UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

IAGO VICTOR PIRES DE SOUZA NUNES

Métodos Quasi-Newton Proximais para Problemas de Otimização Composta

Goiânia
2026

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[X] Dissertação [ ] Tese [ ] Outro*:_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

Exemplos: Estudo de caso ou Revisão sistemática ou outros formatos.

2. Nome completo do autor

Iago Victor Pires de Souza Nunes

3. Título do trabalho

Métodos quasi-Newton proximais para problemas de otimização composta

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [x] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Max Leandro Nobre Goncalves**, **Professor do Magistério Superior**, em 11/03/2026, às 08:56, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Iago Victor Pires De Souza Nunes**, **Discente**, em 12/03/2026, às 18:37, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **6043170** e o código CRC **E51DC269**.

Referência: Processo nº 23070.000948/2026-84 SEI nº 6043170

IAGO VICTOR PIRES DE SOUZA NUNES

Métodos Quasi-Newton Proximais para Problemas de Otimização Composta

Dissertação apresentada ao Programa de Pós–Graduação do Instituto de Matemática e Estatística da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Mestre em Matemática.

Área de concentração: Otimização.

Orientador: Prof. Max Leandro Nobre Gonçalves

Goiânia
2026

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Nunes, Iago Victor Pires de Souza
Métodos Quasi-Newton Proximais para Problemas de Otimização Composta [e-book] / Iago Victor Pires de Souza Nunes. - 2026.
LIII, 53 f.: 2026

Orientador: Prof. Dr. Max Leandro Nobre Gonçalves
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Matemática e Estatística (IME), Programa de Pós-Graduação em Matemática, Goiânia, 2026.
Bibliografia.
Inclui: algoritmos.

1. Otimização Matemática.

I. Gonçalves, Max Leandro Nobre, orient. II. Título.

CDU 51

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº **17** da sessão de Defesa de Dissertação de **Iago Victor Pires de Souza Nunes**, que confere o título de Mestre em **Matemática**, na área de concentração em **Otimização**.

Ao **vigésimo dia do mês de fevereiro do ano de dois mil e vinte e seis**, a partir das **14h00**, na forma de **Videoconferência**, realizou-se a sessão pública de Defesa de Dissertação intitulada **"Métodos quasi-Newton proximais para problemas de otimização composta"**. Os trabalhos foram instalados pelo Orientador, Professor Doutor **Max Leandro Nobre Gonçalves - IME/UFG** com a participação dos demais membros da Banca Examinadora: Professor Doutor **Maicon Marques Alves - DM/UFSC**, membro titular externo; Professor Doutor **Douglas Soares Gonçalves - DM/UFSC,** membro titular externo. Durante a arguição os membros da banca **não fizeram** sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pelo Professor Doutor **Max Leandro Nobre Gonçalves**, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, ao **vigésimo dia do mês de fevereiro do ano de dois mil e vinte e seis**.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Max Leandro Nobre Goncalves**, **Professor do Magistério Superior**, em 20/02/2026, às 15:58, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Douglas Soares Gonçalves**, **Usuário Externo**, em 26/02/2026, às 16:30, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Maicon Marques Alves**, **Usuário Externo**, em 03/03/2026, às 22:08, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5900443** e o código CRC **C6BE67D3**.

Referência: Processo nº 23070.000948/2026-84 SEI nº 5900443

Iago Victor Pires de Souza Nunes

Graduou-se em Engenharia da Computação na UFG, Universidade Federal de Goiás, e em Matemática pela UEG, Universidade Estadual de Goiás. Durante sua graduação, foi monitor no departamento de matemática da Universidade Estadual de Goiás e fez parte do programa de Residência Pedagógica, sendo bolsista pela CAPES. Durante o mestrado no IME-UFG, foi bolsista CAPES.

Dedico esse trabalho a minha família.

Agradecimentos

Agradeço a minha família por ter me dado o suporte necessário para essa fase de minha vida e ao meu orientador por ter disponibilizado seu tempo a me ajudar.

Resumo

Nunes, Iago Victor Pires de Souza. **Métodos Quasi-Newton Proximais para Problemas de Otimização Composta**. Goiânia, 2026. 52p. Dissertação de Mestrado. Instituto de Matemática e Estatística, Universidade Federal de Goiás.

Neste trabalho, analisamos o método proximal quasi-Newton e sua versão acelerada para resolver problemas de otimização composta. O estudo da versão não acelerada é baseado na referência [17], enquanto da versão acelerada é baseado no artigo [3]. É mostrado que o método clássico obtém uma taxa de convergência de $\mathcal{O}(1/k)$, isto é, para uma dada precisão $\varepsilon > 0$, o método gera uma iteração x_k tal que $F(x_k) - F(x^*) < \varepsilon$ em no máximo $\mathcal{O}(1/\varepsilon)$ iterações, onde x^* é um minimizador global do problema mencionado acima. Para a versão acelerada obtemos uma taxa de convergência melhor de $\mathcal{O}(1/k^2)$.

Palavras–chave

Otimização Composta, Métodos Proximais, Métodos Quasi-Newton, Análise de Convergência, Otimização Convexa

Abstract

Nunes, Iago Victor Pires de Souza. **<Proximal Quasi-Newton Methods for Composite Optimization Problems>**. Goiânia, 2026. 52p. MSc. Dissertation. Instituto de Matemática e Estatística, Universidade Federal de Goiás.

In this work, we analyze the proximal quasi-Newton method and its accelerated version to solve composite optimization problems. The study of the non-accelerated version is based on the reference [17], while the accelerated version is based on [3]. It is shown that the classical method obtains a convergence rate of $\mathcal{O}(1/k)$, i.e., for a given precision $\varepsilon > 0$, the method generates an iteration x_k such that $F(x_k) - F(x^*) < \varepsilon$ in at most $\mathcal{O}(1/\varepsilon)$ iterations, where x^* is a global minimizer of the problem mentioned above. For the accelerated version we obtain a better convergence rate of $\mathcal{O}(1/k^2)$.

Keywords

Composite Optimization, Quasi-Newton Methods, Convergence Analysis, Convex Optimization

Sumário

Lista de Algoritmos

CAPÍTULO 1

Introdução

Considere o problema de otimização convexa composto

$$\min_{x \in \mathbb{R}^n} F(x) := f(x) + g(x), \tag{1-1}$$

em que $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ é uma função convexa, semicontínua inferiormente e possivelmente não suave, e $f : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa, continuamente diferenciável e com gradiente Lipschitz contínuo com constante L, isto é, existe uma constante $L > 0$ tal que

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y \in \mathbb{R}^n.$$

Assumimos que o conjunto solução de (1-1), denotado por X^*, é não vazio. Esta classe de problemas, na qual $g(x)$ atua frequentemente como uma função regularizadora, possui ampla aplicação em áreas como aprendizado de máquina, processamento de sinais e estatística. Tal formulação é central e, quando $g(x) = \lambda\|x\|_1$, engloba modelos clássicos como a regressão logística esparsa [18, 22], a seleção esparsa de covariância inversa [4, 13, 16] e o Lasso irrestrito [21].

Um dos métodos fundamentais para solucionar o problema em (1-1) é o Algoritmo do Gradiente Proximal (AGP). Sendo uma variante dos métodos proximais, o AGP destaca-se pela robustez e facilidade de implementação, atingindo uma taxa de convergência global, em termos dos valores funcionais, de $\mathcal{O}(1/k)$ como pode ser visto em [1, 11]. Isso representa um avanço em relação aos métodos clássicos de subgradiente para funções não suaves, cuja taxa é tipicamente de $\mathcal{O}(1/\sqrt{k})$ [10]. Basicamente, a taxa $\mathcal{O}(1/k)$ em termos de valores funcionais significa que o algoritmo requer, no máximo, $\mathcal{O}(1/\varepsilon)$ iterações para gerar uma solução ε-aproximada, isto é, um ponto x_k tal que $F(x_k) - F(x^*) < \varepsilon$ com $x^* \in X^*$.

Visando aprimorar esse método e obter melhores taxas de convergência, foram desenvolvidas estratégias de aceleração baseadas em passos de extrapolações. O Algoritmo do Gradiente Proximal Acelerado (AGPA), proposto originalmente por Nesterov [9] e refinado por Beck e Teboulle [1] no algoritmo conhecido como FISTA (*Fast Iterative Shrinkage-Thresholding Algorithm*), consegue atingir a taxa de $\mathcal{O}(1/k^2)$. Esta é conhe-

cida como a taxa ótima de convergência para métodos que utilizam apenas informações de primeira ordem [8, 10, 15].

Buscando melhorar ainda mais o desempenho dos métodos proximais gradientes, especialmente em cenários práticos e mal condicionados, desenvolveu-se uma abordagem alternativa que incorpora informações de curvatura (segunda ordem) no passo proximal. Tais métodos, denominados Newton Proximal ou Quasi-Newton Proximal, ver por exemplo [2, 7, 13, 17, 20], generalizam a ideia do AGP por basicamente substituir o passo do gradiente pelo passo do Newton ou quasi-Newton juntamente com um ajuste na métrica do subproblema proximal. Enquanto o AGP pode ser interpretado como um método que utiliza uma métrica escalar da forma $H_k = (1/\mu_k)I$, os algoritmos Quasi-Newton Proximais (AQNP) substituem essa matriz por uma aproximação H_k da Hessiana de $f(x)$, simétrica e definida positiva. Em geral, embora a performance prática desses métodos possa ser superior à de suas contrapartes de primeira ordem, suas taxas de convergência global teóricas permanecem, na melhor das hipóteses, as mesmas; ver por exemplo, [17] para mais detalhes.

Neste trabalho, discutimos dois métodos proximal quasi-Newton, uma versão padrão AQNP e uma versão acelerada AQNPA, para resolver o problema (1-1). O estudo da versão não acelerada é baseado na referência [17], enquanto da versão acelerada é baseado no artigo [3]. É mostrado que o AQNP obtém uma taxa de convergência da $\mathcal{O}(1/k)$, isto é, para uma dada precisão $\varepsilon > 0$, o método gera uma iteração x_k tal que $F(x_k) - F(x^*) < \varepsilon$ em no máximo $\mathcal{O}(1/\varepsilon)$ iterações, onde $x^* \in X^*$. Para a versão acelerada AQNPA obtemos uma taxa de convergência melhor, da $\mathcal{O}(1/k^2)$. Vale ressaltar que a obtenção de uma taxa melhor para a versão acelerada depende de uma condição restritiva sobre a sequência de matrizes que aproxima a matriz Hessiana de f. Como consequência, nós também discutimos um algoritmo acelerado com aproximação da Hessiana fixa.

Este trabalho está organizado em cinco capítulos. O Capítulo 1 é esta introdução. O Capítulo 2 estabelece as notações, resultados preliminares e revisa brevemente os algoritmos clássicos de base. O Capítulo 3 apresenta o método AQNP padrão, detalhando sua estrutura e análise de convergência. O Capítulo 4 introduz o AQNPA (Acelerado), discutindo as condições necessárias para a aceleração e propondo a variante de Hessiana Fixa. Por fim, o Capítulo 5 apresenta as conclusões e discussões finais.

CAPÍTULO 2

Preliminares

Uma vez estabelecida a formulação do problema geral de otimização composta em (1-1), este capítulo dedica-se a apresentar a fundamentação teórica necessária para as análises subsequentes. Inicialmente, introduzimos as notações e os conceitos preliminares de análise convexa essenciais para o desenvolvimento do texto. Na sequência, realizamos uma revisão das estratégias algorítmicas clássicas, abrangendo métodos de primeira e segunda ordem que formam a base para a construção dos algoritmos proximais abordados neste trabalho. A exposição teórica fundamenta-se nas obras de referência [5, 14, 12], bem como nos desenvolvimentos apresentados em [3] e [19].

2.1 Conceitos básicos e notações

Nesta seção, apresentamos as notações e os conceitos preliminares sobre análise convexa e subgradientes essenciais para o desenvolvimento deste trabalho.

O produto interno usual em $\mathbb{R}^n$ é denotado por $\langle a, b\rangle := a^T b$, induzindo a norma euclidiana $\|a\| := \sqrt{a^T a}$. Dada uma matriz simétrica definida positiva $H \in \mathbb{R}^{n\times n}$, definimos o produto interno ponderado e a norma induzida, respectivamente, como

$$\langle a, b\rangle_H := a^T H b \quad \text{e} \quad \|a\|_H := \sqrt{a^T H a}.$$

Para matrizes simétricas A e B, a notação $A \succeq B$ (Ordem de Löwner) indica que a matriz diferença $A - B$ é semidefinida positiva. Analogamente, $A \succ B$ indica que $A - B$ é definida positiva. As notações $A \preceq B$ e $A \prec B$ seguem a mesma lógica, implicando que $B - A$ é semidefinida positiva ou definida positiva, respectivamente.

Por fim, em relação aos parâmetros escalares, adotaremos a seguinte convenção ao longo do texto: μ_k denota o tamanho do passo (comum em métodos de primeira ordem), enquanto σ_k denota o parâmetro de curvatura ou convexidade forte da aproximação quadrática. Vale ressaltar que, em geral, essas grandezas são inversamente proporcionais, isto é, $\sigma_k \propto 1/\mu_k$.

