$\begin{aligned}
H_{11} &= -2D_p^T \sum_{l=1}^{N} \left[\operatorname{diag}(Z_l) \otimes Z_l + 2(I_p \otimes Z_l)\,\Delta_p\,(I_p \otimes Z_l) - \operatorname{sym}(Z_l \operatorname{diag}(Z_l)) \otimes I_p\right] D_p, \\
H_{12} &= -2D_p^T \sum_{l=1}^{N} \left[\operatorname{diag}(Z_l) \otimes Z_l^{\perp} + 2(I_p \otimes Z_l)\,\Delta_p\,(I_p \otimes Z_l^{\perp})\right], \\
H_{21} &= -4 \sum_{l=1}^{N} \left[\operatorname{diag}(Z_l) \otimes (Z_l^{\perp})^T + 2(I_p \otimes (Z_l^{\perp})^T)\,\Delta_p\,(I_p \otimes Z_l)\right] D_p, \\
H_{22} &= -4 \sum_{l=1}^{N} \left[\operatorname{diag}(Z_l) \otimes Z_l^{\perp\perp} + 2(I_p \otimes (Z_l^{\perp})^T)\,\Delta_p\,(I_p \otimes Z_l^{\perp}) - \operatorname{sym}(Z_l \operatorname{diag}(Z_l)) \otimes I_{n-p}\right].
\end{aligned} \tag{3.24}$$

Demonstração: Usando a expressão obtida para B_H na Proposição 5 e $\operatorname{veck}(B_H) = \frac{1}{2} D_n^T \operatorname{vec}(B_H)$, obtemos

$$\begin{aligned}
\operatorname{veck}(B_H) = \frac{1}{2} D_p^T \operatorname{vec}(-4 \sum_{l=1}^{N} &\operatorname{skew}((Z_l B + Z_l^{\perp} C) \operatorname{diag}(Z_l) + 2 Z_l \operatorname{diag}(Z_l B + Z_l^{\perp} C) \\
&- B \operatorname{sym}(Z_l \operatorname{diag}(Z_l)))).
\end{aligned}$$

Como $\operatorname{vec}(\operatorname{skew}(W) = \frac{1}{2}(I_{n_2} - T_n) \operatorname{vec}(W)$ para todo $W \in \mathbb{R}^{n\times n}$, obtemos

$$\begin{aligned}
\operatorname{veck}(B_H) &= -D_p^T (I_p - T_p) \sum_{l=1}^{N} \operatorname{vec}((Z_l B + Z_l^{\perp} C) \operatorname{diag}(Z_l) + 2 Z_l \operatorname{diag}(Z_l B + Z_l^{\perp} C) \\
&\quad - B \operatorname{sym}(Z_l \operatorname{diag}(Z_l)))) \\
&= -D_p^T (I_p - T_p) \sum_{l=1}^{N} \operatorname{vec}((Z_l B + Z_l^{\perp} C) \operatorname{diag}(Z_l)) + 2 \operatorname{vec}(Z_l \operatorname{diag}(Z_l B + Z_l^{\perp} C)) \\
&\quad - \operatorname{vec}(B \operatorname{sym}(Z_l \operatorname{diag}(Z_l))).
\end{aligned}$$

Usando a propriedade (3.19), temos

$$\begin{aligned}\operatorname{veck}(B_H) = -D_p^T(I_p - T_p)\sum_{l=1}^{N}((\operatorname{diag}(Z_l)\otimes Z_l)\operatorname{vec}(B) + (\operatorname{diag}(Z_l)\otimes Z_l^{\perp})\operatorname{vec}(C)\\ + 2(I_p\otimes Z_l)\operatorname{vec}(\operatorname{diag}(Z_lB + Z_l^{\perp}C)) - (\operatorname{sym}(Z_l\operatorname{diag}(Z_l))\otimes I_p)\operatorname{vec}(B).\end{aligned}$$

Tome

$$H_{11} = -2D_p^T\sum_{l=1}^{N}\left[\operatorname{diag}(Z_l)\otimes Z_l + 2(I_p\otimes Z_l)\,\Delta_p\,(I_p\otimes Z_l) - \operatorname{sym}(Z_l\operatorname{diag}(Z_l))\otimes I_p\right]D_p \text{ e}$$

$$H_{12} = -2D_p^T\sum_{l=1}^{N}\left[\operatorname{diag}(Z_l)\otimes Z_l^{\perp} + 2(I_p\otimes Z_l)\,\Delta_p\,(I_p\otimes Z_l^{\perp})\right],$$

e obtenha

$$\operatorname{veck}(B_H) = H_{11}\operatorname{veck}(B) + H_{12}\operatorname{vec}(C).$$

Analogamente, usando a expressão para C_H obtida na Proposição 5, temos

$$\begin{aligned}\operatorname{vec}(C_H) = \operatorname{vec}(-4\sum_{l=1}^{N}(((Z_l^{\perp})^TB + Z_l^{\perp\perp}C)\operatorname{diag}(Z_l) + 2(Z_l^{\perp})^T\operatorname{diag}(Z_lB + Z_l^{\perp}C)\\ - C\operatorname{sym}(Z_l\operatorname{diag}(Z_l))))\end{aligned}$$

Pela propriedade (3.19), é válido que

$$\begin{aligned}\operatorname{vec}(C_H) = -4\sum_{l=1}^{N}((\operatorname{diag}(Z_l)\otimes(Z_l^{\perp})^T)\operatorname{vec}(B) + (\operatorname{diag}(Z_l)\otimes Z_l^{\perp\perp})\operatorname{vec}(C)\\ + 2(I_p\otimes(Z_l^{\perp})^T)\operatorname{vec}(\operatorname{diag}(Z_lB + Z_l^{\perp}C)) - (\operatorname{sym}(Z_l\operatorname{diag}(Z_l))\otimes I_{p-n})\operatorname{vec}(C)).\end{aligned}$$

Assim, podemos escrever

$$\operatorname{vec}(C_H) = H_{21}\operatorname{veck}(B) + H_{2}2\operatorname{vec}(C),$$

onde

$$H_{21} = -4\sum_{l=1}^{N}\left[\operatorname{diag}(Z_l)\otimes(Z_l^{\perp})^T + 2(I_p\otimes(Z_l^{\perp})^T)\,\Delta_p\,(I_p\otimes Z_l)\right]D_p \text{ e}$$

$$H_{22} = -4\sum_{l=1}^{N}\left[\operatorname{diag}(Z_l)\otimes Z_l^{\perp\perp} + 2(I_p\otimes(Z_l^{\perp})^T)\,\Delta_p\,(I_p\otimes Z_l^{\perp}) - \operatorname{sym}(Z_l\operatorname{diag}(Z_l))\otimes I_{n-p}\right].$$

□

A Proposição 9 corresponde a um resultado previamente apresentado em [12]. Note que

obter $\xi \in T_Y\mathcal{M}$ como solução da equação $\operatorname{Hess} f(Y)[\xi] = -\operatorname{grad} f(Y)$, onde $Y \in \operatorname{St}(p,n)$ equivale a resolver o sistema

$$\begin{cases} Y^T \operatorname{Hess} f(Y)[\xi] = -Y^T \operatorname{grad} f(Y) \\ Y_\perp^T \operatorname{Hess} f(Y)[\xi] = -Y_\perp^T \operatorname{grad} f(Y), \end{cases}$$

com $Y_\perp$ satisfazeno $Y^T Y_\perp = 0$. Além disso, sabendo que $\operatorname{Hess} f(Y) = YB_H + Y_\perp C_H$, obtemos

$$\begin{cases} Y^T \operatorname{Hess} f(Y)[\xi] = B_H \\ Y_\perp^T \operatorname{Hess} f(Y)[\xi] = C_H. \end{cases} \tag{3.25}$$

Aplicando o operador veck à primeira equação de (3.25), o operador vec à segunda equação e usando (3.23), obtemos

$$H_A \begin{bmatrix} \operatorname{veck}(B) \\ \operatorname{vec}(C) \end{bmatrix} = - \begin{bmatrix} \operatorname{veck}(Y^T \operatorname{grad} f(Y)) \\ \operatorname{vec}(Y_\perp^T \operatorname{grad} f(Y)) \end{bmatrix}, \tag{3.26}$$

onde $\xi = YB + Y_\perp C$. Se H_A for inversível, podemos resolver (3.26) como

$$\begin{bmatrix} \operatorname{veck}(B) \\ \operatorname{vec}(C) \end{bmatrix} = -H_A^{-1} \begin{bmatrix} \operatorname{veck}(Y^T \operatorname{grad} f(Y)) \\ \operatorname{vec}(Y_\perp^T \operatorname{grad} f(Y)) \end{bmatrix}. \tag{3.27}$$

Uma vez obtidos $\operatorname{veck}(B)$ e $\operatorname{vec}(C)$, podemos obter $B \in \operatorname{Skew}(p)$ e $C \in \mathbb{R}^{(n-p)\times p}$. Portanto, conseguimos calcular a solução $\xi = YB + Y_\perp C$ da equação de Newton (3.25).
A seguir, apresenta-se o algoritmo do método de Newton, incluindo a estratégia de vetorização adotada para o cálculo da direção de Newton, proposto por [12].

