Universidade Federal de Minas Gerais
Escola de Engenharia
Departamento de Engenharia Elétrica

Notas de Aula de Otimização

Jaime A. Ramírez
Felipe Campelo
Frederico G. Guimarães
Lucas S. Batista
Ricardo H. C. Takahashi

Sumário

Lista de Figuras

Capítulo 1

Introdução

Neste capítulo, iremos discutir, de maneira preliminar, *o que são* os problemas de *Otimização*, com base sempre em funções matemáticas simples, de apenas duas variáveis, que permitem portanto sua representação gráfica em três dimensões. Iremos mostrar como diferentes tipos de funções irão requerer diferentes estratégias de otimização e, de maneira intuitiva, iremos discutir os princípios que se encontram por trás dos métodos de otimização que serão estudados em detalhe nos próximos capítulos. Leitura complementar pode ser encontrada em [1] - [2].

1.1 O Jogo da Otimização

A *Otimização*, sob o ponto de vista prático, trata do conjunto de métodos capazes de determinar as melhores configurações possíveis para a construção ou o funcionamento de sistemas de interesse para o ser humano. Estamos falando da aplicação de uma mesma teoria, com um mesmo conjunto de métodos e ferramentas, quando:

- um engenheiro eletricista procura o melhor projeto possível para um motor elétrico;
- um engenheiro de controle e automação procura o melhor ajuste possível para os controles de um determinado processo industrial;
- um engenheiro de produção busca a melhor configuração possível para encadear as etapas de fabricação de um produto;
- um matemático computacional estuda modelos quantitativos de epidemias, procurando determinar as melhores políticas de vacinação;
- um cientista da computação estuda o desempenho de uma rede de computadores, e tenta estabelecer a melhor estratégia de tráfego de informação possível, visando maximizar o fluxo global de informação nessa rede;
- um economista procura o melhor *portfolio* de investimentos, que maximiza a expectativa de retorno financeiro;
- um veterinário ou zootecnista procura determinar a melhor política de compras e vendas das cabeças de um rebanho de gado.

Apesar dos contextos completamente distintos, todos estes problemas (e muitos outros) uma vez formulados matematicamente, possuem exatamente a mesma estrutura, e sua solução é obtida essencialmente através da utilização do mesmo conjunto de técnicas: a *Otimização*.

1.1.1 Formulação do Problema de Otimização

Evidentemente, em cada contexto distinto, há um conjunto de informações que cada especialista de cada área deve conhecer, que lhe permite obter uma descrição matemática de cada problema, a partir da situação concreta em questão. Uma vez construído o modelo do problema[1], chegamos sempre[2] à formulação característica do problema de otimização:

$$\begin{aligned} \boldsymbol{x}^* = & \arg \min_{\boldsymbol{x}} f(\boldsymbol{x}) \\ & \text{sujeito a: } \begin{cases} g_i(\boldsymbol{x}) \leq 0, i = 1, \ldots, p \\ h_j(\boldsymbol{x}) = 0, j = 1, \ldots, q \end{cases} \end{aligned} \tag{1.1}$$

Vamos primeiro entender o que significa essa expressão. Como convenção que adotaremos ao longo de todo este livro, as variáveis em negrito significam grandezas vetoriais (ou seja, que representam conjuntos de vários valores) enquanto as variáveis sem negrito significam grandezas escalares (que representam um único valor).

O projeto de um auto-falante

Por exemplo, suponhamos que um engenheiro está projetando um auto-falante como indicado nas Figs. 1.1 e 1.2. Utilizaremos este exemplo para definir conceitualmente o significado da equação (1.1).

O objetivo é encontrar o auto-falante com menor volume (e possivelmente menor preço) que satisfaça algumas características de desempenho e de construção; por exemplo: (i) densidade de fluxo magnético no entreferro (variável x_9) maior do que um valor pré-determinado, (ii) materiais que compõem as regiões do "ferro" e "ímã" tais que seja possível obter o desempenho especificado e o menor volume. Essas caracteríticas são definidas, usualmente, por quem contrata o projeto.

Matematicamente, o problema do auto-falante pode ser definido por:

$$\begin{aligned} & \min f(\boldsymbol{x}) = \text{volume} \\ & \text{sujeito a: } g_1(\boldsymbol{x}) : |\mathbf{B}| \geq \mathbf{B}_{min} \end{aligned} \tag{1.2}$$

onde $\mathbf{B}$ significa a densidade de fluxo magnético. Este exemplo será discutido detalhadamente ao final do capítulo.