O domínio efetivo de F é definido por $\mathsf{dom}(F) := \{x \in \mathbb{R}^n : F(x) < \infty\}$. Já o conjunto de nível de F associado a um ponto $x \in \mathsf{dom}(F)$ é denotado por $\chi_F(x) := \{y \in \mathsf{dom}(F) : F(y) \leq F(x)\}$. O conjunto χ_0 denota o conjunto de nível $\chi_0 := \chi_F(x_0)$ dado algum $x_0 \in \mathsf{dom}(F)$.

Apresentamos, a seguir, resultados auxiliares fundamentais para o desenvolvimento das análises de convergência nos capítulos subsequentes. As demonstrações deste resultados podem ser encontradas nas referências [5], [14]. Algumas que são importantes para discussão subsequente são apresentadas em detalhes.

Teorema 2.1 *Sejam $f : \mathbb{R}^n \to \mathbb{R}$ uma função contínua e $\Omega \subset \mathbb{R}^n$ um conjunto compacto não vazio. Então, f assume um minimizador global em Ω.*

Prova. Inicialmente, mostremos que o conjunto imagem $f(\Omega) = \{f(x) \mid x \in \Omega\}$ é limitado inferiormente. Suponha, por contradição, que não seja. Então, para todo $k \in \mathbb{N}$, existe $x^k \in \Omega$ tal que $f(x^k) \leq -k$. Como a sequência $\{x^k\}$ está contida no compacto Ω, ela admite uma subsequência convergente para um ponto $\overline{x} \in \Omega$. Pela continuidade de f, a imagem dessa subsequência deve convergir para $f(\overline{x})$. No entanto, isso gera uma contradição, pois a suposição inicial implicaria que o limite fosse $-\infty$, enquanto $f(\overline{x})$ é um número real finito. Portanto, $f(\Omega)$ é limitado inferiormente. Defina $f^* = \inf\{f(x) \mid x \in \Omega\}$. Pela definição de ínfimo, para todo $k \in \mathbb{N}$, existe $x^k \in \Omega$ tal que

$$f^* \leq f(x^k) \leq f^* + \frac{1}{k}.$$

Fazendo $k \to \infty$, concluímos que $f(x^k) \to f^*$. Utilizando novamente o argumento de compacidade, existe uma subsequência de $\{x^k\}$ que converge para um ponto $x^* \in \Omega$. Pela unicidade do limite e continuidade de f, obtemos

$$f(x^*) = \lim f(x^k) = f^*.$$

Logo, $f(x^*) \leq f(x)$ para todo $x \in \Omega$, concluindo a demonstração. □

Definição 2.2 *Dizemos que uma função $f : \mathbb{R}^n \to \mathbb{R}$ é coerciva quando*

$$\lim_{\|x\| \to +\infty} f(x) = +\infty,$$

ou seja, para todo $M > 0$, existe $r > 0$ tal que $f(x) > M$ sempre que $\|x\| > r$.

Teorema 2.3 *Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função contínua e coerciva. Então, f admite um minimizador global.*

Prova. Fixe um ponto $a \in \mathbb{R}^n$ e defina $b = f(a)$. Pela coercividade de f, tomando $M = b$, garantimos a existência de um raio $r > 0$ tal que $f(x) > b$ sempre que $\|x\| > r$. Considere o conjunto $B = \{x \in \mathbb{R}^n \mid \|x\| \leq r\}$, que é compacto (fechado e limitado). Pelo Teorema de Weierstrass, a restrição de f ao compacto B assume um mínimo global em algum ponto $x^* \in B$. Logo,

$$f(x^*) \leq f(x), \quad \forall x \in B.$$

Note que $a \in B$. De fato, se $a \notin B$, teríamos $\|a\| > r$, o que implicaria $f(a) > b$, contradizendo a definição $f(a) = b$. Portanto, a está no conjunto B e $f(x^*) \leq f(a) = b$. Finalmente, para qualquer $x \notin B$, temos

$$f(x) > b = f(a) \geq f(x^*).$$

Concluímos que $f(x^*) \leq f(x)$ para todo $x \in \mathbb{R}^n$, provando que x^* é um minimizador global de f. □

Definição 2.4 *Se $D \subset \mathbb{R}^n$ é um conjunto convexo, diz-se que a função $f : D \to \mathbb{R}$ é convexa em D quando para quaisquer $x \in D, y \in D$ e $\alpha \in [0,1]$, tem-se que*

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

A função f é estritamente convexa quando a desigualdade acima é estrita para todos $x \neq y$ e $\alpha \in (0,1)$.

A função f é fortemente convexa com módulo $\gamma > 0$, qunado para quaisquer $x \in D$, $y \in D$ e $\alpha \in [0,1]$ tem-se

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) - \gamma\alpha(1-\alpha)\|x-y\|^2.$$

Teorema 2.5 *Sejam $D \subset \mathbb{R}^n$ um conjunto convexo aberto e $f : D \to \mathbb{R}$ uma função diferenciável em D. Então, as seguintes propriedades são equivalentes:*

(a) A função f é convexa em D.
(b) Para todo $x, y \in D$, tem-se $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.
(c) Para todo $x, y \in D$, tem-se $\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0$ (monotonicidade do gradiente).
(d) Se f é duas vezes diferenciável, a matriz Hessiana de f é semidefinida positiva em todo ponto de D, isto é:

$$\langle \nabla^2 f(x) d, d \rangle \geq 0, \quad \forall x \in D, \forall d \in \mathbb{R}^n.$$

Prova. Demonstraremos a cadeia de implicações $(a) \Rightarrow (b) \Rightarrow (c)$ e $(c) \Rightarrow (b) \Rightarrow (a)$.
(a) $\Rightarrow$ (b): Seja f convexa. Para quaisquer $x, y \in D$ e $\alpha \in (0,1]$, definindo $d = y - x$, temos

$$\begin{aligned} f(x + \alpha d) &= f(\alpha y + (1-\alpha)x) \\ &\leq \alpha f(y) + (1-\alpha) f(x). \end{aligned}$$

Reorganizando os termos, obtemos $\alpha(f(y) - f(x)) \geq f(x + \alpha d) - f(x)$. Dividindo por $\alpha > 0$ e tomando o limite $\alpha \to 0^+$, segue que

$$\begin{aligned} f(y) - f(x) &\geq \lim_{\alpha \to 0^+} \frac{f(x + \alpha d) - f(x)}{\alpha} \\ &= \langle \nabla f(x), d \rangle = \langle \nabla f(x), y - x \rangle, \end{aligned}$$

o que prova (b).
(b) $\Rightarrow$ (c): Permutando x e y na desigualdade de (b), temos

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

Somando esta desigualdade com a original de (b), obtemos

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

Simplificando, temos $\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0$.
(c) $\Rightarrow$ (b): Sejam $x, y \in D$. Pelo Teorema do Valor Médio, existe $\tau \in (0,1)$ tal que, definindo $z_\tau = x + \tau(y - x)$, temos

$$f(y) - f(x) = \langle \nabla f(z_\tau), y - x \rangle. \tag{2-1}$$

Aplicando a hipótese (c) para os pontos z_τ e x, temos $\langle \nabla f(z_\tau) - \nabla f(x), z_\tau - x \rangle \geq 0$. Como $z_\tau - x = \tau(y - x)$ e $\tau > 0$, segue que

$$\begin{aligned} \langle \nabla f(z_\tau) - \nabla f(x), y - x \rangle &\geq 0 \\ \langle \nabla f(z_\tau), y - x \rangle &\geq \langle \nabla f(x), y - x \rangle. \end{aligned}$$

Substituindo em (2-1), obtemos $f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle$.
(b) $\Rightarrow$ (a): Sejam $x, y \in D$ e $\alpha \in [0,1]$. Defina $z = (1-\alpha)x + \alpha y$. Aplicando (b) em x e z, e depois em y e z

$$\begin{aligned} f(x) &\geq f(z) + \langle \nabla f(z), x - z \rangle, \\ f(y) &\geq f(z) + \langle \nabla f(z), y - z \rangle. \end{aligned}$$

Multiplicando a primeira por $(1-\alpha)$ e a segunda por α, e somando

$$\begin{aligned}(1-\alpha)f(x)+\alpha f(y) &\geq f(z)+\langle\nabla f(z),(1-\alpha)(x-z)+\alpha(y-z)\rangle\\ &= f(z)+\langle\nabla f(z),0\rangle = f(z),\end{aligned}$$

o que confirma a convexidade.

Equivalência com (d): Suponha f duas vezes diferenciável. Mostremos $(b)\Rightarrow(d)$. Fixe $x\in D$ e $d\in\mathbb{R}^n$. Para α pequeno, $x+\alpha d\in D$. Por (b)

$$f(x+\alpha d)-f(x)-\alpha\langle\nabla f(x),d\rangle\geq 0.$$

Pela expansão de Taylor de segunda ordem

$$\frac{\alpha^2}{2}\langle\nabla^2 f(x)d,d\rangle+o(\alpha^2)\geq 0.$$

Dividindo por α^2 e fazendo $\alpha\to 0$, obtemos $\langle\nabla^2 f(x)d,d\rangle\geq 0$.

A volta $(d)\Rightarrow(b)$ segue da expansão de Taylor com resto de Lagrange: existe ξ no segmento entre x e y tal que

$$f(y)=f(x)+\langle\nabla f(x),y-x\rangle+\frac{1}{2}\langle\nabla^2 f(\xi)(y-x),y-x\rangle.$$

Como a Hessiana é semidefinida positiva, o termo quadrático é ≥ 0, implicando (b). □

Definição 2.6 *Dizemos que uma função diferenciável $f:\mathbb{R}^n\to\mathbb{R}$ possui gradiente Lipschitz contínuo se existir uma constante $L>0$ tal que*

$$\|\nabla f(x)-\nabla f(y)\|_2\leq L\|x-y\|_2,\quad\forall x,y\in\mathbb{R}^n.$$

A constante L é denominada constante de Lipschitz global do gradiente ∇f.

Conforme mencionado anteriormente, a função $g(x)$ no problema geral pode ser não suave, isto é, pode não possuir gradiente em todos os pontos. Para contornar essa limitação e permitir a análise de otimalidade, introduzimos os conceitos de subgradiente e subdiferencial.

Definição 2.7 *Seja $f:\mathbb{R}^n\to\mathbb{R}$ uma função convexa. Um vetor $v\in\mathbb{R}^n$ é dito um subgradiente de f no ponto $x\in\mathbb{R}^n$ se satisfizer a desigualdade:*

$$f(z)\geq f(x)+\langle v,z-x\rangle,\quad\forall z\in\mathbb{R}^n.$$

Definição 2.8 *O conjunto de todos os subgradientes de f em x é denominado subdiferencial de f em x e é denotado por $\partial f(x)$. Ou seja:*

$$\partial f(x) := \{v \in \mathbb{R}^n : f(z) \geq f(x) + \langle v, z - x\rangle, \forall z \in \mathbb{R}^n\}.$$

Vale notar que a desigualdade que define o subgradiente é uma generalização direta da propriedade de caracterização de primeira ordem de funções convexas diferenciáveis (Item (b) do Teorema 2.5), substituindo o gradiente $\nabla f(x)$ pelo vetor $v \in \partial f(x)$.

2.2 Operador Proximal

Nesta seção, será apresentado o operador proximal e descrita brevemente sua funcionalidade dentro do contexto de otimização.

O Algoritmo do Ponto Proximal é um método iterativo fundamental para minimizar funções convexas (mesmo que não diferenciáveis). A partir de um ponto inicial $x_0 \in \mathbb{R}^n$, o método gera recursivamente uma sequência $\{x_k\}$ que converge para um minimizador do problema.

Definição 2.9 *Seja $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ uma função convexa semicontínua inferiormente. O operador proximal de f em um ponto v, com parâmetro $\mu > 0$, é definido como:*

$$prox_\mu(v) := \arg\min_{u \in \mathbb{R}^n} \left\{ f(u) + \frac{1}{2\mu}\|u - v\|_2^2 \right\}. \tag{2-2}$$

Esse operador é a base para todos os métodos proximais. No contexto do algoritmo, a atualização é dada por

$$x_{k+1} = \text{prox}_{\mu_k}(x_k). \tag{2-3}$$

O Teorema 2.10 a seguir assegura a boa definição da sequência gerada por (2-2).