Algoritmo 4: Método de Newton Riemanniano vetorizado

1 Escolha uma retração R de $\mathcal{M}$, um ponto inicial $Y_0 \in \mathcal{M}$ e $k = 0$.

2 Calcule $Y_\perp^{(k)}$ que satisfaça

$$\left(Y^{(k)}\right)^T Y_\perp^{(k)} = 0 \quad \text{e} \quad \left(Y_\perp^{(k)}\right)^T Y_\perp^{(k)} = I_{n-p}.$$

3 Calcule $Z_l^{(k)} = \left(Y^{(k)}\right)^T A_l Y^{(k)}, Z_l^{\perp(k)} = \left(Y^{(k)}\right)^T A_l Y_\perp^{(k)}, Z_l^{\perp\perp(k)} = \left(Y_\perp^{(k)}\right)^T A_l Y_\perp^{(k)}$, para $l = 1, 2, \ldots, N$.

4 Calcule $\left(Y^{(k)}\right)^T \operatorname{grad} f\left(Y^{(k)}\right)$ e $\left(Y_\perp^{(k)}\right)^T \operatorname{grad} f\left(Y^{(k)}\right)$ por

$$\left(Y^{(k)}\right)^T \operatorname{grad} f\left(Y^{(k)}\right) = -4\,\mathrm{skew}\left(\sum_{l=1}^{N} Z_l^{(k)}\ \mathrm{diag}(Z_l^{(k)})\right), \text{ e}$$

$$\left(Y_\perp^{(k)}\right)^T \operatorname{grad} f\left(Y^{(k)}\right) = -4\sum_{l=1}^{N}\left((Z_l^{\perp(k)})^T\ \mathrm{diag}(Z_l^{(k)})\right).$$

5 Calcule as matrizes $H_{11}^{(k)}, H_{12}^{(k)}, H_{21}^{(k)}, H_{22}^{(k)}$ usando (3.24), respectivamente, com $Z_l = Z_l^{(k)}$, $Z_l^\perp = Z_l^{\perp(k)}$, $Z_l^{\perp\perp} = Z_l^{\perp\perp(k)}$.

6 Calcule $b^{(k)} \in \mathbb{R}^{p(p-1)/2}$ e $c^{(k)} \in \mathbb{R}^{p(n-p)}$, usando

$$\begin{pmatrix} b^{(k)} \\ c^{(k)} \end{pmatrix} = -\begin{pmatrix} H_{11}^{(k)} & H_{12}^{(k)} \\ H_{21}^{(k)} & H_{22}^{(k)} \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{veck}\left((Y^{(k)})^T \operatorname{grad} f(Y^{(k)})\right) \\ \mathrm{vec}\left((Y_\perp^{(k)})^T \operatorname{grad} f(Y^{(k)})\right) \end{pmatrix}.$$

7 Calcule $B^{(k)} \in \mathrm{Skew(p)}$ e $C^{(k)} \in \mathbb{R}^{(n-p)\times p}$ tal que $\mathrm{veck}(B^{(k)}) = b^{(k)}$ e $\mathrm{vec}(C^{(k)}) = c^{(k)}$.

8 Calcule $\xi_k = Y_k B_k + Y_{\perp k} C_k$.

9 Calcule $Y_{k+1} = \mathrm{R}_{Y_k}(\xi_k)$.

10 Tome $k := k + 1$ e retorne ao passo 2.

Capítulo 4

Método de Newton Riemanniano Amortecido

Neste capítulo, apresentamos a busca de Armijo, escolhida para equipar o método de Newton Riemanniano, bem como o algoritmo completo do método com tal alteração. Além disso, provamos o resultado de convergência associado a esse método de Newton amortecido e analisamos os experimentos numéricos realizados ao comparar o desempenho dos métodos de Newton com e sem busca de Armijo na resolução do problema de diagonalização conjunta.

4.1 Busca de Armijo

Até o momento, a atualização dos pontos da sequência (Y_k) foi considerada apenas como a aplicação direta de uma retração a uma direção ξ_k pertencente ao espaço tangente em Y_k. Entretanto, é possível introduzir um parâmetro $\alpha_k > 0$, antes da aplicação da retração. Assim, o novo ponto é definido por

$$Y_{k+1} = \mathrm{R}_{Y_k}(\alpha_k \xi_k),$$

onde R é uma retração e $\alpha_k > 0$ é denominado comprimento de passo.
A escolha adequada desse parâmetro é realizada por meio de um procedimento denominado *busca*. O objetivo da busca é definir um critério que permita selecionar α_k de forma eficiente, pois passos muito grandes podem comprometer a convergência do método, enquanto passos excessivamente pequenos tendem a tornar a convergência lenta.
Uma das estratégias utilizadas é a *busca de Armijo*, que assegura que o valor da função objetivo decresça suficientemente a cada iteração. Formalmente, a condição de Armijo é expressa por

$$f(Y) - f\big(\mathrm{R}_Y(\alpha_k \eta)\big) \geqslant -\sigma \langle \operatorname{grad} f(Y), \alpha_k \eta \rangle, \tag{4.1}$$

em que $\mathrm{R}_Y(\alpha_k\eta)$ é uma retração sobre $\mathrm{St}(p,n)$, $\langle \operatorname{grad} f(Y), \alpha_k\eta\rangle = \alpha_k \operatorname{tr}(\operatorname{grad} f(Y)^T\eta)$, onde tr é o traço de uma matriz, $\sigma \in (0,1)$ e η pertence ao espaço tangente de $\mathrm{St}(p,n)$ em Y.
Observe que, se η é de fato uma direção de descida, então $\langle \operatorname{grad} f(Y), \eta\rangle < 0$. Consequentemente,

$$-\sigma\langle \operatorname{grad} f(Y), \alpha_k\eta\rangle > 0, \quad \forall \alpha_k > 0, \sigma \in (0,1).$$

Portanto, quando a condição de Armijo é satisfeita, obtemos

$$f(Y) - f(Y_{k+1}) = f(Y) - f\big(\mathrm{R}_Y(\alpha_k\eta)\big) \geqslant -\sigma\langle \operatorname{grad} f(Y), \alpha_k\eta\rangle > 0. \tag{4.2}$$

Em outras palavras, a condição de Armijo impõe, necessariamente, o decréscimo de f.
A seguir, apresentamos o Método de newton Riemanniano vetorizado, detalhado no Algorimo 4, equipado com a busca de Armijo. Quando o método de Newton é combinado com uma busca, ele passa a ser conhecido como método de Newton amortecido.

Algoritmo 5: Método de Newton Riemanniano vetorizado equipado com a busca de Armijo

1 Escolha uma retração R de $\mathcal{M}$, um ponto inicial $Y_0 \in \mathcal{M}$ e $k = 0$.

2 Calcule $Y_\perp^{(k)}$ que satisfaça

$$\left(Y^{(k)}\right)^T Y_\perp^{(k)} = 0 \quad \text{e} \quad \left(Y_\perp^{(k)}\right)^T Y_\perp^{(k)} = I_{n-p}.$$

3 Calcule $Z_l^{(k)} = \left(Y^{(k)}\right)^T A_l Y^{(k)}, Z_l^{\perp(k)} = \left(Y^{(k)}\right)^T A_l Y_\perp^{(k)}, Z_l^{\perp\perp(k)} = \left(Y_\perp^{(k)}\right)^T A_l Y_\perp^{(k)}$, para $l = 1, 2, \ldots, N$.