[1] O leitor não deve se enganar: a construção do modelo matemático do problema muitas vezes é a parte mais difícil de todo o processo. Estamos saltando esta parte porque a Otimização começa exatamente quando o modelo da situação está pronto.

[2] OK, você está certo: *quase* sempre.

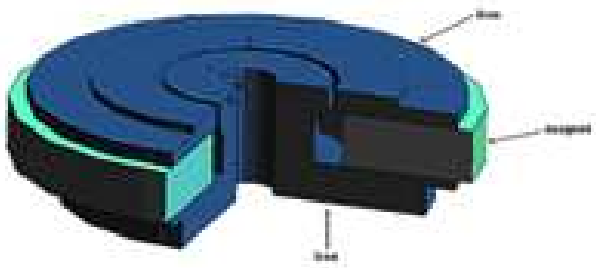

Figura 1.1: Ilustração de um alto-falante constituído por uma estrutura composta por “ferro” (parte azul escuro) e “ímã” (parte azul claro).

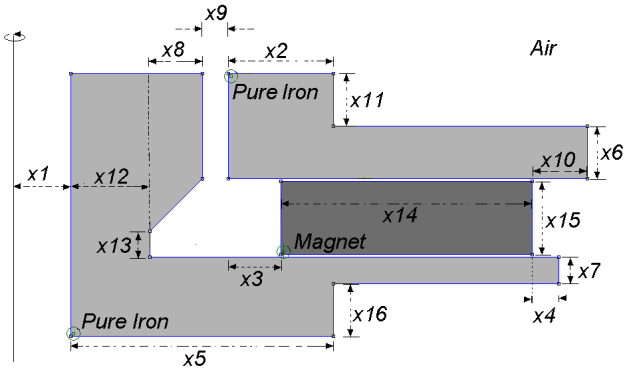

Figura 1.2: Ilustração do alto-falante em 2D com indicação das regiões de “ferro”, “ímã” e das variáveis $\boldsymbol{x}$ de projeto.

O vetor de variáveis de otimização

O vetor $\boldsymbol{x}$ é o *vetor de variáveis de otimização*, que representa o conjunto das variáveis cujos valores procuramos especificar através do processo de otimização.

No exemplo do projeto do auto-falante, o objetivo é encontrar os valores (dimensões) das dezesseis variáveis $\boldsymbol{x}$, que seria neste caso representado por:

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{16} \end{bmatrix} \tag{1.3}$$

Uma vez especificados esses dezesseis valores, para construir o auto-falante basta "seguir a receita" implícita em $\boldsymbol{x}$: obter as peças de "ferro" e "ímã" com as dimensões especificadas.

Nós optamos pelo exemplo do auto-falante, pois é muito útil para ilustrar o fato de que os elementos do vetor $\boldsymbol{x}$ possuem usualmente um significado bastante concreto, ligado à estrutura do problema que está sendo representado. De maneira genérica, se o vetor $\boldsymbol{x}$ possui n variáveis reais, dizemos que $\boldsymbol{x} \in \mathbb{R}^n$.

Nem sempre o vetor de variáveis de otimização é composto de variáveis reais. Muitas vezes, as variáveis são números inteiros, por exemplo, quando estamos estabelecendo quantas máquinas serão utilizadas para trabalhar em determinada etapa de um processo de fabricação. Outras vezes as variáveis são até mesmo binárias: por exemplo, ao se estudar o problema da formação de uma malha viária ligando diversas cidades, deve-se decidir se determinada estrada ligando diretamente duas cidades será ou não construída (só existiriam, nesse caso, as opções *sim* ou *não*).

A diferença mais importante entre os problemas de otimização, que conduz a técnicas de resolução com fundamentações bastante distintas, é aquela que separa os problemas em que as variáveis de otimização são reais dos problemas que apresentam variáveis de otimização discretas (binárias ou inteiras). Neste livro, iremos estudar apenas os problemas com variáveis reais.

A função objetivo

A próxima entidade presente na expressão (1.1) que devemos discutir é a chamada *função objetivo*, $f(\cdot)$. Essa entidade representa o índice de desempenho do sistema, cujo valor, por convenção, queremos minimizar para atingirmos o desempenho ótimo.

Um índice que muito frequentemente desejamos minimizar é o *custo* de fabricação de um equipamento. No exemplo em questão, o volume do alto-falante está associado ao custo; ou seja quanto menor o volume do alto-falante menor também será a quantidade de material utilizado (ímãs permanentes são caros), e consequentemente menor será o custo final do equipamento. Por essa razão, nesse exemplo, $f(\cdot) = \mathit{volume}$. As especificações possíveis do *volume* do alto-falante estão contidas no vetor $\boldsymbol{x}$, ou seja, para cada conjunto de diferentes valores que esse vetor assumir haverá um custo de fabricação diferente envolvido.