Teorema 2.10 *Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função convexa. A sequência $\{x_k\} \subset \mathbb{R}^n$ gerada pela iteração proximal está bem definida e é caracterizada pela relação:*

$$\frac{1}{\mu_k}(x_k - x_{k+1}) \in \partial f(x_{k+1}). \tag{2-4}$$

Prova. Para demonstrar a boa definição da sequência $\{x_k\}$, definimos, para cada iteração $k > 0$, a função auxiliar

$$\phi_k(u) = f(u) + \frac{1}{2\mu_k}\|u - x_k\|^2.$$

Note que ϕ_k é uma regularização aproximada de f. Primeiramente, mostremos que ϕ_k possui um minimizador. Dado x_k, existe $s_k \in \partial f(x_k)$ tal que, pela convexidade de f, tem-se $f(u) \geq f(x_k) + \langle s_k, u - x_k \rangle$. Substituindo em $\phi_k(u)$

$$\phi_k(u) \geq f(x_k) + \langle s_k, u - x_k \rangle + \frac{1}{2\mu_k}\|u - x_k\|^2.$$

Usando a desigualdade de Cauchy-Schwarz e evidenciando $\|u - x_k\|$ (com $u \neq x_k$) temos

$$\phi_k(u) \geq \|u - x_k\| \left(\frac{f(x_k)}{\|u - x_k\|} - \|s_k\| + \frac{1}{2\mu_k}\|u - x_k\| \right).$$

Passando ao limite quando $\|u\| \to +\infty$, obtemos

$$\lim_{\|u\| \to +\infty} \phi_k(u) = +\infty.$$

Isso significa que ϕ_k é coerciva. Como ϕ_k é contínua e coerciva, ela possui pelo menos um minimizador global. Além disso, como a função $\|\cdot\|^2$ é estritamente convexa e f é convexa, ϕ_k é estritamente convexa, o que garante que o minimizador x_{k+1} é único. Portanto, a sequência $\{x_k\}$ está bem definida. Sendo x_{k+1} o minimizador de ϕ_k, a condição de otimalidade implica que $0 \in \partial\phi_k(x_{k+1})$. Calculando o subdiferencial da soma

$$0 \in \partial f(x_{k+1}) + \nabla \left(\frac{1}{2\mu_k}\|u - x_k\|^2 \right) \Big|_{u = x_{k+1}}.$$

A derivada do termo quadrático é $(x_{k+1} - x_k)/\mu_k$. Logo

$$0 \in \partial f(x_{k+1}) + \frac{1}{\mu_k}(x_{k+1} - x_k) \iff \frac{1}{\mu_k}(x_k - x_{k+1}) \in \partial f(x_{k+1}).$$

□

2.3 Métodos de Primeira Ordem

Nesta seção, revisamos as estratégias fundamentais de otimização de primeira ordem que servem de alicerce para os algoritmos desenvolvidos neste trabalho. Partindo do clássico Método do Gradiente para funções diferenciáveis, estendemos a discussão para o Método do Gradiente Proximal (adequado para o problema composto definido em (1-1)) e, finalmente, apresentamos a variante acelerada FISTA, que supera as limitações de convergência dos métodos básicos.

2.3.1 Método do Gradiente

Uma das estratégias iterativas mais antigas e fundamentais para minimizar uma função diferenciável é o Método do Gradiente, também conhecido como Método de Cauchy ou *Steepest Descent*.

A ideia central do método consiste em gerar uma sequência de pontos $\{x_k\}$ movendo-se, a cada iteração, na direção oposta ao vetor gradiente da função objetivo, isto é, $d_k = -\nabla f(x_k)$. A justificativa teórica para essa escolha reside no fato de que, localmente, a direção do gradiente negativo corresponde à direção de máxima descida da função f em relação à norma euclidiana.

O Algoritmo 2.1 descreve a estrutura geral do método.

Algoritmo 2.1: Método do Gradiente

1 **Entrada:** $x_0 \in \mathbb{R}^n$ e tolerância $\epsilon > 0$;
2 $k \leftarrow 0$;
3 **while** $\|\nabla f(x_k)\| > \epsilon$ **do**
4 Defina a direção de busca: $d_k = -\nabla f(x_k)$;
5 Obtenha um tamanho de passo $\mu_k > 0$ tal que $f(x_k + \mu_k d_k) < f(x_k)$;
6 Atualize o iterado: $x_{k+1} = x_k + \mu_k d_k$;
7 $k \leftarrow k + 1$;
8 **end**
9 **Saída:** x_k (aproximação do minimizador).

2.3.2 Gradiente Proximal

O Método do Gradiente Proximal (AGP) pode ser interpretado como uma extensão do método do gradiente clássico para minimizar a função composta $F(x) := f(x) + g(x)$. Cada iteração do AGP consiste em um passo de gradiente na parte suave f, seguido pela aplicação do operador proximal associado à parte não suave g. Matematicamente, o iterado é calculado como:

$$\begin{aligned} p_{\mu_k}(x_k) &:= \operatorname{prox}_{\mu_k}(x_k - \mu_k \nabla f(x_k)) \\ &:= \arg\min_{u \in \mathbb{R}^n} \left\{ f(x_k) + \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{2\mu_k} \|u - x_k\|^2 + g(u) \right\}. \end{aligned} \tag{2-5}$$

A função objetivo do problema de minimização em (2-5) é denominada Aproximação Quadrática Composta de $F(x)$ em torno de x_k. Generalizando este conceito para uma matriz simétrica definida positiva H, definimos o modelo quadrático em um ponto v como

$$Q_H(u, v) := f(v) + \langle \nabla f(v), u - v \rangle + \frac{1}{2}\|u - v\|_H^2 + g(u). \tag{2-6}$$

Observe que, no caso do AGP, a matriz de curvatura é um múltiplo da identidade, isto é, $H = I/\mu_k$. Já para métodos do tipo Newton Proximal, teremos $H \approx \nabla^2 f(x)$. Dessa forma, podemos reescrever o passo proximal de forma unificada como

$$p_H(v) := \arg\min_{u \in \mathbb{R}^n} Q_H(u, v). \tag{2-7}$$

O Algoritmo 2.2 descreve o procedimento, incorporando uma estratégia de *backtracking* (busca linear) sobre o parâmetro de passo μ_k. O objetivo do backtracking é garantir que o passo gerado satisfaça a seguinte condição de decréscimo suficiente

$$F(p_{\mu_k}(x_k)) \leq Q_{\mu_k}(p_{\mu_k}(x_k), x_k), \tag{2-8}$$

onde denotamos Q_{μ_k} como o modelo Q_H com $H = I/\mu_k$.

Essa condição é fundamental para a análise de convergência global do método e é satisfeita quando $\mu_k \leq 1/L$, onde L é a constante de Lipschitz do gradiente de f. O uso do backtracking justifica-se pois L pode ser desconhecida ou estimar $\mu_k \leq 1/L$ pode resultar em passos excessivamente conservadores, retardando a convergência.

Algoritmo 2.2: Algoritmo do Gradiente Proximal (AGP)

Entrada: $x_0 \in \mathbb{R}^n$, $\beta \in (0,1)$ e $\mu_{init} > 0$;
for $k = 0, 1, 2, \ldots$ **do**
  Escolha uma estimativa inicial para o passo μ_k (ex: $\mu_k = \mu_{k-1}$ ou μ_{init});
  Calcule o candidato: $p_{\mu_k}(x_k) := \arg\min_{u \in \mathbb{R}^n} Q_{\mu_k}(u, x_k)$;
  while $F(p_{\mu_k}(x_k)) > Q_{\mu_k}(p_{\mu_k}(x_k), x_k)$ **do**
    Reduza o passo: $\mu_k \leftarrow \beta \mu_k$;
    Recalcule o candidato: $p_{\mu_k}(x_k) := \arg\min_{u \in \mathbb{R}^n} Q_{\mu_k}(u, x_k)$;
  end
  Atualize o iterado: $x_{k+1} \leftarrow p_{\mu_k}(x_k)$;
end
Saída: x_{k+1} (solução aproximada).

2.3.3 Gradiente Proximal Acelerado

A variante acelerada do AGP, denominada AGPA (Algoritmo do Gradiente Proximal Acelerado), distingue-se pela utilização de um passo de "momentum". Ao invés de avaliar o modelo quadrático Q_{H_k} no iterado atual x_{k-1}, o algoritmo utiliza um ponto de busca extrapolado y_k, construído a partir de uma combinação linear de iterações anteriores.

Essa abordagem, popularizada pelo algoritmo FISTA (*Fast Iterative Shrinkage-*

Thresholding Algorithm) proposto em [1], define o ponto de busca como

$$y_k = x_{k-1} + \alpha_{k-1}(x_{k-1} - x_{k-2}). \tag{2-9}$$

A sequência de parâmetros $\{\alpha_k\}$ é escolhida especificamente para garantir a aceleração da taxa de convergência global de $\mathcal{O}(1/k)$ (do AGP) para $\mathcal{O}(1/k^2)$, utilizando apenas informações de primeira ordem.

O Algoritmo 2.3 detalha o funcionamento do FISTA. Nele, o parâmetro α_k é definido pela sequência $\alpha_k = (t_k - 1)/t_{k+1}$, onde t_{k+1} é atualizado recursivamente.

Vale ressaltar que a versão original de [1] exige que o passo μ_k seja não-crescente ($\mu_{k+1} \leq \mu_k$) durante o *backtracking*. Contudo, estudos mais recentes, como o apresentado em [15], propõem generalizações para o parâmetro t_{k+1} que permitem relaxar esta condição (aceitando $\mu_{k+1} > \mu_k$) sem comprometer a taxa de convergência quadrática.

Algoritmo 2.3: FISTA (Fast Iterative Shrinkage-Thresholding Algorithm)

Entrada: $x_0 \in \mathbb{R}^n$, $\beta \in (0,1)$ e $\mu_1^0 > 0$;
Inicialização: $y_1 = x_0$, $t_1 = 1$;
for $k = 1, 2, \ldots$ **do**
  Defina a tentativa de passo: $\mu_k := \mu_k^0$;
  Calcule o candidato: $p_{\mu_k}(y_k) := \arg\min_{u \in \mathbb{R}^n} Q_{\mu_k}(u, y_k)$;
  while $F(p_{\mu_k}(y_k)) > Q_{\mu_k}(p_{\mu_k}(y_k), y_k)$ **do**
    Reduza o passo: $\mu_k \leftarrow \beta \mu_k$;
    Recalcule: $p_{\mu_k}(y_k) := \arg\min_{u \in \mathbb{R}^n} Q_{\mu_k}(u, y_k)$;
  end
  Atualize o iterado: $x_k \leftarrow p_{\mu_k}(y_k)$;
  Defina o passo inicial da próxima iteração: $\mu_{k+1}^0 := \mu_k$;
  Atualize o parâmetro de aceleração:

$$t_{k+1} := \frac{1}{2}\left(1 + \sqrt{1 + 4t_k^2}\right) \tag{2-10}$$

  Calcule o novo ponto de busca extrapolado:

$$y_{k+1} := x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}) \tag{2-11}$$

end
Saída: x_k.

2.4 Métodos de Segunda Ordem

Nesta seção, abordamos os métodos de segunda ordem, classe de algoritmos que utiliza informações de curvatura da função (via matriz Hessiana) para determinar a direção de busca. Em comparação aos métodos de primeira ordem, estas estratégias geralmente apresentam uma taxa de convergência superior (quadrática ou superlinear), exigindo menos iterações para atingir a solução. No entanto, o custo computacional por iteração tende a ser mais elevado devido ao cálculo e manipulação da matriz Hessiana. A seguir, discutimos o método de Newton clássico e introduzimos os métodos Quasi-Newton.

2.4.1 Método de Newton

O Método de Newton é a estratégia fundamental para otimização não linear irrestrita de segunda ordem. Diferente do método do gradiente, que aproxima a função localmente por um plano (linearização), o Método de Newton utiliza uma aproximação quadrática da função objetivo f em torno do iterado atual x_k.

A direção de busca d_k (denominada direção de Newton) é obtida minimizando essa aproximação quadrática, o que resulta na solução do sistema

$$\nabla^2 f(x_k) d_k = -\nabla f(x_k), \tag{2-12}$$

ou, equivalentemente, assumindo que a Hessiana é invertível

$$d_k = -(\nabla^2 f(x_k))^{-1} \nabla f(x_k). \tag{2-13}$$

A atualização do iterado segue a forma padrão $x_{k+1} = x_k + \mu_k d_k$, onde μ_k é o tamanho do passo. Quando $\mu_k = 1$ é aceito, dizemos que foi dado um passo de Newton puro. O Algoritmo 2.4 descreve a estrutura geral do método.

Algoritmo 2.4: Método de Newton

Entrada: $x_0 \in \mathbb{R}^n$ e tolerância $\epsilon > 0$;
$k \leftarrow 0$;
while $\|\nabla f(x_k)\| > \epsilon$ **do**
    Calcule a Hessiana $\nabla^2 f(x_k)$ e o gradiente $\nabla f(x_k)$;
    Obtenha a direção d_k resolvendo o sistema linear:

$$\nabla^2 f(x_k) d_k = -\nabla f(x_k)$$

    Defina o tamanho do passo $\mu_k > 0$ (via busca linear);
    Atualize: $x_{k+1} = x_k + \mu_k d_k$;
    $k \leftarrow k + 1$;
end
Saída: x_k.

Vale ressaltar um aspecto computacional crucial: na prática, nunca se calcula explicitamente a matriz inversa $(\nabla^2 f(x_k))^{-1}$. A direção d_k é obtida resolvendo o sistema de equações lineares $\nabla^2 f(x_k) d_k = -\nabla f(x_k)$ (por exemplo, via fatoração de Cholesky), o que é computacionalmente mais eficiente e numericamente mais estável.

2.4.2 Métodos Quasi-Newton

Os métodos Quasi-Newton representam uma alternativa eficiente ao Método de Newton clássico. A principal vantagem desta classe de algoritmos reside no fato de não exigirem o cálculo explícito da Hessiana exata $\nabla^2 f(x_k)$. Mais especificamente, enquanto a resolução do sistema linear em (2-12) exige, em geral, $\mathcal{O}(n^3)$ operações aritméticas por iteração devido a necessidade de fatoração da matriz, os métodos quasi-Newton permitem obter a direção de busca resolvendo o sistema aproximado $B_k d_k = -\nabla f(x_k)$ com um custo computacional de apenas $\mathcal{O}(n^2)$ operações [12]. Essa drástica redução no custo por iteração é alcançada mantendo, contudo, uma taxa de convergência superlinear (superior à do método do gradiente).