4 Calcule $\left(Y^{(k)}\right)^T \operatorname{grad} f\left(Y^{(k)}\right)$ e $\left(Y_\perp^{(k)}\right)^T \operatorname{grad} f\left(Y^{(k)}\right)$ por

$$\left(Y^{(k)}\right)^T \operatorname{grad} f\left(Y^{(k)}\right) = -4 \operatorname{skew}\left(\sum_{l=1}^{N} Z_l^{(k)} \operatorname{diag}(Z_l^{(k)})\right), \text{ e}$$

$$\left(Y_\perp^{(k)}\right)^T \operatorname{grad} f\left(Y^{(k)}\right) = -4 \sum_{l=1}^{N} \left((Z_l^{\perp(k)})^T \operatorname{diag}(Z_l^{(k)})\right).$$

5 Calcule as matrizes $H_{11}^{(k)}, H_{12}^{(k)}, H_{21}^{(k)}, H_{22}^{(k)}$ usando (3.24), respectivamente, com $Z_l = Z_l^{(k)}$, $Z_l^\perp = Z_l^{\perp(k)}$, $Z_l^{\perp\perp} = Z_l^{\perp\perp(k)}$.

6 Calcule $b^{(k)} \in \mathbb{R}^{p(p-1)/2}$ e $c^{(k)} \in \mathbb{R}^{(p(p-1)/2}$, usando

$$\begin{pmatrix} b^{(k)} \\ c^{(k)} \end{pmatrix} = - \begin{pmatrix} H_{11}^{(k)} & H_{12}^{(k)} \\ H_{21}^{(k)} & H_{22}^{(k)} \end{pmatrix}^{-1} \begin{pmatrix} \operatorname{veck}\left((Y^{(k)})^T \operatorname{grad} f(Y^{(k)})\right) \\ \operatorname{vec}\left((Y_\perp^{(k)})^T \operatorname{grad} f(Y^{(k)})\right) \end{pmatrix}.$$

7 Calcule $B^{(k)} \in \operatorname{Skew}(\mathrm{p})$ e $C^{(k)} \in \mathbb{R}^{(n-p)\times p}$ tal que $\operatorname{veck}(B^{(k)}) = b^{(k)}$ e $\operatorname{vec}(C^{(k)}) = c^{(k)}$.

8 Calcule $\xi_k = Y_k B_k + Y_{\perp k} C_k$

9 Calcule $\alpha_k > 0$ tal que

$$f(Y_k) - f\left(\mathrm{R}_{Y_k}(\alpha_k \xi_k)\right) \geqslant -\sigma \langle \operatorname{grad} f(Y_k), \alpha_k \xi_k \rangle,$$

onde $\sigma \in (0, 1]$.

10 Calcule $Y_{k+1} = \mathrm{R}_{Y_k}(\alpha_k \xi_k)$.

11 Tome $k := k + 1$ e retorne ao passo 2.

A convergência do método de Newton Riemanniano vetorizado se mantém rápida, preservando as propriedades locais do método de Newton, mesmo com a introdução da busca

de Armijo. Esta busca linear desempenha um papel importante no algoritmo, já que o comprimento do passo α_k é escolhido de forma a garantir que a função objetivo decresça a cada iteração. Ao controlar o tamanho do passo, a condição de Armijo evita, por exemplo, oscilações excessivas.
O teorema de convergência do método de Newton Riemanniano amortecido é um caso particular do Teorema 5. A seguir, apresentamos o enunciado do teorema de convergência local do Método de Newton Riemanniano equipado com a busca de Armijo, cuja demonstração pode ser consultada em [16].

Teorema 6. Seja $\mathcal{M}$ uma variedade Riemanniana, $\Omega \subset \mathcal{M}$ um aberto, e $X : \Omega \to T\mathcal{M}$ uma função suave. Seja R uma retração sobre $\mathcal{M}$. Se $Y_* \in \Omega$ é um ponto de acumulação de uma sequência (Y_k), gerada pelo Algoritmo 6, então $\operatorname{grad} f(Y_*) = 0$. Além disso, assumindo que $\operatorname{Hess} f$ seja não singular em Y_* e que $0 < \theta < 1/\operatorname{cond}(\nabla X(\bar{p}))$, onde $\operatorname{cond}(X) = ||X|| \cdot ||X^{-1}||$, a sequência (Y_k) converge superlinearmente para Y_*.

Neste trabalho, propomos comparar o método de Newton Riemanniano vetorizado clássico, descrito no Algoritmo 4, com sua versão amortecida, apresentada no Algoritmo 5. A principal diferença entre os dois algoritmos está no cálculo do ponto Y_{k+1}: enquanto no Algoritmo 4 temos $Y_{k+1} = \mathrm{R}_{Y_k}(\xi_k)$, no Algoritmo 5 o ponto é atualizado como $Y_{k+1} = \mathrm{R}_{Y_k}(\alpha_k \xi_k)$, em que α_k é um comprimento de passo escolhido para satisfazer a condição de Armijo. A seguir, apresentam-se os experimentos numéricos comparando os dois métodos na minimização da função (3.3).

4.2 Experimentos numéricos

Os experimentos numéricos conduzidos ao longo deste trabalho foram direcionados à minimização da função $f : \mathrm{St}(p, n) \to \mathbb{R}$, onde

$$f(Y) = -\sum_{l=1}^{N} \left\| \operatorname{diag}\left(Y^T A_l Y\right) \right\|_F^2 ,$$

a qual está associada ao problema de diagonalização conjunta das matrizes simétricas A_l $(l = 1, \ldots, N)$, de ordem n.
Para a realização dos testes, usamos a linguagem de programação Julia e seguimos as recomendações apresentadas em [12] para a construção das matrizes A_l, com o intuito de garantir que a solução do problema seja conhecida.
A geração das N matrizes simétricas foram realizadas a partir de N matrizes diagonais $n \times n$, $\Lambda^{(1)}, \Lambda^{(2)}, \ldots, \Lambda^{(N)}$, cujos elementos diagonais $\lambda_1^{(i)}, \ldots, \lambda_n^{(i)}$ $(i = 1, \ldots, N)$, são positivos e dispostos em ordem decrescente. Em seguida, é escolhida, de forma aleatória,

uma matriz ortogonal $P \in \mathbb{R}^{n\times n}$. As matrizes A_ℓ são então dadas por

$$A_\ell = P\Lambda^{(\ell)}P^T, \quad \ell = 1, \ldots, N.$$

Cada A_ℓ, assim obtida, é simétrica e preserva os autovalores das diagonais de $\Lambda^{(\ell)}$. Com a solução $W = PI_{n,p}$ obtida, é possível aplicar e comparar os métodos propostos, avaliando sua eficiência na aproximação de W.

Nos experimentos apresentados, fixamos $n = 10$, $p = 7$ e $N = 10$, aplicando o método de Newton Riemanniano proposto em [12] (Algoritmo 4) e o método de Newton Riemanniano com busca de Armijo (Algoritmo 5). O desempenho desses métodos foi avaliado a partir de diferentes pontos iniciais na variedade, construídos da seguinte maneira: escolhe-se uma direção aleatória $\eta \in T_W \operatorname{St}(p, n)$ e um escalar $\beta \in \mathbb{R}$, e define-se o chute inicial como

$$Y_0 = \mathrm{R}_W(\beta\eta) \in \operatorname{St}(p, n).$$

A retração empregada tanto para atualizar os pontos das sequências geradas pelos métodos quanto para construir os pontos iniciais foi a retração qf. Essa retração consiste em calcular a fatoração QR da matriz $Y + \xi$, onde $Y \in \operatorname{St}(n, p)$ é o ponto atual na variedade e $\xi \in T_Y \operatorname{St}(n, p)$ é o vetor tangente associado. Na decomposição $Y + \xi = QR$, $Q \in \operatorname{St}(n, p)$ possui colunas ortonormais e R é uma matriz triangular superior com elementos diagonais positivos. Define-se, então,

$$\mathrm{R}_Y(\xi) = \mathrm{qf}(Y + \xi),$$

onde $\mathrm{qf}(\cdot)$ representa o fator Q da decomposição QR.