Ao invés de utilizar a Hessiana exata, esses métodos constroem e atualizam iterativamente uma aproximação matricial $B_k \approx \nabla^2 f(x_k)$. Essa atualização baseia-se na variação observada nos gradientes entre iterações sucessivas para capturar a curvatura da função ao longo da direção de busca. Dentre as diversas fórmulas de atualização propostas na literatura, temos a fórmula BFGS.

BFGS

O método BFGS (nomeado em homenagem a seus criadores: Broyden, Fletcher, Goldfarb e Shanno) utiliza uma estratégia de atualização de posto-2 para construir a aproximação da Hessiana. A fórmula de atualização para a matriz B_{k+1}, dada a aproximação anterior B_k, é definida por

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}, \tag{2-14}$$

onde os vetores de deslocamento e variação do gradiente são, respectivamente

$$s_k = x_{k+1} - x_k \quad \text{e} \quad y_k = \nabla f(x_{k+1}) - \nabla f(x_k). \tag{2-15}$$

Essa fórmula satisfaz a chamada *equação da secante* ($B_{k+1} s_k = y_k$) e preserva a simetria da matriz. Além disso, uma propriedade fundamental do BFGS é a hereditariedade da definição positiva: se a aproximação inicial B_0 for definida positiva e a condição de curvatura $s_k^T y_k > 0$ for satisfeita, todas as matrizes subsequentes B_k serão definidas positivas.

L-BFGS

O método L-BFGS (*Limited-memory BFGS*) é uma adaptação do algoritmo BFGS projetada especificamente para problemas de grande porte, onde o custo de armazenar ou manipular uma matriz densa $n \times n$ é proibitivo.

Diferente do BFGS padrão, que carrega a matriz B_k completa, o L-BFGS armazena a aproximação da Hessiana implicitamente. A ideia central consiste em manter na memória apenas um conjunto limitado de m pares de vetores de correção $\{s_i, y_i\}$ das iterações mais recentes (com m tipicamente entre 3 e 20). As informações de curvatura mais antigas, consideradas menos relevantes para o comportamento local atual da função, são descartadas para economizar memória e custo computacional.

Operacionalmente, produtos matriciais envolvendo B_k (ou sua inversa) são computados através de algoritmos recursivos de dois laços (*two-loop recursion*) que utilizam apenas os pares armazenados $\{s_i, y_i\}$ e uma matriz inicial esparsa (geralmente um múltiplo da identidade), sem nunca construir a matriz densa explicitamente. Embora o L-BFGS apresente, teoricamente, uma taxa de convergência linear (devido à perda de informação antiga), na prática, ele exibe um desempenho muito superior ao método do gradiente e frequentemente próximo ao do BFGS completo para muitas aplicações.

Vale ressaltar que as atualizações BFGS e L-BFGS descritas acima foram originalmente desenvolvidas para problemas de otimização suave (sem a componente não

diferenciável $g(x)$). No contexto deste trabalho, utilizaremos essas fórmulas não para definir a direção de descida direta $d_k = -B_k^{-1}\nabla f_k$, mas sim para construir a matriz de métrica H_k na aproximação quadrática (2-5). Nos capítulos seguintes, detalharemos como essas aproximações são integradas à estrutura do Algoritmo Quasi-Newton Proximal para resolver o problema composto (1-1).

CAPÍTULO 3

Método de Quasi-Newton Proximal

Neste capítulo, dedicamo-nos à construção e análise de um método Quasi-Newton Proximal para a resolução do problema de otimização composta (1-1). Apresentamos a estrutura formal do algoritmo, detalhando as estratégias de atualização da matriz de aproximação e do parâmetro de passo, seguidas pela análise teórica de sua taxa de convergência. Este capítulo é baseado no artigo [17].

Os métodos Quasi-Newton Proximais podem ser vistos como uma generalização do Método de Newton, integrando informações de curvatura de segunda ordem (aproximadas) diretamente na estrutura do operador proximal. Diferentemente do método de Newton clássico, que exigiria o cálculo da Hessiana exata, esta classe de algoritmos utiliza uma matriz simétrica definida positiva H_k para aproximar a geometria local da função suave f.

A cada iteração k, o método considera uma aproximação quadrática do termo suave em torno do iterado atual x_k adicionado à função não-suave:

$$Q_{H_k}(u, x_k) := f(x_k) + \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{2}\|u - x_k\|^2_{H_k} + g(u). \tag{3-1}$$

A estrutura algorítmica geral pode ser descrita em quatro etapas fundamentais:

- **Modelagem:** A função suave $f(x)$ é aproximada localmente pelo modelo quadrático convexo $Q_{H_k}(u, x_k)$ (excluindo a parte não suave $g(u)$ que é mantida exata).
- **Subproblema:** Um método de otimização interno (iterativo) é aplicado para minimizar o modelo Q_{H_k}, gerando um ponto teste candidato a x_{k+1}.
- **Critério de Aceitação:** O ponto teste é aceito como novo iterado x_{k+1} se satisfizer uma condição de decréscimo suficiente (descrita adiante).
- **Atualização:** Caso o ponto não seja aceito, o parâmetro de regularização (passo) ou a matriz H_k são ajustados para aumentar a convexidade do modelo, e um novo ponto teste é calculado.

Utilizando a notação de minimização do modelo, reescrevemos o passo proximal

generalizado como:

$$p_H(v) := \arg\min_{u \in \mathbb{R}^n} Q_H(u, v). \tag{3-2}$$

Os Algoritmos 3.1 e 3.2 a seguir formalizam, respectivamente, a estrutura externa (Outer Loop) e o procedimento de busca linear (Inner Loop) do método proposto.

Algoritmo 3.1: Método de Quasi-Newton Proximal (Estrutura Geral)

Entrada: $x_0 \in \mathrm{dom}(F)$ e parâmetro de descida $\rho \in (0, 1]$;
for $k = 0, 1, 2, \ldots$ **do**
    Escolha uma estimativa de passo inicial $\mu_k > 0$;
    Escolha uma matriz de aproximação $G_k \succeq 0$;
    Execute o procedimento de busca linear (Algoritmo 3.2) com parâmetros (μ_k, G_k, x_k, ρ) para obter o par efetivo (H_k, x_{k+1});
end

Algoritmo 3.2: Busca Linear e Atualização do Parâmetro Proximal

Entrada: μ, G, x, ρ, e escolha um fator de redução $\beta \in (0, 1)$;
Defina $H := G + \frac{1}{\mu} I$;
Calcule o candidato $p(x) := p_H(x)$, onde p_H é definido em (3-2);
while $F(p(x)) > F(x) - \rho(F(x) - Q_H(p(x), x))$ **do**
    Reduza o passo: $\mu \leftarrow \beta\mu$;
    Atualize a matriz: $H := G + \frac{1}{\mu} I$;
    Recalcule o candidato: $p(x) := p_H(x)$;
end
Retorne: H e $p(x)$.

É importante notar que a matriz utilizada no modelo é construída como $H_k = G_k + \frac{1}{\mu_k} I$. Como impomos $G_k \succeq 0$ (semidefinida positiva), a matriz resultante H_k é necessariamente definida positiva, com seu menor autovalor limitado inferiormente por um constante positiva. Portanto, controlar o tamanho do passo μ_k (reduzindo-o) equivale a aumentar o parâmetro de convexidade forte do modelo, tornando o passo mais conservador e garantindo a condição de decréscimo.

De fato, para fins teóricos, qualquer método de construção de H_k é admissível, desde que a matriz resultante satisfaça as cotas espectrais $mI \preceq H_k \preceq MI$ (com $m, M > 0$) e que o ponto gerado satisfaça a condição de decréscimo suficiente especificada no Algoritmo 3.2.

Vale ressaltar que a etapa principal do Algoritmo 3.1, que consiste no cálculo do operador proximal generalizado $x_{k+1} = p_{H_k}(x_k)$, está bem definida para todas as iterações. A função objetivo do subproblema, dada por Q_{H_k}, é a soma de uma função quadrática definida positiva e da função convexa $g(u)$ resultando numa função fortemente convexa

em $\mathbb{R}^n$. Pela teoria clássica de otimização convexa, o problema de minimizar uma função fortemente convexa, fechada e própria sobre todo o espaço $\mathbb{R}^n$ possui, necessariamente, um único minimizador global. Portanto, o próximo iterado x_{k+1} existe e é único. Também é importante mencionar que a boa definição do Algoritmo 3.2 é assegurada devido à propriedade de o gradiente de f ser Lipschitz contínuo, o que garante que o loop seja completado em um número finito de passos.

Para realizar a análise de convergência do algoritmo, além das condições estabelecidas para o problema (1-1), assumiremos as seguintes hipóteses adicionais:

Hipótese 1 *(i) A função g é Lipschitz contínua em $\mathbb{R}^n$ (ou no conjunto de interesse χ_0), com constante $L_g > 0$. Equivalentemente, a norma de qualquer subgradiente de g é limitada, isto é:*

$$\|\nu\| \leq L_g, \quad \forall \nu \in \partial g(x), \forall x \in \mathrm{dom}(g).$$

(ii) Existem constantes positivas $0 < m \leq M$ tais que as matrizes de aproximação satisfazem:

$$mI \preceq H_k \preceq MI, \quad \forall k \geq 0.$$

(iii) Existe uma constante $D > 0$ tal que $\sup_{x^* \in X^*} \|x_k - x^*\| \leq D$ para todo k.

A Hipótese 1(iii), é naturalmente satisfeita caso o conjunto de nível inicial, definido por $\chi_0 = \{x \in \mathrm{dom}(F) : F(x) \leq F(x_0)\}$, seja limitado. Como o método garante o decréscimo monótono da função objetivo a cada iteração, toda sequência gerada x_k permanece contida no conjunto χ_0. Se χ_0 for um conjunto limitado, a distância máxima entre qualquer iterado x_k e uma solução ótima x^* (que também pertence a χ_0) será necessariamente limitada por uma constante D>0. Vale ressaltar que essa condição de limitação do conjunto de nível é garantida, por exemplo, sempre que a função objetivo $F(x)$ for coerciva.

A seguir, apresentamos um resultado fundamental baseado na condição de otimalidade de primeira ordem do subproblema proximal.

Lema 3.1 *Para qualquer $v \in \mathbb{R}^n$, definindo $p = p_H(v)$, existe um subgradiente $\nu_g(p) \in \partial g(p)$ tal que:*

$$\nabla f(v) + H(p - v) + \nu_g(p) = 0.$$

Prova. Pela definição do operador proximal generalizado, temos:

$$p_H(v) = \arg\min_{u \in \mathbb{R}^n} Q_H(u, v) := \left\{ f(v) + \langle \nabla f(v), u - v \rangle + \frac{1}{2}\|u - v\|_H^2 + g(u) \right\}.$$

A condição de otimalidade para este problema de minimização convexa irrestrita requer que o zero pertença ao subdiferencial da função objetivo em relação a u, avaliado no ponto ótimo $p_H(v)$. O gradiente da parte diferenciável em relação a u é $\nabla f(v) + H(p_H(v) - v)$. Assim, a condição de otimalidade é dada por:

$$0 \in \nabla f(v) + H(p_H(v) - v) + \partial g(p_H(v)).$$

Isto implica a existência de um vetor $\nu_g(p_H(v)) \in \partial g(p_H(v))$ que satisfaz a igualdade enunciada. □

A análise a seguir estabelece uma cota superior para o erro da função objetivo utilizando apenas informações de primeira ordem, o que será crucial para demonstrar a taxa de convergência sublinear.

Lema 3.2 *Para quaisquer três pontos $u, v, w \in \mathrm{dom}(F)$, e para qualquer subgradiente $\nu_g(v) \in \partial g(v)$, vale a desigualdade:*

$$F(u) - F(w) \leq \|\nabla f(u) + \nu_g(v)\|\|u - w\| + 2L_g\|u - v\|.$$

Prova. Utilizando a convexidade da função diferenciável f, temos:

$$f(u) - f(w) \leq \langle \nabla f(u), u - w \rangle. \tag{3-3}$$

Pela convexidade de g, para qualquer subgradiente $\nu_g(v) \in \partial g(v)$, temos $g(w) \geq g(v) + \langle \nu_g(v), w - v \rangle$, o que implica:

$$g(v) - g(w) \leq \langle \nu_g(v), v - w \rangle. \tag{3-4}$$

Por outro lado, usando a definição de F, obtemos:

$$\begin{aligned} F(u) - F(w) &= f(u) - f(w) + g(u) - g(w) \\ &= f(u) - f(w) + [g(u) - g(v)] + [g(v) - g(w)], \end{aligned}$$

o que combinado com as desigualdades (3-3) e (3-4) implica

$$F(u) - F(w) \leq \langle \nabla f(u), u - w \rangle + [g(u) - g(v)] + \langle \nu_g(v), v - w \rangle.$$

Ou equivalentemente,

$$F(u) - F(w) \leq \langle \nabla f(u) + \nu_g(v), u - w \rangle + [g(u) - g(v)] + \langle \nu_g(v), v - u \rangle.$$

Agora, utilizamos a propriedade Lipschitz de g (Hipótese 1(i)), temos que $|g(u) - g(v)| \leq L_g\|u - v\|$ e, pela desigualdade de Cauchy-Schwarz combinada com a limitação do subgradiente ($\|\nu_g(v)\| \leq L_g$), temos

$$\begin{aligned} F(u) - F(w) &\leq \|\nabla f(u) + \nu_g(v)\|\|u - w\| + L_g\|u - v\| + L_g\|u - v\| \\ &= \|\nabla f(u) + \nu_g(v)\|\|u - w\| + 2L_g\|u - v\|, \end{aligned}$$

o que prova a desigualdade desejada. □

O próximo resultado garante que o modelo quadrático decresce suficientemente a cada iteração e estabelece uma relação entre o tamanho do passo e a norma do gradiente composto (resíduo de otimalidade).