Adotamos como critério de parada a norma do gradiente Riemanniano, exigindo que $\|\operatorname{grad} f(Y_k)\| < 10^{-8}$, e estabelecemos um limite máximo de 500 iteradas. Assim, cada método encerra sua execução ao atingir a tolerância prescrita para a norma do gradiente ou ao ultrapassar o número máximo permitido de iterações. Nesse último caso, entendemos que a sequência gerada pelo método não convergiu.

Para o método de Newton amortecido, os passos α devem satisfazer a condição de Armijo; enquanto a condição não é atendida, α é reduzido até satisfazê-la. Para evitar que o comprimento de passo se torne excessivamente pequeno, fixamos a tolerância mínima em 10^{-2}. Quando essa tolerância é atingida, exigimos que a busca de Armijo retorne o último comprimento de passo computado, o que pode ocasionar eventuais não decrescimentos da função objetivo em iterações pontuais. No entanto, essa estratégia mostrou-se adequada, pois, sem ela, observamos que o comprimento de passo reduzia rapidamente em algumas iterações, retardando a convergência do método. Com as modificações implementadas, ainda foi possível observar um comportamento predominantemente decrescente da função objetivo ao longo da maior parte das iterações.

Na Figura 4.1, apresentamos os resultados obtidos a partir de 10 direções distintas no

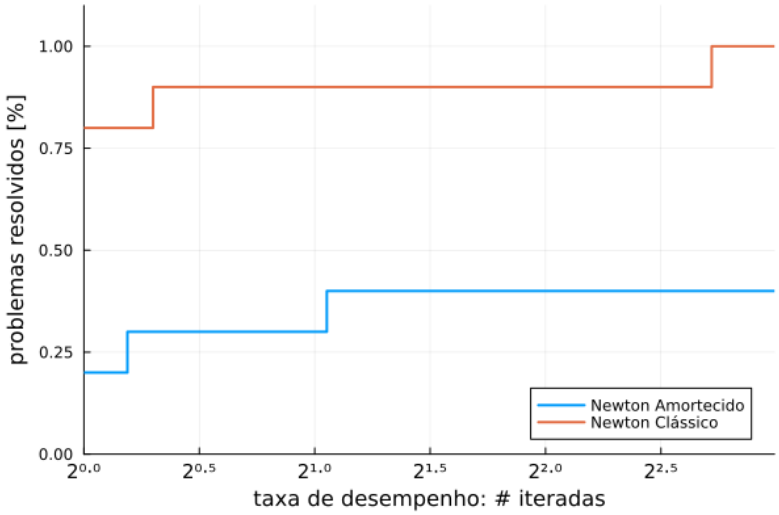

Figura 4.1: Comparação de desempenho entre e método de Newton clássico e o método de Newton Amortecido, variando as direções no espaco tangente $T_W \,\mathrm{St}(7, 10)$.

espaço tangente, considerando vetores unitários, isto é, tomando $\beta = 1$. Nesse cenário, cada método foi avaliado em 10 chutes iniciais diferentes na variedade $\mathrm{St}(10, 7)$. Consideramos como critério de comparação dos métodos o número de iterações.

Verificamos que o método de Newton clássico supera o amortecido em praticamente todos os níveis da taxa de desempenho considerados. Logo no início do gráfico, à esquerda, é possível notar que o método clássico já resolve cerca de 80% dos problemas com o menor número de iterações em relação ao método de Newton amortecido. Observamos ainda que o método de Newton clássico atinge rapidamente a marca de cerca de 90% de problemas resolvidos com melhor deesmpenho em relação ao método de newton amortecido.

Note que o método de Newton amortecido resolve 20% dos problemas com seu menor número de iteradas. O método alcança aproximadamente 40% dos problemas resolvidos apenas em taxas de desempenho maiores, e não ultrapassa cerca de 40% a 45% de resolução total.

Y_0	$\|W - Y_0\|$
Ponto Inicial 1	0.93007
Ponto Inicial 2	0.90784
Ponto Inicial 3	0.90023
Ponto Inicial 4	0.91900
Ponto Inicial 5	0.94717
Ponto Inicial 6	0.93346
Ponto Inicial 7	0.92699
Ponto Inicial 8	0.92982
Ponto Inicial 9	0.95966
Ponto Inicial 10	0.92348

Tabela 4.1: Norma da diferença entre a solução W e 10 pontos iniciais aleatórios.

Na Tabela 4.1, apresentamos a norma da diferença entre a solução W e os pontos iniciais utilizados no primeiro experimento, obtidos a partir das diferentes direções testadas. Entre esses casos, observamos que, no Ponto Inicial 1, o método de Newton amortecido apresentou desempenho superior ao método de Newton clássico. Ele alcançou uma solução com norma do gradiente menor, indicando maior precisão, e também exigiu menos iterações para convergir.

As Tabelas 4.2 e 4.3 exibem as últimas cinco iterações de cada método, contendo os valores da função objetivo avaliados nos pontos correspondentes e as respectivas normas do gradiente Riemanniano. Observa-se que, no método de Newton clássico, a norma do gradiente apresenta uma variação mais acentuada entre as iterações comparado ao método de Newton equipado com a busca.

Iteração (k)	$\|\|\operatorname{grad} f(Y_k)\|\|$
91	0.14641
92	0.02486
93	0.00268
94	6.05880×10^{-5}
95	1.02222×10^{-8}

Tabela 4.2: Últimas 5 iterações do método de Newton clássico.

Iteração (k)	$\|\|\operatorname{grad} f(Y_k)\|\|$
74	1.31008×10^{-5}
75	6.55043×10^{-6}
76	$3.27521\text{e}\times10^{-6}$
77	$1.63760\text{e}\times10^{-6}$
78	$6.08310\text{e}\times10^{-13}$

Tabela 4.3: Últimas 5 iterações do método de Newton amortecido.

Fixando uma direção aleatória em $T_Y\,\mathrm{St}(7,10)$ e variando os chutes iniciais por meio dos valores

$$\beta_s = \frac{1}{10^{s-1}}, \qquad s = 1, 2, \ldots, 100,$$

obtivemos uma lista de pontos iniciais, cuja norma da diferença entre cada $Y_0 = \mathrm{R}_Y(\beta_s\eta)$ e o ponto W, permanece dentro de um intervalo bastante pequeno.

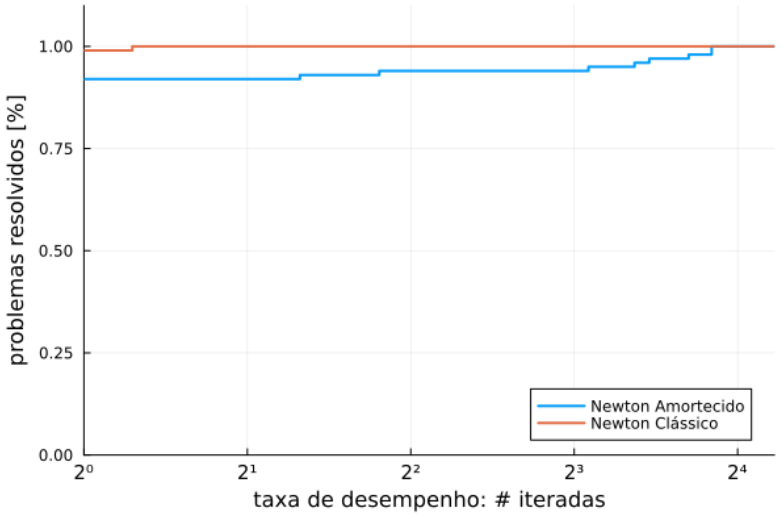

Figura 4.2: Comparação de desempenho entre e método de Newton clássico e o método de Newton Amortecido, fixando a direção e variando β.

Observando a Figura 4.2, verificamos que o método de Newton amortecido apresenta desempenho melhor do que no primeiro teste (Figura 4.1), embora ainda inferior ao do método clássico. Este comportamente nos motivou a reproduzir outro teste considerando pontos iniciais pouco mais distantes da solução. Para isso, fixamos uma única direção escolhida aleatoriamente em $T_Y \operatorname{St}(7, 10)$ e adotamos $\beta = 15$. O ponto inicial gerado com esses parâmetros apresentou norma da diferença em relação à solução (W) igual a aproximadamente (3,30225), um valor superior aos observados na Tabela 4.1. Nesse cenário, o método amortecido demandou 48 iterações para atingir o critério de convergência, enquanto o método de Newton clássico demandou 58 iterações. Tal resultado nos induz a pensar que o método de newton amortecido tem grande potencial ao consideramos pontos iniciais mais distantes da solução.

Iteração	$\|\operatorname{grad} f(Y)\|$
496	1.35766×10^{-5}
497	1.33644×10^{-5}
498	1.315565×10^{-5}
499	1.29500×10^{-5}
500	1.27477×10^{-5}

Tabela 4.4: Últimas 5 iterações do método de Newton amortecido para o ponto inicial inicial 5, considerando máximo de iteradas igual a 500.

Iteração	$\|\operatorname{grad} f(Y)\|$
946	1.13514×10^{-8}
947	1.11741×10^{-8}
948	1.09995×10^{-8}
949	1.08276×10^{-8}
950	1.06584×10^{-8}
951	1.04919×10^{-8}
952	1.03279×10^{-8}
953	1.01666×10^{-8}
954	1.00077×10^{-8}
955	9.85138×10^{-9}

Tabela 4.5: Últimas 10 iterações do método de Newton amortecido para o Ponto Inicial 5, considerando máximo de iteradas igual a 1000.

Durante os experimentos também foi possível observar que, apesar do decrescimento da função objetivo e da norma do gradiente, esta última diminuía de forma bastante lenta quando comparada ao comportamento do método de Newton clássico, no qual a norma do gradiente costuma oscilar de maneira mais acentuada entre iterações. Em diversos casos, notou-se, por exemplo, que a norma do gradiente permanecia em torno de 10^{-8} por várias iterações antes de atingir a ordem de 10^{-9}. Esse comportamento fazia com que o critério de convergência demorasse a ser satisfeito, ainda que a sequência já estivesse efetivamente

muito próxima da solução. A seguir, apresentamos um exemplo em que esse fenômeno é evidente.

Na Tabela 4.4, exibimos as cinco últimas iterações do método de Newton amortecido para o Ponto Inicial 5, cuja norma da diferença para a solução exata pode ser consultada na Tabela 4.1.

Ao aumentar o número máximo de iterações para 1000, a sequência convergiu após 955 iterações, cujas dez últimas são apresentadas na Tabela 4.5. A partir de uma análise mais detalhada, observou-se que a norma do gradiente já se encontrava da ordem de 10^{-8} desde a iteração 808. Esse comportamento não está relacionado à escolha da tolerância, pois a ordem 10^{-7} foi atingida ainda na iteração 662 e permaneceu praticamente estável até a iteração 801, padrão que também se repetiu para os demais testes.

Capítulo 5

Separação de imagens sobrepostas

Neste capítulo, empregamos o método de Newton Riemanniano amortecido, equipado com a busca de Armijo, para resolver um problema de separação de imagens sobrepostas, definido na variedade de Stiefel. Comparamos o referido método com o Newton clássico apresentado no Capítulo 3, considerando número de iteradas para tal comparação.

5.1 Análise de componente independente

A busca pelo aprimoramento de imagens é incentivada pelas diversas aplicações em que o reconhecimento e a análise de informações visuais são fundamentais, como em sistemas de vigilância e diagnósticos médicos. Em alguns contextos, as imagens capturadas podem conter sobreposições indesejadas que dificultam a identificação de elementos originais de tais imagens. Neste caso, é de interesse buscar uma aproximação razoável das imagens originais, a partir apenas das informações dessas sobreposições.
De modo geral, o problema de separação às cegas de fontes se resume em encontar uma representação linear em que os componentes sejam estatísticamente independentes.
Seja, portanto, $X_1, X_2, \ldots, X_n$ um conjunto de variáveis aleatórias definidas no mesmo espaço de probabilidade. Dizemos que essas variáveis são independentes se, e somente se, quaiasquer eventos determinados por qualquer grupo de variáveis aleatórias distintas são independentes, [4]. Por definição, as variáveis aleatórias X_i são independentes se

$$P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2)\cdots P(X_n),$$

onde P denota a medida de probabilidade associada ao espaço considerado.
A Análise de Componentes Independentes (ACI) é uma técnica estatística e computacional voltada à identificação de fatores que dão origem a conjuntos de sinais, medições ou variáveis aleatórias, [3]. A ideia central consiste em supor que as observações disponíveis correspondem a combinações de certas fontes desconhecidas, sendo tanto essas fontes quanto o sistema de mistura igualmente desconhecidos. No modelo adotado, essas

variáveis devem ser não gaussianas e estatisticamente independentes entre si; tais variáveis recebem o nome de componentes independentes, [3]. O papel da ACI é justamente estimar essas fontes/variáveis ocultas a partir dos dados observados.

Para ilustrar o modelo básico, considere que existam n sinais independentes $s_1, s_2, \ldots, s_n$. O que se observa, entretanto, não são esses sinais diretamente, mas sim n combinações misturadas $x_1, x_2, \ldots, x_n$, que podem ser expressas como $x = As$, onde

$$x = (x_1, x_2, \ldots, x_n)^T, \qquad s = (s_1, s_2, \ldots, s_n)^T,$$

e A é uma matriz de mistura quadrada $n \times n$, desconhecida.

A tarefa, portanto, é estimar uma matriz de separação B capaz de inverter (ao menos aproximadamente) o processo de mistura feito por meio da matriz A. Dessa forma, o vetor $z = Bx$ deve representar uma reconstrução, o mais fiel possível, das fontes originais s. Idealmente, se o modelo for exato e $B = A^{-1}$, então $z = s$.

Na prática, contudo, nem A nem s são conhecidos, de modo que o objetivo passa a ser encontrar uma matriz B que maximize a independência estatística entre os componentes de z. Essa busca é realizada com base em medidas de distribuição, que servem como critérios para avaliar o grau de independência obtido, [3].

O problema de ICA é frequentemente resolvido minimizando uma função objetivo, denominada função de contraste. Uma opção é a função de contraste JADE (Joint Approximate Diagonalization of Eigen-matrices), denotada por ϕ, que corresponde à soma dos cumulantes de quarta ordem dos elementos $z_1, z_2, \ldots, z_n$ de z, [12].

Em particular, vale a pena relembrar que, para uma variável aleatória contínua z_i com função densidade de probabilidade $f_{z_i}(z_i)$, a esperança matemática (ou valor esperado) é definida por

$$E(z_i) = \int_{-\infty}^{\infty} z_i \, f_{z_i}(z_i) \, dz_i,$$

onde $f_{z_i}(z_i) \geqslant 0$ e $\int_{-\infty}^{\infty} f_{z_i}(z_i) \, dz_i = 1$.