Lema 3.3 *Seja x_{k+1} a solução do subproblema proximal em x_k com matriz H_k. Então, o decréscimo do modelo satisfaz:*

$$Q_{H_k}(x_k, x_k) - Q_{H_k}(x_{k+1}, x_k) \geq \frac{m}{2}\|x_{k+1} - x_k\|^2.$$

Além disso, existe um subgradiente $\nu_g(x_{k+1}) \in \partial g(x_{k+1})$ tal que:

$$\frac{1}{M}\|\nabla f(x_k) + \nu_g(x_{k+1})\| \leq \|x_{k+1} - x_k\| \leq \frac{1}{m}\|\nabla f(x_k) + \nu_g(x_{k+1})\|. \tag{3-5}$$

Prova. A função $Q_{H_k}(\cdot, x_k)$ é fortemente convexa. Como x_{k+1} é o minimizador global de $Q_{H_k}(\cdot, x_k)$, segue da propriedade de convexidade forte que, para qualquer u:

$$Q_{H_k}(u, x_k) \geq Q_{H_k}(x_{k+1}, x_k) + \frac{1}{2}\|u - x_{k+1}\|^2_{H_k}.$$

Tomando $u = x_k$ e utilizando a propriedade da norma induzida pela matriz ($\|v\|^2_{H_k} \geq m\|v\|^2$), obtemos:

$$\begin{aligned} Q_{H_k}(x_k, x_k) - Q_{H_k}(x_{k+1}, x_k) &\geq \frac{1}{2}\|x_k - x_{k+1}\|^2_{H_k} \\ &\geq \frac{m}{2}\|x_k - x_{k+1}\|^2, \end{aligned}$$

o que prova a primeira parte do lema. Para a segunda parte, lembramos a condição de otimalidade de primeira ordem para x_{k+1}:

$$\nabla f(x_k) + H_k(x_{k+1} - x_k) + \nu_g(x_{k+1}) = 0, \quad \text{para algum } \nu_g(x_{k+1}) \in \partial g(x_{k+1}).$$

Ou equivalentemente, $H_k(x_{k+1} - x_k) = -(\nabla f(x_k) + \nu_g(x_{k+1}))$. Denotando $v := x_{k+1} - x_k$,

segue que

$$m\|v\|^2 \leq \|v\|_{H_k}^2 = \langle H_k v, v\rangle = -\langle \nabla f(x_k) + \nu_g(x_{k+1}), v\rangle \leq \|\nabla f(x_k) + \nu_g(x_{k+1})\|\|v\|,$$

a qual implica a segunda desigualdade em (3-5). Para a primeira desigualdade, como H_k é definida positiva (em particular, invertível), temos

$$v = x_{k+1} - x_k = -H_k^{-1}(\nabla f(x_k) + \nu_g(x_{k+1}))$$

Como $H_k \preceq MI$ temos que $H_k^{-1} \succeq (1/M)I$. Daí,

$$\langle H_k^{-1}(\nabla f(x_k) + \nu_g(x_{k+1})), \nabla f(x_k) + \nu_g(x_{k+1})\rangle \geq \frac{1}{M}\|\nabla f(x_k) + \nu_g(x_{k+1})\|^2.$$

Substituindo $H_k^{-1}(\nabla f(x_k) + \nu_g x_{k+1}) = -v$, obtemos

$$\frac{1}{M}\|\nabla f(x_k) + \nu_g(x_{k+1})\|^2 \leq \langle -v, \nabla f(x_k) + \nu_g(x_{k+1})\rangle \leq \|v\|\|(\nabla f(x_k) + \nu_g(x_{k+1})\|$$

o que implica a primeira desigualdade em (3-5). □

Por fim, o Teorema 3.4 estabelece a convergência global de ordem $\mathcal{O}(1/k)$ para o Algoritmo 3.1.

Teorema 3.4 *A sequência $\{x_k\}$ gerada pelo Algoritmo 3.1 satisfaz*

$$F(x_k) - F(x^*) \leq \frac{C}{k}, \quad \forall k \geq 1,$$

onde $x^* \in X^*$ e a constante C é definida por

$$C := \frac{2M^2(DM + 2L_g)^2}{\rho m^3},$$

e D, m, M e L_g são como na Hipótese 1.

Prova. Defina $\Delta F_k := F(x_k) - F(x^*)$. Aplicando o Lema 3.2 com $u = x_k$, $w = x^*$ e $v = x_{k+1}$, obtemos

$$\Delta F_k \leq \|\nabla f(x_k) + \nu_g(x_{k+1})\|\|x_k - x^*\| + 2L_g\|x_k - x_{k+1}\|.$$

Pela Hipótese 1(iii), temos $\|x_k - x^*\| \leq D$. Por outro lado, do Lema 3.3, sabemos que $\|x_{k+1} - x_k\| \leq \|\nabla f(x_k) + \nu_g(x_{k+1})\|/m$. Combinando estas desigualdades, obtemos

$$\Delta F_k \leq \|\nabla f(x_k) + \nu_g(x_{k+1})\|\left(D + \frac{2L_g}{m}\right) \leq \|\nabla f(x_k) + \nu_g(x_{k+1})\|\left(\frac{MD + 2L_g}{m}\right)$$

onde a última desigualdade é devido ao fato de que $mD \leq MD$. Portanto,

$$\|\nabla f(x_k) + \nu_g(x_{k+1})\|^2 \geq \frac{\Delta F_k^2 m^2}{(MD + 2L_g)^2}. \tag{3-6}$$

Por outro lado, segue da definição de ΔF_k, do Algoritmo 3.2 e do fato que $Q_{H_k}(x_k, x_k) = F(x_k)$, que

$$\Delta F_k - \Delta F_{k+1} = F(x_k) - F(x_{k+1}) \geq \rho(Q_{H_k}(x_k, x_k) - Q_{H_k}(x_{k+1}, x_k)),$$

o que combinado com Lema 3.3 implica que

$$\Delta F_k - \Delta F_{k+1} \geq \frac{\rho m}{2M^2}\|\nabla f(x_k) + \nu_g(x_{k+1})\|^2.$$

Combinando a última desigualdade com (3-6), temos

$$\Delta F_k - \Delta F_{k+1} \geq \left(\frac{\rho m}{2M^2}\right)\frac{m^2}{(MD + 2L_g)^2}(\Delta F_k)^2,$$

ou, equivalentemente,

$$\Delta F_k - \Delta F_{k+1} \geq \underbrace{\frac{\rho m^3}{2M^2(DM + 2L_g)^2}}_{1/C}(\Delta F_k)^2.$$

Dividindo a desigualdade acima por $\Delta F_{k+1}\Delta F_k$

$$\frac{1}{\Delta F_{k+1}} - \frac{1}{\Delta F_k} \geq \frac{1}{C}\frac{\Delta F_k}{\Delta F_{k+1}} \geq \frac{1}{C}$$

Fazendo a soma da última desigualdade de $k = 0, 1, \ldots k-1$, temos

$$\frac{1}{\Delta F_k} \geq \frac{1}{\Delta F_k} - \frac{1}{\Delta F_0} \geq \sum_{i=0}^{k-1}\frac{1}{C} = \frac{k}{C},$$

o que implica a desigualdade desejada. □

CAPÍTULO 4

Método Quasi-Newton Proximal Acelerado

Neste capítulo, abordamos uma variação acelerada do método quasi-Newton proximal apresentado no Capítulo 3. Discutimos as hipóteses necessárias, a estrutura do algoritmo e a análise teórica de sua taxa de convergência. Este capítulo é baseado no artigo [3].

A premissa central do método quasi-Newton proximal acelerado é incorporar um passo de extrapolação ao método padrão para obter taxas de convergência superiores. Em [6], os autores introduziram essa abordagem, alcançando uma taxa de convergência de $\mathcal{O}(1/k^2)$ para o Newton proximal. No entanto, esse resultado teórico depende de condições bastante restritivas. Especificamente, exige-se que a estimativa da Hessiana H_k satisfaça a condição de monotonicidade $0 \prec H_k \preceq H_{k-1}$ a cada iteração k.

Simultaneamente, as matrizes H_k devem ser escolhidas de modo a garantir que o modelo quadrático seja uma função majorante de F (suficientemente curva). Essas duas condições podem ser conflitantes, a menos que a sequência $\{H_k\}$ seja composta por matrizes com autovalores excessivamente grandes, o que resultaria em passos muito curtos. No caso particular em que $H_k = I/\mu_k$, a condição $H_k \preceq H_{k-1}$ implica $\mu_k \geq \mu_{k-1}$, ou seja, tamanhos de passo não decrescentes. Isso contradiz a estratégia padrão do algoritmo FISTA, do qual a aceleração é derivada, que geralmente requer $\mu_k \leq \mu_{k-1}$ (passos não crescentes).

Aqui, baseamo-nos na abordagem de [3], que introduz uma versão do método quasi-Newton proximal acelerado com restrições relaxadas. Essa estratégia fundamenta-se na variante do FISTA proposta em [15], a qual permite a não-monotonicidade do parâmetro de passo, possibilitando uma escolha mais flexível e eficiente para a sequência de matrizes H_k.

Para realizar a análise de convergência do algoritmo, além das condições estabelecidas para o problema (1-1), assumiremos a seguinte hipótese adicional:

Hipótese 2 *Existem constantes positivas m e M tais que, para todo $k \geq 0$, tem-se $mI \preceq H_k \preceq MI$.*

O Algoritmo 4.1 descreve o método adotado. Note que este algoritmo permite atualizar dinamicamente o ponto de extrapolação y_k durante a fase de *backtracking*, o que é crucial para garantir a convergência sem impor a monotonicidade estrita das matrizes.

Novamente, a seguinte aproximação quadrática do termo suave em torno do iterado atual x_k adicionado com a função não-suave será necessário:

$$Q_{H_k}(u, x_k) := f(x_k) + \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{2}\|u - x_k\|^2_{H_k} + g(u). \tag{4-1}$$

Algoritmo 4.1: Algoritmo Quasi-Newton Proximal Acelerado

Inicialização: Defina $t_1 = 1$, $\theta_0 = 1$, escolha um escalar inicial $\sigma_0 = \sigma_1^0 > 0$, defina $y_1 = x_0 \in \mathrm{dom}(F)$ (com $x_{-1} = x_0$) e uma matriz inicial $H_0 = H_1 \succ 0$. Escolha o fator de redução $\beta \in (0,1)$;

for $k = 1, 2, \ldots$ **do**

- Defina a estimativa inicial: $\sigma_k = \sigma_k^0$;
- Calcule o candidato: $p_{H_k}(y_k) = \arg\min_{u \in \mathbb{R}^n} Q_{H_k}(u, y_k)$;
- **while** $F(p_{H_k}(y_k)) > Q_{H_k}(p_{H_k}(y_k), y_k)$ **do**
  - Aumente a curvatura do modelo: $H_k \leftarrow \frac{1}{\beta} H_k$;
  - Ajuste σ_k (se necessário) para manter a coerência da aceleração, de modo que $\sigma_k H_k \preceq \sigma_{k-1} H_{k-1}$;
  - Atualize o parâmetro $\theta_{k-1} = \frac{\sigma_{k-1}}{\sigma_k}$;
  - Recalcule t_k e y_k utilizando as equações (4-2) e (4-3) com o novo θ_{k-1};
  - Recalcule o candidato: $p_{H_k}(y_k) = \arg\min_{u \in \mathbb{R}^n} Q_{H_k}(u, y_k)$;
- **end**
- Defina o novo iterado: $x_k = p_{H_k}(y_k)$;
- Escolha $\sigma_{k+1}^0 > 0$ e H_{k+1} satisfazendo $\sigma_{k+1}^0 H_{k+1} \preceq \sigma_k^0 H_k$;
- Defina $\theta_k = \frac{\sigma_k}{\sigma_{k+1}^0}$;
- Calcule os parâmetros de aceleração para a próxima iteração:

$$t_{k+1} := \frac{1}{2}\left(1 + \sqrt{1 + 4\theta_k t_k^2}\right) \tag{4-2}$$

  e o ponto extrapolado:

$$y_{k+1} := x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}) \tag{4-3}$$

end

A condição central imposta do Algoritmo 4.1 na sequência de matrizes $\{H_k\}$ é a desigualdade $\sigma_{k+1} H_{k+1} \preceq \sigma_k H_k$. O parâmetro de relaxamento $\theta_k = \sigma_k / \sigma_{k+1}$ desempenha

um papel fundamental, pois impacta diretamente o cálculo do escalar de aceleração t_{k+1} em (4-2).

Uma consequência computacional importante dessa abordagem reside na fase de *backtracking*. Como a atualização dos parâmetros ocorre dentro do loop interno, pode ser necessário recalcular o ponto extrapolado y_k — e, consequentemente, avaliar o gradiente $\nabla f(y_k)$ — múltiplas vezes em uma única iteração, elevando o custo computacional se comparado a métodos não acelerados.