De forma análoga, para duas variáveis aleatórias contínuas z_i e z_j com função densidade de probabilidade conjunta $f_{z_i,z_j}(z_i, z_j)$, o valor esperado de $z_i z_j$ é dado por

$$E(z_i z_j) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z_i z_j \, f_{z_i,z_j}(z_i, z_j) \, dz_i \, dz_j.$$

Além disso, se z_i e z_j são independentes, então

$$E(z_i z_j) = E(z_i) E(z_j).$$

O valor esperado de de uma variável aleatória z_i é usada no cálculo dos cumulantes, que são medidas de dependência estatística entre variáveis. A independência entre componentes implica em cumulantes de ordem superior iguais a zero, [3]. Conforme descrito por [13],

tais cumulantes são calculados como

$$C_{ikl}(z) = \mathrm{E}(z_i z_j z_k z_l) - \mathrm{E}(z_i z_j)\,\mathrm{E}(z_k z_l) - \mathrm{E}(z_i z_k)\,\mathrm{E}(z_j z_l) - \mathrm{E}(z_i z_l)\,\mathrm{E}(z_j z_k).$$

A função JADE ϕ de uma variável aleatória z é definida como,

$$\phi(\mathbf{z}) = \sum_{\substack{i,j,k,l \\ i \neq j}} \left(C_{ijkl}(z)\right)^2.$$

Para reformular o problema como um problema de diagonalização conjunta, definem-se as matrizes de cumulantes de acordo com [13]. A matriz de cumulantes $Q^z(M)$, associada a uma dada matriz $n \times n$, $M = (m_{ij})$, é definida de modo que o seu elemento (i,j)-ésimo seja dado por

$$(Q^z(M))_{ij} = \sum_{k,l=1}^{n} C_{ijkl}(z) m_{kl}.$$

Desejamos procurar uma matriz de separação $B \in \mathrm{St}(p,n)$, com $p = n$. Utilizando $z = Bx$, conforme [12], temos que

$$\phi(z) = \sum_{k \leqslant l} ||\,\mathrm{off}(Q^z(M_{kl}))||_F^2,$$

onde $\mathrm{off}(A)$ denota a parte fora da diagonal da matriz A, e

$$M_{kl} = \begin{cases} E_{kl}^{(n\times n)}, & \text{se } k = l, \\ \dfrac{E_{kl}^{(n\times n)} + E_{lk}^{(n\times n)}}{\sqrt{2}}, & \text{se } k < l. \end{cases}$$

Definimos, portanto, $A_1, A_2, \ldots, A_N$ como $Q^x(M_{kl})$, para $k \leqslant l$, e definirmos $Y = B^T$.
Nosso objetivo, ao resolver o problema de separação de imagens sobrepostas, consiste em diagonalizar conjuntamente as matrizes $A_1, A_2, \ldots, A_N$, ou seja, resolver o problema de minimizar (3.4).
Buscando obter um ponto inicial adequado para a geração da sequência, seguimos a seguinte estratégia: tomamos uma matriz $\overline{A} \approx A$, tal que $\overline{A} = A + 0.001 \times \mathrm{rand}(n)$, onde $\mathrm{rand}(n)$ denota uma matriz de nordem n aleatória. Note que, em geral $\overline{A}^{-1} \notin \mathrm{St}(p,n)$ $(n = p)$. Portanto, dada $\overline{A}$, calculamos $B_0 := \mathrm{qf}(\overline{A}^{-1})$, onde $\mathrm{qf}(\cdot)$ denota o fator Q da decomposição QR. Assim, obtemos o ponto inicial $Y_0 := B_0^T \in \mathrm{St}(n,p)$, onde $p = n$. Esta estratégia, sugerida por [12] é razoável pois construímos a sobreposição das imagens para o experimento e portanto, conhecemos a matriz de mistura.
Dado o ponto inicial Y_0, aplicamos o método de Newton apresentado no Algoritmo 6 para

obter uma solução ótima Y_N e a matriz de separação correspondente $B_N = Y_N^T$. Por fim, calculamos $Z = B_N X$ e estimamos as imagens separadas.

Algoritmo 6: Método de Newton Riemanniano vetorizado para $n = p$

1 Escolha uma retração R de $\mathcal{M}$, um ponto inicial $Y_0 \in \mathcal{M}$ e $k = 0$.

2 Calcule $Z_l^{(k)} = \left(Y^{(k)}\right)^T A_l Y^{(k)}$, para $l = 1, 2, \ldots, N$.

3 Calcule $\left(Y^{(k)}\right)^T \operatorname{grad} f\left(Y^{(k)}\right)$ por

$$\left(Y^{(k)}\right)^T \operatorname{grad} f\left(Y^{(k)}\right) = -4\,\mathrm{skew}\left(\sum_{l=1}^{N} Z_l^{(k)}\ \mathrm{diag}(Z_l^{(k)})\right).$$

4 Calcule as matrizes $H_{11}^{(k)}$, usando

$$H_{11} = -2D_p^T \sum_{l=1}^{N} \left[\mathrm{diag}(Z_l) \otimes Z_l + 2(I_p \otimes Z_l)\,\Delta_p\,(I_p \otimes Z_l) - \mathrm{sym}(Z_l\,\mathrm{diag}(Z_l)) \otimes I_p\right] D_p,$$

com $Z_l = Z_l^{(k)}$.

5 Calcule $b^{(k)} \in \mathbb{R}^{p(p-1)/2}$, usando

$$\left(b^{(k)}\right) = -\left(H_{11}^{(k)}\right)^{-1} \left(\mathrm{veck}\left((Y^{(k)})^T \operatorname{grad} f(Y^{(k)})\right)\right).$$

6 Calcule $B^{(k)} \in \mathrm{Skew(p)}$ tal que $\mathrm{veck}(B^{(k)}) = b^{(k)}$.

7 Calcule $\xi_k = Y_k B_k$.

8 Calcule $Y_{k+1} = \mathrm{R}_{Y_k}(\alpha_k \xi_k)$.

9 Tome $k := k + 1$ e retorne ao item 2.

Note que, quando $n = p$, não existe subespaço ortogonal não trivial a Y dentro da variedade $\mathrm{St}(p, p)$. Nesse caso, a matriz $Y_\perp$, que no Algoritmo 5 representa uma base ortonormal para o complemento de Y, não existe, pois a dimensão desse complemento é zero. Como consequência, todos os termos que dependem de $Y_\perp$ desaparecem, isto é, não há termos $Z_l^\perp$ ou $Z_l^{\perp\perp}$, não há projeção do gradiente na direção de $Y_\perp$, e Hessiana reduz-se exclusivamente ao bloco H_{11}. Assim, o Algoritmo 6 é um caso particular do Algoritmo 5, para o caso $n = p$.

Observamos ainda que, o método de Newton vetorizado, apresentado no Algoritmo 6 em [12], diferencia-se do método de Newton amortecido principalmente na etapa de retração. Enquanto em [12] a retração é realizada por $\mathrm{R}_Y(\xi)$, neste estudo adotamos $\mathrm{R}_Y(\alpha\eta)$, em que α representa o comprimento de passo determinado pela regra de Armijo. Nesse contexto, quando $\alpha = 1$, o Algoritmo 6 corresponde ao método de Newton clássico. Por outro lado,

para $\alpha > 0$ satisfazendo a condição de Armijo, obtemos a versão amortecida do método de Newton.

5.2 Experimentos numéricos

Para esse experimento, consideramos $n = 3$ imagens obtidas de [14], mostradas nas Figuras 5.1, 5.2 e 5.3. Essas imagens foram carregadas de modo a ter o mesmo número de pixels 500×500. Em virtude das imagens serem coloridas, foi necessário transformá-las em preto e branco.

Figura 5.1: Imagem original 1.

Figura 5.2: Imagem original 2.

Figura 5.3: Imagem original 3.

Inicialmente, tratamos de construir as imagens sobrepostas para posteriormente realizar o processo de separação. As matrizes correspondentes às imagens apresentadas nas Figuras 5.1, 5.2 e 5.3, foram vetorizadas empilhando as colunas de cada uma e obtendo respectivamente 3 vetores I_1, I_2 e I_3, com dimensão 500^2. Em seguida, esses vetores foram organizados em uma matriz $S = [I_1^T, I_2^T, I_3^T]^T$ de dimensão 3×500^2. Por fim, calculamos a mistura $X = MS$, onde $M \in \mathrm{St}(n, n)$ é a matriz de mistura. Este produto gerou as imagens sobrepostas apresentadas nas Figuras 5.4, 5.5 e 5.6.

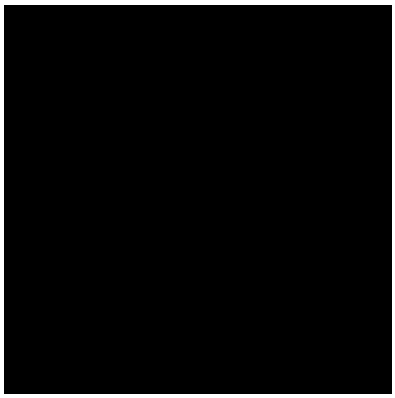

Figura 5.4: Imagem sobreposta 1.