Vale ressaltar que o algoritmo não fixa uma regra explícita para a construção da aproximação H_k, exigindo apenas que a condição de monotonicidade escalonada ($\sigma_{k+1}H_{k+1} \preceq \sigma_k H_k$) seja satisfeita. No entanto, existe uma restrição de causalidade: em geral a Hessiana exata avaliada no ponto futuro y_{k+1} não pode ser utilizada como H_{k+1}, uma vez que a definição de y_{k+1} depende, via t_{k+1} e θ_k, da escolha prévia de H_{k+1}.

Ao desativarmos o mecanismo de aceleração no Algoritmo 4.1, o que equivale a fixar o parâmetro $t_k = 1$ para todas as iterações, fazendo com que o ponto de extrapolação coincida com o iterado atual ($y_{k+1} = x_k$). A iteração do método acelerado é similar para a estrutura do Método Quasi-Newton Proximal padrão discutido no Capítulo 3. A distinção remanescente entre as duas abordagens reside, portanto, exclusivamente na estratégia de construção da matriz de aproximação H_k: enquanto o método do Capítulo 3 prescreve uma atualização baseada na regularização proximal explícita ($H_k = G_k + I/\mu_k$), o método acelerado deste capítulo permite uma construção mais flexível, desde que a sequência de matrizes respeite as condições impostas ($\sigma_{k+1}H_{k+1} \preceq \sigma_k H_k$) necessárias para garantir a estabilidade da aceleração.

Uma escolha trivial para a sequência $\{H_k\}$ é definir $H_k = I/\mu_k$. Essa escolha reduz o Algoritmo 4.1 à versão do FISTA apresentada na Seção 2.3.3, incorporando o *backtracking* completo do tamanho de passo proposto em [15]. Portanto, o Algoritmo 4.1 também pode ser visto como uma generalização direta do FISTA para métricas variáveis.

Alternativamente, uma escolha estruturada para a aproximação da Hessiana é definir $H_k = H/\sigma_k$, onde H é uma matriz definida positiva fixa (independente de k). Essa estrutura satisfaz automaticamente a condição de monotonicidade $\sigma_{k+1}H_{k+1} \preceq \sigma_k H_k$ (desde que H_{k+1} e H_k compartilhem a mesma matriz base H, a desigualdade se reduz a uma identidade ou desigualdade escalar simples, dependendo de como σ_k é atualizado). O Algoritmo 4.1 especializa-se, então, na versão simplificada descrita no Algoritmo 4.2.

Algoritmo 4.2: Algoritmo de Quasi-Newton Proximal Acelerado com Hessiana Fixa

Inicialização: $t_1 = 1$, $\theta_0 = 1$, $\sigma_1^0 > 0$, $y_1 = x_0 \in \mathrm{dom}(F)$ (com $x_{-1} = x_0$), matriz fixa $H \succ 0$ e fator $\beta \in (0,1)$;
for $k = 1, 2, \ldots$ **do**
  Defina a estimativa inicial: $\sigma_k = \sigma_k^0$;
  Calcule a matriz atual: $H_k = \frac{1}{\sigma_k} H$;
  Calcule o candidato: $p_{H_k}(y_k) = \arg\min_{u \in \mathbb{R}^n} Q_{H_k}(u, y_k)$;
  while $F(p_{H_k}(y_k)) > Q_{H_k}(p_{H_k}(y_k), y_k)$ **do**
    Reduza o parâmetro (aumentando a curvatura de H_k): $\sigma_k = \beta \sigma_k$;
    Atualize o parâmetro $\theta_{k-1} = \frac{\sigma_{k-1}}{\sigma_k}$;
    Recalcule t_k e y_k utilizando as equações (4-4) e (4-5);
    Atualize a matriz: $H_k = \frac{1}{\sigma_k} H$;
    Recalcule o candidato: $p_{H_k}(y_k) = \arg\min_{u \in \mathbb{R}^n} Q_{H_k}(u, y_k)$;
  end
  Defina o novo iterado: $x_k = p_{H_k}(y_k)$;
  Escolha $\sigma_{k+1}^0 > 0$, defina $\theta_k = \frac{\sigma_k}{\sigma_{k+1}^0}$;
  Calcule os parâmetros de aceleração:

$$t_{k+1} = \frac{1}{2}\left(1 + \sqrt{1 + 4\theta_k t_k^2}\right) \tag{4-4}$$

$$y_{k+1} = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}) \tag{4-5}$$

end

A boa definição do candidato $p_{H_k}(y_k)$ envolve a minimização do modelo Q_{H_k}. Embora o ponto de referência y_k mude dinamicamente devido ao processo de *backtracking*, a estrutura do subproblema permanece inalterada. A condição imposta às matrizes de aproximação, $H_k \succeq mI$ (ou, no caso de Hessiana Fixa, $H_k = H/\sigma_k$ com $\sigma_k > 0$), assegura que a matriz da métrica seja sempre definida positiva. Isso confere ao modelo Q_{H_k} a propriedade de convexidade forte em relação à variável de decisão u. Dessa forma, independentemente do valor de y_k ou da iteração corrente, o subproblema admite uma solução única, garantindo que o algoritmo possa sempre computar o próximo passo de forma unívoca.

A seguir, apresentamos a análise de convergência para o caso geral (Algoritmo 4.1), onde a aproximação H_k é construída por um método genérico. Vale ressaltar que para um esquema genérico de atualização de H_k, pode ser difícil garantir uma cota inferior

positiva uniforme para a sequência $\{\sigma_k\}$. Essa dificuldade teórica é uma das principais motivações para o estudo da variante de Hessiana Fixa (Algoritmo 4.2), onde o controle sobre σ_k é mais direto.

Para a análise, utilizaremos o Lema 3.1 (condição de otimalidade) e o resultado a seguir, que estabelece uma cota inferior para o decréscimo da função objetivo em termos da geometria induzida por H.

Lema 4.1 *Suponha que para um dado $v \in \mathbb{R}^n$ e uma matriz definida positiva H, a condição de descida $F(p_H(v)) \leq Q_H(p_H(v), v)$ seja satisfeita. Então, para qualquer $x \in \mathbb{R}^n$, vale a desigualdade*

$$F(x) - F(p_H(v)) \geq \frac{1}{2}\|p_H(v) - v\|_H^2 + \langle v - x, p_H(v) - v\rangle_H.$$

Prova. Como $F(p_H(v)) \leq Q_H(p_H(v), v)$, temos

$$F(x) - F(p_H(v)) \geq F(x) - Q_H(p_H(v), v).$$

Pela convexidade de f e g

$$\begin{aligned} f(x) &\geq f(v) + \langle \nabla f(v), x - v\rangle, \\ g(x) &\geq g(p_H(v)) + \langle \nu_g(p_H(v)), x - p_H(v)\rangle, \end{aligned}$$

onde $\nu_g(p_H(v))$ é o subgradiente dado pelo Lema 3.1. Expandindo a definição de $Q_H(p_H(v), v)$

$$Q_H(p_H(v), v) = f(v) + \langle \nabla f(v), p_H(v) - v\rangle + \frac{1}{2}\|p_H(v) - v\|_H^2 + g(p_H(v)).$$

Substituindo essas relações na desigualdade inicial

$$\begin{aligned} F(x) - F(p_H(v)) &\geq \left[f(v) + \langle \nabla f(v), x - v\rangle + g(p_H(v)) + \langle \nu_g, x - p_H(v)\rangle\right] \\ &\quad - \left[f(v) + \langle \nabla f(v), p_H(v) - v\rangle + \frac{1}{2}\|p_H(v) - v\|_H^2 + g(p_H(v))\right] \\ &= \langle \nabla f(v), x - p_H(v)\rangle + \langle \nu_g, x - p_H(v)\rangle - \frac{1}{2}\|p_H(v) - v\|_H^2 \\ &= \langle \nabla f(v) + \nu_g(p_H(v)), x - p_H(v)\rangle - \frac{1}{2}\|p_H(v) - v\|_H^2. \end{aligned}$$

Pelo Lema 3.1, sabemos que $\nabla f(v) + \nu_g(p_H(v)) = -H(p_H(v) - v) = H(v - p_H(v))$. Substituindo este termo

$$F(x) - F(p_H(v)) \geq \langle H(v - p_H(v)), x - p_H(v)\rangle - \frac{1}{2}\|p_H(v) - v\|_H^2$$

$$= \langle v - p_H(v), x - p_H(v) \rangle_H - \frac{1}{2}\|p_H(v) - v\|_H^2.$$

Para obter a desigualdade desejada, note que $x - p_H(v) = (x - v) + (v - p_H(v))$. Assim

$$\begin{aligned}\langle v - p_H(v), x - p_H(v) \rangle_H &= \langle v - p_H(v), x - v \rangle_H + \langle v - p_H(v), v - p_H(v) \rangle_H \\ &= \langle v - p_H(v), x - v \rangle_H + \|v - p_H(v)\|_H^2.\end{aligned}$$

Substituindo na desigualdade acima, obtemos

$$\begin{aligned}F(x) - F(p_H(v)) &\geq \langle v - p_H(v), x - v \rangle_H + \|v - p_H(v)\|_H^2 - \frac{1}{2}\|p_H(v) - v\|_H^2 \\ &= \langle v - x, p_H(v) - v \rangle_H + \frac{1}{2}\|p_H(v) - v\|_H^2,\end{aligned}$$

o que conclui a prova do lema. □

O próximo resultado relaciona a redução no valor da função objetivo com a distância percorrida pelo passo proximal na norma induzida por H.

Corolário 4.2 *Sejam $v \in \mathbb{R}^n$ e $H \succ 0$ tais que $F(p_H(v)) \leq Q_H(p_H(v), v)$. Então, para qualquer $x \in \mathbb{R}^n$, tem-se*

$$2(F(x) - F(p_H(v))) \geq \|p_H(v) - x\|_H^2 - \|v - x\|_H^2.$$

Prova. Segue do Lema 4.1 que

$$F(x) - F(p_H(v)) \geq \frac{1}{2}\|p_H(v) - v\|_H^2 + \langle v - x, p_H(v) - v \rangle_H,$$

o que é equivalente a

$$2(F(x) - F(p_H(v))) \geq \|p_H(v) - v\|_H^2 + 2\langle p_H(v) - v, v - x \rangle_H.$$

Utilizando a identidade algébrica $\|b - a\|^2 + 2\langle b - a, a - c \rangle = \|b - c\|^2 - \|a - c\|^2$ com $a = v$, $b = p_H(v)$ e $c = x$, obtemos

$$2(F(x) - F(p_H(v))) \geq \|p_H(v) - x\|_H^2 - \|v - x\|_H^2,$$

o que conclui a prova do resultado. □

O próximo Lema discute relações invariantes cruciais que são mantidas durante todo o algoritmo. Essas propriedades são fundamentais para garantir que a aceleração não viole a estabilidade do método.

Lema 4.3 *Em cada iteração do Algoritmo 4.1, as seguintes relações são válidas:*

1. $\sigma_k H_k \succeq \sigma_{k+1} H_{k+1}$;
2. $\sigma_k t_k^2 \geq \sigma_{k+1} t_{k+1}(t_{k+1} - 1)$.

Prova. O primeiro item é garantido explicitamente pela construção do Algoritmo 4.1 (no passo de atualização de H_{k+1}). Resta provar o segundo item. Da definição de t_{k+1} em (4-2), temos

$$2t_{k+1} - 1 = \sqrt{1 + 4\theta_k t_k^2}.$$

Daí,

$$(2t_{k+1} - 1)^2 = 1 + 4\theta_k t_k^2,$$

o que implica que

$$t_{k+1}(t_{k+1} - 1) = \theta_k t_k^2.$$

Como $\theta_k = \sigma_k / \sigma_{k+1}^0$ e o algoritmo assegura que $\sigma_{k+1} \geq \sigma_{k+1}^0$, temos a relação $\theta_k \leq \sigma_k / \sigma_{k+1}$. Portanto

$$t_{k+1}(t_{k+1} - 1) = \theta_k t_k^2 \leq \frac{\sigma_k}{\sigma_{k+1}} t_k^2.$$

Multiplicando por σ_{k+1}, concluímos que $\sigma_{k+1} t_{k+1}(t_{k+1} - 1) \leq \sigma_k t_k^2$. □

O Lema 4.4 estabelece o decréscimo dessa sequência, o que é imprescindível para provar a taxa de convergência $\mathcal{O}(1/k^2)$. A partir de agora, considere $x^* \in X^*$.