Figura 5.5: Imagem sobreposta 2.

Figura 5.6: Imagem sobreposta 3.

Ao observar a imagem apresentada na Figura 5.4, notamos que ela aparece completamente preta. Esse comportamento está relacionado à forma como a máquina interpreta as matrizes associadas às imagens. As imagens originais são matrizes cujos elementos pertencem ao intervalo $[0, 1]$, uma vez que representam intensidades em tons de preto e branco. Já a matriz de mistura, escolhida aleatoriamente em $\mathrm{St}(n, n)$, pode produzir combinações lineares que resultem em valores fora desse intervalo, inclusive negativos. Diante disso, ao tentar exibir a imagem, a máquina interpreta tais valores como zero, gerando uma visualização inteiramente preta. Contudo, a matriz correspondente contém corretamente as informações da sobreposição; a limitação é exclusivamente visual. Assim, os dados numéricos que permanecem íntegros, foram utilizados normalmente no processo de separação, sem qualquer prejuízo sobre os resultados.

No estudo numérico, tomamos a tolerância para a norma do gradiente igual a 10^{-15}. Isto é, quando $||\operatorname{grad} f(Y_k)|| < 10^{-15}$, é declarado convergência do método. Se a quantidade de iterações for maior que 50, ou o comprimento de passo se tornar menor que 10^{-7}, o algoritmo é interrompido. Além disso, novamente optamos pela retração qf, já usada em experimentos anteriores. As aproximações das imagens originais obtidas a partir do método de Newton amortecido e método de Newton clássico estão dispostas na Figura 5.7 e Figura 5.8, respectivamente. O código foi executado em um computador com processador 11th Gen Intel(R) Core(TM) i3-1115G4, utilizando a linguagem de programação Julia.

Figura 5.7: Aproximações das imagens originais 1, 2 e 3, respectivamente, obtidas via método de Newton amortecido.

Observa-se que ambos os métodos produziram aproximações satisfatórias das imagens originais, em tempos muito próximos: o Método de Newton Clássico levou 1,634 segundos, enquanto o Método de Newton Amortecido levou 1,617 segundos. Entretanto, a diferença no número de iterações foi considerável, favorecendo o método amortecido. O método de Newton clássico precisou de 40 iterações para convergir, enquanto o método de Newton amortecido alcançou o mesmo resultado com apenas 23 iterações.

Com o intuito de avaliar de forma mais precisa o desempenho dos métodos na aproximação das imagens originais, construímos histogramas para cada imagem e para suas respectivas



Figura 5.8: Aproximações das imagens originais 1, 2 e 3, respectivamente, obtidas via método de Newton clássico.

aproximações obtidas por ambos os métodos. Isto é, desejamos analisar o conteúdo de cada imagem aproximada de forma numérica, observando como seus tons de cinza se distribuem ao longo da matriz da imagem aproximada em comparação à imagem original. Para gerar o histograma, percorremos todos os elementos da matriz da imagem e registramos quantas vezes cada intensidade de tom de cinza aparece. Como cada pixel assume um valor entre 0 e 255 (sendo 0 totalmente preto e 255 totalmente branco), o gráfico resultante é composto por 256 barras, cada uma correspondente à frequência de um nível específico de cinza. Barras mais altas indicam intensidades predominantes na imagem, enquanto barras mais baixas indicam tons pouco presentes.

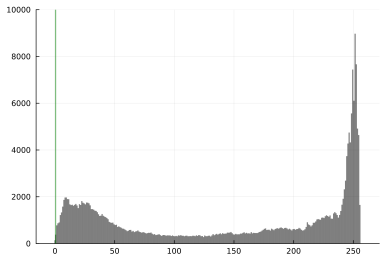

Figura 5.9: Comparação entre os histogramas da imagem sobreposta 1 e imagem original 1.

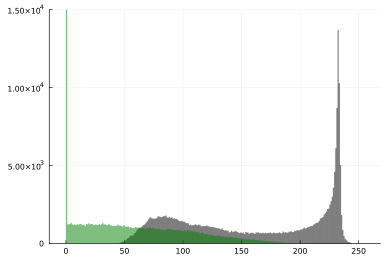

Figura 5.10: Comparação entre os histogramas da imagem sobreposta 2 e imagem original 2.

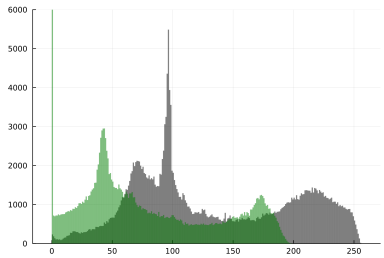

Figura 5.11: Comparação entre os histogramas da imagem sobreposta 3 e imagem original 3.

Inicialmente, geramos os histogramas das imagens sobrepostas apresentadas nas Figuras 5.4, 5.5 e 5.6, e os comparamos com os histogramas das imagens originais apresentadas nas Figuras 5.1, 5.2 e 5.3, respectivamente. Esses resultados estão reunidos nas Figuras 5.9, 5.10 e 5.11, onde o histograma em cinza representa cada imagem original e o histograma em verde corresponde à sua sobreposição.

O objetivo dessa comparação é evidenciar o quanto a mistura altera a distribuição dos tons de cinza, algo que pode ser verificado claramente quando observamos os histogramas. Em particular, para o primeiro par de histogramas apresentado na Figura 5.9, observa-se um pico acentuado na intensidade 0 no gráfico verde, refletindo o fato de que a primeira

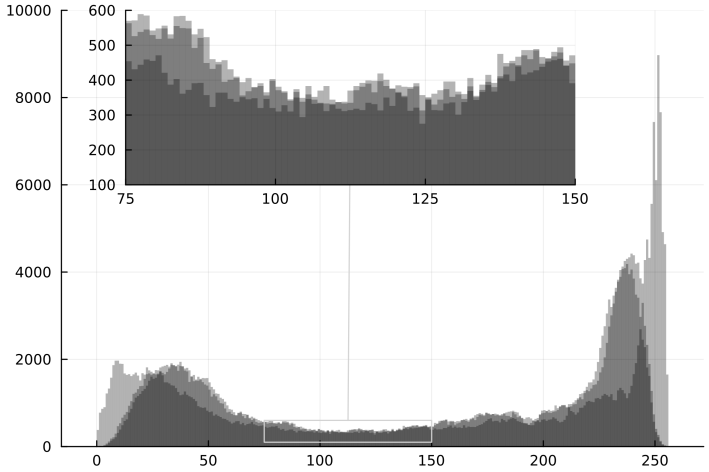

Figura 5.12: Histogramas da imagem original 1 e de suas aproximações obtidas pelos métodos de Newton clássico e Newton amortecido.

imagem sobreposta se apresentou totalmente preta, o que já era esperado.

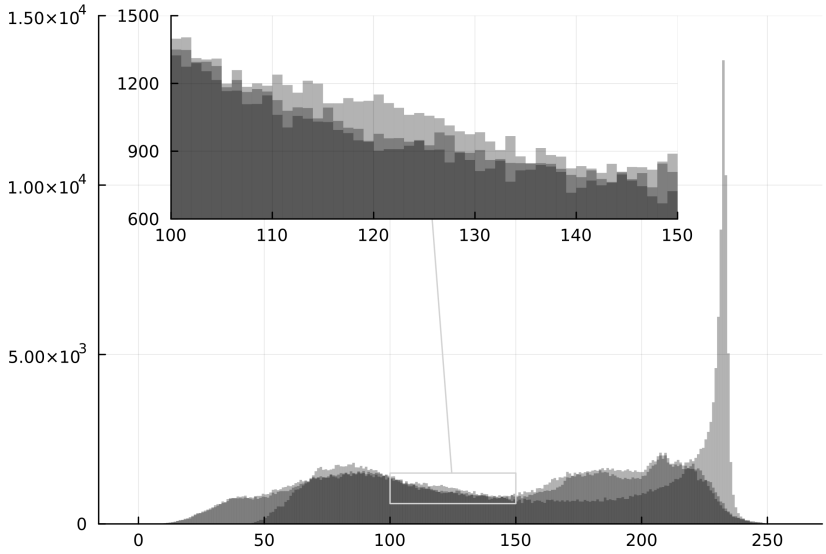

Figura 5.13: Histogramas da imagem original 2 e de suas aproximações obtidas pelos métodos de Newton clássico e Newton amortecido.

Além da análise das imagens sobrepostas, realizamos também uma comparação entre cada imagem original e suas respectivas aproximações obtidas pelos métodos de Newton amortecido e Newton clássico. Nas Figuras 5.12, 5.13 e 5.14, o histograma de tom mais claro corresponde à imagem original, o tom intermediário representa a aproximação produzida pelo método de Newton amortecido e o tom mais escuro indica a aproximação obtida pelo método de Newton clássico.

É evidente que a comparação por meio dos histogramas tornou mais nítida as diferenças que não eram perceptíveis quando observamos apenas as imagens aproximadas apresentadas nas Figuras 5.7 e 5.8, as quais pareciam visualmente muito semelhantes. Pelas Figuras

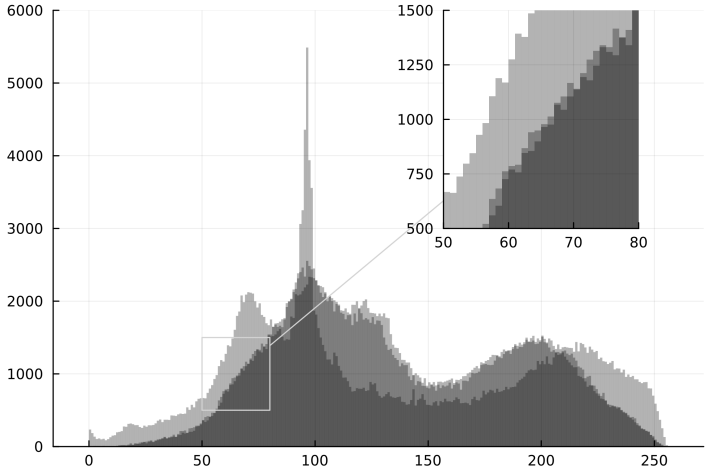

Figura 5.14: Histogramas da imagem original 3 e de suas aproximações obtidas pelos métodos de Newton clássico e Newton amortecido.

5.12, 5.13 e 5.14, nota-se que o método de Newton amortecido produz aproximações mais fiéis ao histograma da imagem original, reproduzindo com maior precisão as variações presentes na distribuição de tons de cinza.

Embora, em alguns intervalos, os histogramas das aproximações produzidas pelos dois métodos pareçam muito próximos entre si, e também relativamente próximos do histograma da imagem original, uma análise mais cuidadosa revela que o método de Newton amortecido mantém uma precisão superior. Mesmo nas regiões em que os histogramas se sobrepõem, pequenas variações evidenciam que o método de Newton amortecido aproxima de maneira mais precisa o histograma da imagem original.

Conclusão

Os experimentos numéricos realizados neste trabalho permitiram avaliar o desempenho do método de Newton Riemanniano clássico e do método de Newton Riemanniano amortecido, equipado com a busca de Armijo na resolução do problema de diagonalização conjunta de matrizes sobre a variedade de Stiefel. A análise considerou diferentes pontos iniciais, obtidos a partir de variações nas direções do espaço tangente à variedade e do parâmetro β, escolhido de modo a gerar diferentes chutes na variedade.

Verificou-se ainda que, quando o ponto inicial foi tomado mais distante da solução, o método de Newton amortecido apresentou desempenho superior ao do método de Newton clássico. Por outro lado, em situações nas quais os pontos iniciais já eram favoráveis ao método de Newton, o comportamento esperado, onde o método de Newton clássico convergiu mais rapidamente.

Outro ponto importante observado é que o método amortecido apresenta uma tendência mais estável de redução da norma do gradiente a cada iteração, enquanto o método de Newton clássico, apesar de mais eficiente em certos casos, pode exibir oscilações mais pronunciadas ao longo da convergência. Dessa forma, o método amortecido pode se mostrar mais assertivo em cenários nos quais o ponto inicial não é ideal, ampliando o raio de pontos aceitáveis para alcançar a solução.

Por fim, ao aplicar ambos os métodos ao problema de separação de imagens sobrepostas, em que as matrizes a serem diagonalizadas carregam as informações das sobreposições, verificou-se uma clara superioridade do método de Newton Riemanniano amortecido em relação ao método clássico, mesmo quando se utilizou um ponto inicial adequado. Esse resultado indica que o uso do amortecimento pode ser particularmente vantajoso em certos problemas.

Em síntese, os resultados obtidos demonstram que, embora o método de Newton Riemanniano clássico possa apresentar desempenho superior quando bem inicializado, o método amortecido se mostra mais estável e previsível, especialmente em contextos em que não se dispõe de um ponto inicial próximo à solução ou em que é possível estabelecer um número máximo de iteradas expressivo.

Bibliografia

[1] A. A. Ribeiro; E. W. Karas. *Otimização contínua: aspectos teóricos e computacionais*. Cengage Learning, 1ª edição, 2013.

[2] A. Izmailov; M. Solodov. *Otimização - volume 2: Métodos computacionais*. IMPA, 3ª edição, 2018.

[3] A. Hyvärienen; J. Karhunen; E. Oja.*Independent Component Analysis*. John Wiley & Sons, inc, 2001.

[4] B. R. James. *Probabilidade. Um Curso em Nível Intermediário*. IMPA, 2015.

[5] E. L. Lima. *Álgebra Linear*. IMPA, 1ª edição, 2014.

[6] E. L. Lima. *Curso de Análise - volume 1*. IMPA, 1ª edição, 2014.

[7] E. L. Lima. *Curso de Análise - volume 2*. IMPA, 1ª edição, 2014.

[8] E. L. Lima. *Elementos de topologia geral*. SBM, 4ª edição, 2024.

[9] E. W. Grafarend; S. Zwanzig; J. L. Awange. *Applications of Linear and Nonlinear Models: fixed effects, random effects, and mixed models.* Springer, 2022.

[10] G. H. Golub; C. F. V. Loan. *Matrix Computations*. Johns Hopkins University Press, 2ª edição, 2013.

[11] H. Farid; E. H. Adelson, *Separating reflections from images by use of independent component analysis*. 1999, J. Opt. Soc. Am. A 16, 2136-2145.

[12] H. Sato. *Riemannian Newton-type methods for joint diagonalization on the Stiefel manifold with application to independent component analysis*. 2017, Optimization 66.

[13] J. F. Cardoso; C.N.R.S; E.N.S.T *Blind signal separation: statistical principles*. 1998, Proceedings of the IEEE.

[14] Laurent de Vito. ICA for Demixing Images. https : / / github . com / ldv1 / ICA _ for _ demixing_images/tree/master. Acessado em: 13 de março de 2025.

[15] L. W. Tu. *An Introduction to Manifolds*. Springer, 2º edição, 2011.

[16] M. A. A. Bortoloti; T. A. Fernandes; O. P. Ferreira. *An efficient damped Newton-type algorithm with globalization strategy on Riemannian manifolds*. 2022, Journal of Computational and Applied Mathematics.

[17] M. P. Carmo. *Geometria Riemanniana*. IMPA, 5º edição, 2015.

[18] M. R. Spiegel. *Probabilidade e estatística*. Pearson, 2004.

[19] N. Boumal. *An Introduction to optimization on smooth manifolds*. pre-publication, Princeton University Press, 2013.

[20] P. A. Absil; R. M; R. S. *Optimization Algorithms on Matrix Manifolds*.Princeton University Press, 2007.

[21] R. A. Horn; C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

[22] S. C. Poltroniere; E. M. Soler; a A. B. Afonso. *Joint Approximate Diagonalization of Symmetric Real Matrices of Order 2*. Tendências em Matemática Aplicada e Computacional, 2016.

[23] W. de Melo. *Topologia das Variedades*. SBM, 1º edição, 2019.