Lema 4.4 *Para todo $k \geq 1$, definem-se as sequências*

$$v_k := F(x_k) - F(x^*), \quad u_k := t_k x_k - (t_k - 1)x_{k-1} - x^*.$$

Então, vale a desigualdade

$$2\sigma_k t_k^2 v_k + \|u_k\|_{\sigma_k H_k}^2 \geq 2\sigma_{k+1} t_{k+1}^2 v_{k+1} + \|u_{k+1}\|_{\sigma_{k+1} H_{k+1}}^2.$$

Prova. Aplicamos o Corolário 4.2 com os parâmetros $v = y_{k+1}$, $p_H(v) = x_{k+1}$ e $H = H_{k+1}$ e $x = x_k$

$$2(F(x_k) - F(x_{k+1})) \geq \|x_{k+1} - y_{k+1}\|_{H_{k+1}}^2 + 2\langle x_{k+1} - y_{k+1}, y_{k+1} - x_k \rangle_{H_{k+1}}. \tag{4-6}$$

Note que $F(x_k) - F(x_{k+1}) = (F(x_k) - F(x^*)) - (F(x_{k+1}) - F(x^*)) = v_k - v_{k+1}$. Multiplicando a inequação (4-6) por $\sigma_{k+1}(t_{k+1} - 1)$

$$\begin{aligned} 2\sigma_{k+1}(t_{k+1} - 1)(v_k - v_{k+1}) \geq \sigma_{k+1}(t_{k+1} - 1)\|x_{k+1} - y_{k+1}\|_{H_{k+1}}^2 \\ + 2\sigma_{k+1}(t_{k+1} - 1)\langle x_{k+1} - y_{k+1}, y_{k+1} - x_k \rangle_{H_{k+1}}. \end{aligned} \tag{4-7}$$

Novamente, aplicando o Corolário 4.2 com $v = y_{k+1}$, $p_H(v) = x_{k+1}$ e $H = H_{k+1}$ e $x = x^*$

$$2(F(x^*) - F(x_{k+1})) \geq \|x_{k+1} - y_{k+1}\|^2_{H_{k+1}} + 2\langle x_{k+1} - y_{k+1}, y_{k+1} - x^*\rangle_{H_{k+1}}.$$

Como $F(x^*) - F(x_{k+1}) = -v_{k+1}$, multiplicando por σ_{k+1}, temos

$$-2\sigma_{k+1}v_{k+1} \geq \sigma_{k+1}\|x_{k+1} - y_{k+1}\|^2_{H_{k+1}} + 2\sigma_{k+1}\langle x_{k+1} - y_{k+1}, y_{k+1} - x^*\rangle_{H_{k+1}}. \tag{4-8}$$

Somando (4-7) e (4-8) obtemos

Lado Esquerdo

$$\begin{aligned}\text{LHS} &= 2\sigma_{k+1}(t_{k+1} - 1)(v_k - v_{k+1}) - 2\sigma_{k+1}v_{k+1} \\ &= 2\sigma_{k+1}[(t_{k+1} - 1)v_k - (t_{k+1} - 1 + 1)v_{k+1}] \\ &= 2\sigma_{k+1}[(t_{k+1} - 1)v_k - t_{k+1}v_{k+1}].\end{aligned}$$

Multiplicando por t_{k+1} e usando o Lema 4.3 ($\sigma_{k+1}t_{k+1}(t_{k+1} - 1) \leq \sigma_k t_k^2$)

$$\begin{aligned}t_{k+1} \cdot \text{LHS} &= 2[\sigma_{k+1}t_{k+1}(t_{k+1} - 1)v_k - \sigma_{k+1}t_{k+1}^2 v_{k+1}] \\ &\leq 2\sigma_k t_k^2 v_k - 2\sigma_{k+1}t_{k+1}^2 v_{k+1}.\end{aligned} \tag{4-9}$$

Lado Direito Agrupando os termos quadráticos e os produtos internos multiplicados por t_{k+1}

$$\begin{aligned}t_{k+1} \cdot \text{RHS} &= \sigma_{k+1}t_{k+1}[(t_{k+1} - 1) + 1]\|x_{k+1} - y_{k+1}\|^2_{H_{k+1}} \\ &\quad + 2\sigma_{k+1}t_{k+1}\langle x_{k+1} - y_{k+1}, (t_{k+1} - 1)(y_{k+1} - x_k) + (y_{k+1} - x^*)\rangle_{H_{k+1}} \\ &= \sigma_{k+1}t_{k+1}^2\|x_{k+1} - y_{k+1}\|^2_{H_{k+1}} \\ &\quad + 2\sigma_{k+1}\langle x_{k+1} - y_{k+1}, t_{k+1}y_{k+1} - (t_{k+1} - 1)x_k - x^*\rangle_{t_{k+1}H_{k+1}}.\end{aligned}$$

Agora, aplicamos a identidade $\|b - c\|^2 - \|a - c\|^2 = \|b - a\|^2 + 2\langle b - a, a - c\rangle$ no espaço com métrica ponderada por $\sigma_{k+1}H_{k+1}$. Definimos

$$\begin{aligned}a &:= t_{k+1}y_{k+1}, \\ b &:= t_{k+1}x_{k+1}, \\ c &:= (t_{k+1} - 1)x_k + x^*.\end{aligned}$$

Note que $b - a = t_{k+1}(x_{k+1} - y_{k+1})$. O termo quadrático do RHS é exatamente $\|b - a\|^2_{\sigma_{k+1}H_{k+1}}$. Analisemos os vetores $b - c$ e $a - c$.

$$b - c = t_{k+1}x_{k+1} - (t_{k+1} - 1)x_k - x^* = u_{k+1}.$$

$$a - c = t_{k+1}y_{k+1} - (t_{k+1} - 1)x_k - x^*.$$

Da definição de y_{k+1} em (4-3), temos $t_{k+1}y_{k+1} = t_{k+1}x_k + (t_k - 1)(x_k - x_{k-1})$. Logo

$$\begin{aligned} a - c &= [t_{k+1}x_k + (t_k - 1)(x_k - x_{k-1})] - (t_{k+1} - 1)x_k - x^* \\ &= x_k(t_{k+1} - t_{k+1} + 1) + t_k x_k - x_k - (t_k - 1)x_{k-1} - x^* \\ &= t_k x_k - (t_k - 1)x_{k-1} - x^* = u_k. \end{aligned}$$

Portanto, o lado direito torna-se

$$\|u_{k+1}\|^2_{\sigma_{k+1}H_{k+1}} - \|u_k\|^2_{\sigma_{k+1}H_{k+1}}.$$

Combinando as desigualdades

$$2\sigma_k t_k^2 v_k - 2\sigma_{k+1}t_{k+1}^2 v_{k+1} \geq \|u_{k+1}\|^2_{\sigma_{k+1}H_{k+1}} - \|u_k\|^2_{\sigma_{k+1}H_{k+1}}.$$

Rearranjando os termos

$$2\sigma_k t_k^2 v_k + \|u_k\|^2_{\sigma_{k+1}H_{k+1}} \geq 2\sigma_{k+1}t_{k+1}^2 v_{k+1} + \|u_{k+1}\|^2_{\sigma_{k+1}H_{k+1}}.$$

Finalmente, utilizamos a propriedade do Lema 4.3 que $\sigma_k H_k \succeq \sigma_{k+1}H_{k+1}$. Isso implica que $\|u_k\|^2_{\sigma_k H_k} \geq \|u_k\|^2_{\sigma_{k+1}H_{k+1}}$. Substituindo no lado esquerdo, obtemos o resultado desejado. □

No Teorema 4.5, estabelecemos a cota superior para o erro da função objetivo, demonstrando que ele decai inversamente proporcional ao termo de aceleração $\sigma_k t_k^2$.

Teorema 4.5 *A sequência de iterações $\{x_k\}$ gerada pelo Algoritmo 4.1 satisfaz*

$$F(x_k) - F(x^*) \leq \frac{\|x_0 - x^*\|^2_{\sigma_0 H_0}}{2\sigma_k t_k^2}.$$

Prova. Defina $\phi_k = 2\sigma_k t_k^2 v_k + \|u_k\|^2_{\sigma_k H_k}$, onde $v_k = F(x_k) - F(x^*)$ e $u_k = t_k x_k - (t_k - 1)x_{k-1} - x^*$. Pelo Lema 4.4, sabemos que a sequência $\{\phi_k\}$ é não crescente, ou seja, $\phi_k \leq \phi_{k-1} \leq \cdots \leq \phi_1$. Portanto

$$2\sigma_k t_k^2 v_k + \|u_k\|^2_{\sigma_k H_k} \leq \phi_1 = 2\sigma_1 t_1^2 v_1 + \|u_1\|^2_{\sigma_1 H_1}.$$

Como a matriz $\sigma_k H_k$ é definida positiva, o termo $\|u_k\|^2_{\sigma_k H_k}$ é sempre não negativo. Assim, podemos descartá-lo para obter a desigualdade

$$2\sigma_k t_k^2 v_k \leq 2\sigma_1 t_1^2 v_1 + \|u_1\|^2_{\sigma_1 H_1}. \tag{4-10}$$

Na primeira iteração ($k = 1$), temos por definição que $t_1 = 1$. Substituindo na expressão de u_1, obtemos

$$u_1 = 1 \cdot x_1 - (1-1)x_0 - x^* = x_1 - x^*.$$

Logo, o lado direito da desigualdade (4-10) torna-se

$$\phi_1 = 2\sigma_1 v_1 + \|x_1 - x^*\|^2_{\sigma_1 H_1}.$$

Agora, aplicamos o Corolário 4.2 para a iteração inicial. Tomando $v = y_1 = x_1$, $p_H(v) = x_1$, $H = H_0$ e $x = x^*$, temos

$$2(F(x^*) - F(x_1)) \geq \|x_1 - x^*\|^2_{H_1} - \|x_0 - x^*\|^2_{H_1}.$$

Sabendo que $F(x^*) - F(x_1) = -v_1$ e multiplicando toda a expressão por σ_1

$$\begin{aligned} -2\sigma_1 v_1 &\geq \sigma_1 \|x_1 - x^*\|^2_{H_1} - \sigma_1 \|x_0 - x^*\|^2_{H_1} \\ &= \|x_1 - x^*\|^2_{\sigma_1 H_1} - \|x_0 - x^*\|^2_{\sigma_1 H_1}. \end{aligned}$$

Rearranjando os termos para isolar a parte que aparece em ϕ_1

$$2\sigma_1 v_1 + \|x_1 - x^*\|^2_{\sigma_1 H_1} \leq \|x_0 - x^*\|^2_{\sigma_1 H_1}. \tag{4-11}$$

Substituindo (4-11) em (4-10), concluímos que

$$2\sigma_k t_k^2 v_k \leq \|x_0 - x^*\|^2_{\sigma_1 H_1}.$$

Isolando v_k

$$F(x_k) - F(x^*) \leq \frac{\|x_0 - x^*\|^2_{\sigma_1 H_1}}{2\sigma_k t_k^2},$$

o que, combinado com o fato que $\sigma_1 H_1 \preceq \sigma_0 H_0$, prova a desigualdade desejada. □

O Teorema 4.5 demonstra que o erro decai com a taxa $1/(\sigma_k t_k^2)$. Para garantir a convergência global acelerada de ordem $\mathcal{O}(1/k^2)$, é necessário assegurar que o denominador $\sigma_k t_k^2$ cresça quadraticamente com k. O Lema 4.6 a seguir fornece essa garantia.

Lema 4.6 *A sequência $\{\sigma_k, t_k\}$ gerada pelo Algoritmo 4.1 satisfaz:*

$$\sigma_k t_k^2 \geq \left(\frac{1}{2} \sum_{i=1}^{k} \sqrt{\sigma_i} \right)^2.$$

Prova. A prova segue por indução em k.

Base ($k = 1$) Temos $t_1 = 1$. Logo, $\sqrt{\sigma_1} t_1 = \sqrt{\sigma_1}$. O lado direito da desigualdade é $\sqrt{\sigma_1}/2$.

Como $\sqrt{\sigma_1} \geq \sqrt{\sigma_1}/2$ é trivialmente verdade (pois $\sigma_1 > 0$), a base se verifica.
Passo Indutivo Suponha que a afirmação seja válida para k, isto é, $\sqrt{\sigma_k}t_k \geq \frac{1}{2}\sum_{i=1}^{k}\sqrt{\sigma_i}$. Devemos provar para $k+1$. Da definição de t_{k+1}

$$t_{k+1} = \frac{1}{2} + \sqrt{\frac{1}{4} + \theta_k t_k^2} \geq \frac{1}{2} + \sqrt{\theta_k t_k^2}.$$

Lembrando que no algoritmo $\theta_k = \sigma_k/\sigma_{k+1}^0$ e, após o ajuste, σ_{k+1} é tal que a relação efetiva utilizada na prova do Lema 4.3 garante a coerência. Usando a cota conservadora $\theta_k \approx \sigma_k/\sigma_{k+1}$ (ou a relação algébrica direta $t_{k+1} \approx (t_k\sqrt{\sigma_k})/\sqrt{\sigma_{k+1}}$ para grandes k)

$$t_{k+1} \geq \frac{1}{2} + \sqrt{\frac{\sigma_k}{\sigma_{k+1}}}\, t_k.$$

Multiplicando por $\sqrt{\sigma_{k+1}}$,

$$\sqrt{\sigma_{k+1}}t_{k+1} \geq \frac{\sqrt{\sigma_{k+1}}}{2} + \sqrt{\sigma_k}t_k.$$

Aplicando a hipótese de indução no termo $\sqrt{\sigma_k}t_k$,

$$\sqrt{\sigma_{k+1}}t_{k+1} \geq \frac{\sqrt{\sigma_{k+1}}}{2} + \frac{1}{2}\sum_{i=1}^{k}\sqrt{\sigma_i} = \frac{1}{2}\sum_{i=1}^{k+1}\sqrt{\sigma_i}.$$

Elevando ambos os lados ao quadrado, obtemos o resultado desejado. □

Assumindo que a sequência $\{\sigma_k\}$ é limitada inferiormente por uma constante positiva $\underline{\sigma}$, podemos estabelecer a limitação desejada para o termo de aceleração $\sigma_k t_k^2$, garantindo a taxa de convergência ótima.

Teorema 4.7 *Suponha que, para todo $k \geq 1$, tenhamos $\sigma_k \geq \underline{\sigma}$ (onde $\underline{\sigma}$ é uma constante positiva). Então,*

$$F(x_k) - F(x^*) \leq \frac{2\|x_0 - x^*\|^2_{\sigma_0 H_0}}{\underline{\sigma}k^2}, \quad \forall k \geq 1.$$

Prova. Partimos da hipótese que $\sigma_i \geq \underline{\sigma}$ para todo i. Utilizando o Lema 4.6, temos

$$\begin{aligned}\sigma_k t_k^2 &\geq \left(\frac{\sum_{i=1}^{k}\sqrt{\sigma_i}}{2}\right)^2 \\ &\geq \left(\frac{\sum_{i=1}^{k}\sqrt{\underline{\sigma}}}{2}\right)^2 = \left(\frac{k\sqrt{\underline{\sigma}}}{2}\right)^2 = \frac{k^2\underline{\sigma}}{4}.\end{aligned} \tag{4-12}$$

Invertendo a desigualdade, obtemos

$$\frac{1}{\sigma_k t_k^2} \leq \frac{4}{k^2 \underline{\sigma}}. \tag{4-13}$$

Pelo Teorema 4.5, sabemos que $F(x_k) - F(x^*) \leq (\|x_0 - x^*\|^2_{\sigma_0 H_0})/(2\sigma_k t_k^2)$. Substituindo o limite obtido em (4-13)

$$\begin{aligned} F(x_k) - F(x^*) &\leq \frac{\|x_0 - x^*\|^2_{\sigma_0 H_0}}{2} \cdot \frac{4}{k^2 \underline{\sigma}} \\ &= \frac{2\|x_0 - x^*\|^2_{\sigma_0 H_0}}{k^2 \underline{\sigma}}, \end{aligned}$$

o que conclui a prova. □

É importante notar que a hipótese de existência de um limitante inferior uniforme $\underline{\sigma} > 0$ (juntamente com as condições do Lema 4.3) pode não ser satisfeita se utilizarmos uma aproximação genérica para a Hessiana. Nesse cenário, o Teorema 4.7 não se aplica e a convergência acelerada $\mathcal{O}(1/k^2)$ não é garantida. Para ilustrar isso, considere as seguintes sequências de matrizes

$$H_{2k} = \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix} \text{ e } H_{2k+1} = \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix}.$$

Claramente, $\sigma_{2k+1} \leq \sigma_{2k}/10$ e $\sigma_{2k} \leq \sigma_{2k-1}/10$, e $\sigma_k \leq 10^{-k}$. Nesse caso, baseado no resultado do Teorema 4.5, não podemos garantir nenhum resultado de convergência.

Entretanto, é possível demonstrar que o algoritmo preserva a convergência global (embora com taxa reduzida) mesmo quando $\sigma_k \to 0$, desde que esse decaimento não seja excessivamente rápido. O Teorema 4.8 a seguir estabelece que, se σ_k decair a uma taxa não superior a $\mathcal{O}(1/k)$, o algoritmo converge com taxa sublinear $\mathcal{O}(1/k)$, recuperando o desempenho do método não acelerado.

Teorema 4.8 *Se para todas as iterações do Algoritmo 4.1 tivermos $\sigma_k \geq \underline{\sigma}/k$ para alguma constante $\underline{\sigma} > 0$, então, para todo $k \geq 1$:*

$$F(x_k) - F(x^*) \leq \frac{2\|x_0 - x^*\|^2_{\sigma_0 H_0}}{\underline{\sigma} k}.$$

Prova. Assumimos que $\sigma_i \geq \underline{\sigma}/i$ para todo $i = 1, \ldots, k$. Retomando a cota inferior do Lema 4.6

$$\sum_{i=1}^{k} \sqrt{\sigma_i} \geq \sum_{i=1}^{k} \sqrt{\frac{\underline{\sigma}}{i}} = \sqrt{\underline{\sigma}} \sum_{i=1}^{k} \frac{1}{\sqrt{i}}.$$

Para limitar a soma harmônica generalizada, observamos que a função $1/\sqrt{i}$ é decrescente. Logo, o menor termo da soma é o último $(1/\sqrt{k})$. Portanto

$$\sum_{i=1}^{k} \frac{1}{\sqrt{i}} \geq k \cdot \frac{1}{\sqrt{k}} = \sqrt{k}.$$

Substituindo essa estimativa na desigualdade do Lema 4.6

$$\begin{aligned} \sigma_k t_k^2 &\geq \left(\frac{1}{2} \sum_{i=1}^{k} \sqrt{\sigma_i} \right)^2 \\ &\geq \left(\frac{\sqrt{\underline{\sigma}}\sqrt{k}}{2} \right)^2 = \frac{k\underline{\sigma}}{4}. \end{aligned}$$

Invertendo a desigualdade, temos

$$\frac{1}{\sigma_k t_k^2} \leq \frac{4}{k\underline{\sigma}}.$$

Finalmente, aplicando o Teorema 4.5

$$\begin{aligned} F(x_k) - F(x^*) &\leq \frac{\|x_0 - x^*\|_{\sigma_0 H_0}^2}{2\sigma_k t_k^2} \\ &\leq \frac{\|x_0 - x^*\|_{\sigma_0 H_0}^2}{2} \cdot \frac{4}{k\underline{\sigma}} \\ &= \frac{2\|x_0 - x^*\|_{\sigma_0 H_0}^2}{k\underline{\sigma}}, \end{aligned}$$

o que prova a desigualdade desejada. □

Finalmente, podemos estabelecer a taxa de convergência acelerada para o Algoritmo 4.2. Neste caso específico, a estrutura de Hessiana Fixa nos permite garantir uma cota inferior global para σ_k, o que satisfaz a hipótese exigida pelo Teorema 4.7.

Lema 4.9 *No Algoritmo 4.2, assumindo que $mI \preceq H$, temos que $\sigma_k \geq (\beta m)/L$ para todo k. Consequentemente, a taxa de convergência ótima de $\mathcal{O}(1/k^2)$ é alcançada.*

Prova. O Algoritmo 4.2 define a matriz de aproximação como $H_k = H/\sigma_k$. Sabemos, pela propriedade de Lipschitz do gradiente de f, que a condição de decréscimo suficiente $F(p_{H_k}(y)) \leq Q_{H_k}(p_{H_k}(y), y)$ é sempre satisfeita se a matriz do modelo dominar a curvatura da função, isto é, se $H_k \succeq LI$. Substituindo a definição de H_k e utilizando a hipótese $H \succeq mI$

$$H_k = \frac{1}{\sigma_k} H \succeq \frac{m}{\sigma_k} I.$$

Portanto, uma condição suficiente para garantir $H_k \succeq LI$ é

$$\frac{m}{\sigma_k} \geq L \iff \sigma_k \leq \frac{m}{L}.$$

O algoritmo inicia com uma estimativa σ e, enquanto a condição de decréscimo não for satisfeita, reduz o parâmetro multiplicando-o por β (ou seja, $\sigma_{novo} = \beta\sigma_{atual}$). Como a condição de decréscimo é garantidamente satisfeita sempre que $\sigma \leq m/L$, o loop de backtracking interromperá a redução assim que σ entrar nessa região segura. Isso implica que o algoritmo jamais reduzirá σ_k para um valor arbitrariamente pequeno se não for necessário. Matematicamente, se o loop para em um valor σ_k, significa que o valor anterior (que falhou na verificação) era maior que m/L. No pior cenário, o valor imediatamente anterior era arbitrariamente próximo de m/L, e ao multiplicarmos por β, obtemos o menor valor possível aceito

$$\sigma_k \geq \frac{\beta m}{L}.$$

Definindo $\underline{\sigma} = (\beta m)/L$, temos uma cota inferior positiva uniforme para a sequência. Como o Algoritmo 4.2 é um caso particular do Algoritmo 4.1 que satisfaz $\sigma_k \geq \underline{\sigma}$, aplicamos o Teorema 4.7 para concluir que a taxa de convergência é $\mathcal{O}(1/k^2)$. □

A análise desenvolvida nesta seção considerou que os subproblemas proximais são resolvidos de forma exata. Contudo, em aplicações práticas, pode ser necessário utilizar métodos iterativos para resolver esses subproblemas. Conforme discutido em [3], a análise pode ser estendida para o caso inexato de maneira direta. Os autores afirmam que, se um algoritmo com convergência linear for utilizado para resolver os subproblemas (respeitando critérios adequados de erro), a taxa de convergência global $\mathcal{O}(1/k^2)$ do método Quasi-Newton Proximal Acelerado é preservada.

CAPÍTULO 5

Conclusão

Neste trabalho, realizamos a análise de dois métodos para a resolução de problemas de otimização composta com base nas referências [3] e [17]. Para o primeiro, algoritmo quasi-Newton Proximal (AQNP), foi mostrado que este possui uma taxa de convergência sublinear da $\mathcal{O}(1/k)$, enquanto para o segundo, uma versão acelerada do primeiro denotado por AQNPA, foi mostrada uma taxa de convergência melhor da $\mathcal{O}(1/k^2)$. Um passo futuro seria checar a performance prática desses algoritmos em diferentes aplicações para verificar se o método com melhor taxa de convergência é de fato o mais eficiente na prática.

Referências Bibliográficas

[1] BECK, A.; TEBOULLE, M. **A fast iterative shrinkage-thresholding algorithm for linear inverse problems**. *SIAM*, 2:183–202, 2009.

[2] BYRD, R.; NOCEDAL, J.; OZTOPRAK, F. **An inexact successive quadratic approximation method for convex l_1-regularized optimization**. *Math. Program.*, 157:375–396, 2016.

[3] GHANBARI, H.; SCHEINBERG, K. **Proximal quasi-newton methods for regularized convex optimization with linear and accelerated sublinear convergence rates**. *Computational Optimization and Applications*, 69:597–627, 2018.

[4] HSIEH, C.; SUSTIK, M.; DHILON, I.; RAVIKUMAR, P. **Sparse inverse covariance matrix estimation using quadratic approximation**. In: *NIPS*, p. 2330–2338, 2011.

[5] IZMAILOV, A.; SOLODOV, M. **Otimização - volume 1: Condições de Otimalidade, Elementos de Análise Convexa e de Dualidade**. IMPA, Rio de Janeiro, 3 edition, 2014.

[6] JIANG, K.; SUN, D.; TOH, K. **An inexact accelerated proximal gradient method for large scale linearly constrained convex SDP**. *SIAM*, 22:1042–1064, 2012.

[7] LEE, J.; SUN, Y.; SAUNDERS, M. **Proximal newton-type methods for convex optimization for minimizing composite functions**. *SIAM J. Optim.*, 24:1420–1443, 2014.

[8] NEMIROVSKI, A.; YUDIN, D. **Informational Complexity and Efficient Methods for Solution of Convex Extremal Problems**. Wiley, New York, 1983.

[9] NESTEROV, Y. **A method for solving the convex programming problem with convergence rate $\mathcal{O}(1/k^2)$**. *Sov. Math. Dokl.*, 27:372–376, 1983.

[10] NESTEROV, Y. **Introductory Lectures on Convex Programming: A Basic Course**. Springer, New York, 2004.

[11] NESTEROV, Y. **Gradient methods for minimizing composite objective function**. *Math. Program.*, 140:125–161, 2013.

[12] NOCEDAL, J.; WRIGHT, S. **Numerical Optimization**. Springer Series in Operations Research. Springer, New York, 2nd edition, 2006.

[13] OLSEN, P.; OZTOPRAK, F.; NOCEDAL, J.; RENNIE, S. **Newton-like methods for sparse inverse covariance estimation**. In: *NIPS*, p. 764–772, 2012.

[14] RIBEIRO, A. A.; KARAS, E. W. **Otimização contínua: aspectos teóricos e computacionais**. Cengage Learning, São Paulo, 2013.

[15] SCHEINBERG, K.; GOLDFARB, D.; BAI, X. **Fast first-order methods for composite convex optimization with backtracking**. *Found. Math.*, 14:389–417, 2014.

[16] SCHEINBERG, K.; RISH, I. **A greedy coordinate ascent method for sparse inverse covariance selection problem**. Technical report, SINCO, Technical Report, 2009.

[17] SCHEINBERG, K.; TANG, X. **Practical inexact proximal quasi-newton method with global complexity analysis**. *Math. Program.*, 160:495–529, 2016.

[18] SHALEV-SHWARTZ, S.; TEWARI, A. **Stochastic methods for l_1-regularized loss minimization**. In: *ICML*, p. 929–936, 2009.

[19] SILVA, S. R. P. D. **Algoritmo de ponto proximal para otimização em $\mathbb{R}^n$**. Dissertação de mestrado, Universidade Federal de Goiás, Goiânia, 1999. Dissertação de Mestrado em Matemática, Instituto de Matemática e Estatística.

[20] SRA, S.; NOWOZIN, S.; WRIGHT, S. **Optimization for Machine Learning**. MIT Press, Cambridge, 2011.

[21] TIBSHIRANI, R. **Regression shrinkage and selection via the lasso**. *J. R. Stat. Soc. Ser. B Methodol.*, 58:267–288, 1996.

[22] YUAN, G.; CHANG, K.; HSIEH, C.; LIN, C. **A comparison of optimization methods and software for large-scale l_1-regularized linear classification**. *JMLR*, 11:3183–3234, 2010.