Um outro exemplo que poderia ser imaginado consiste na fabricação de um motor: de cada maneira diferente que o mesmo for projetado, terá custos de fabricação

diferentes. Nesse caso, a *função objetivo* $f(\boldsymbol{x})$, será uma função que, para cada conjunto de valores que estiver especificado no vetor $\boldsymbol{x}$, irá fornecer o custo de fabricação do equipamento descrito por esse vetor.

Devido a essa interpretação de custo financeiro, muitas vezes a *função objetivo* é chamada, dentro de livros de otimização, de *função custo*.

Outros índices de desempenho de sistemas que muitas vezes queremos minimizar são: consumo de combustível (em automóveis, por exemplo), ruído de funcionamento (em motores), probabilidade de defeitos (em todo tipo de equipamento), etc. Todos eles, claramente, dependem de como o equipamento foi construído, ou seja, são funções do vetor $\boldsymbol{x}$.

Muitas vezes, entretanto, desejamos *maximizar* e não *minimizar* algum índice de desempenho de um sistema. Queremos, por exemplo, maximizar a expectativa de lucro em um *portfolio* de investimentos, assim como o tempo de vida útil de um equipamento, ou a capacidade de produção de uma fábrica. Para simplificar a tarefa de elaborar a teoria matemática da Otimização, *iremos manter a convenção de sempre formular um problema de otimização como um problema de minimização.* Nos casos em que deseja-se fazer uma maximização, devido ao significado do índice de desempenho escolhido, basta minimizarmos a função que se deseja maximizar multiplicada por -1. Ou seja, se se deseja maximizar a função $p(\boldsymbol{x})$, basta fazer $f(\boldsymbol{x}) = -p(\boldsymbol{x})$, de forma que ao determinarmos o vetor $\boldsymbol{x}$ que minimiza $f(\cdot)$, este será também, por consequência, o vetor que maximiza $p(\cdot)$.

Em linguagem matemática, dizemos que $f(\cdot) : \mathbb{R}^n \mapsto \mathbb{R}$. Isso significa que $f(\cdot)$ é uma função de um vetor de n variáveis reais (pertencente ao *espaço* $\mathbb{R}^n$), e a própria função $f(\cdot)$ retorna um valor que é real. As diferentes características que essa função pode ter, assim como as consequências disso para a elaboração de estratégias de otimização são os temas das próximas seções deste capítulo.

A solução ótima

Da maneira como delimitamos o problema exemplo, supondo que o vetor de variáveis de otimização $\boldsymbol{x}$ seja composto de variáveis reais, existem infinitas maneiras diferentes de especificar o alto-falante a ser construído. Em outras palavras, há um número infinito de valores que as variáveis $x_1, \ldots, x_{16}$ podem assumir, o que resulta em um conjunto infinito de possibilidades de construção do alto-falante.

Diante disso, qual é a melhor especificação possível, $\boldsymbol{x}^*$, que o auto-falante pode assumir; ou seja, qual é a especificação que faz com que ele tenha o menor volume e satisfaça a condição $\mathbf{B} > \mathbf{B}_{min}$?

A resposta a tal pergunta é exatamente aquilo que a *Otimização* procura encontrar, por meio de suas técnicas. Em palavras:

> *O vetor ótimo $\boldsymbol{x}^*$ é igual ao* **argumento** *da função $f(\cdot)$ que faz com que essa função atinja seu* **mínimo** *valor.*

Essa é a forma como deve ser lida a primeira linha da expressão (1.1). Posto isso, como encontrar esse vetor $\boldsymbol{x}^*$? Esse é o assunto deste livro.

As restrições

Para terminarmos de entender a formulação contida na expressão (1.1), ainda falta entendermos o significado da desigualdade e da igualdade a que está *sujeito* o resultado da otimização. Essas são as chamadas *restrições* do problema. Elas significam o conjunto dos requisitos que o resultado do projeto deve atender para ser admissível enquanto solução.

O exemplo em questão possui uma restrição de desigualdade $g_1(\boldsymbol{x})$ que especifica o valor mínimo da densidade de fluxo magnético a ser observado na região definida pela variável x_9 (entre ferro) para o qual o alto-falante tem desempenho satisfatório.

Outros tipos de restrição têm significado bastante óbvio: no exemplo do alto-falante seria natural impor também que todas as variáveis sejam positivas, ou $x_1, \ldots, x_{16} > 0$. Embora, se substituído na expressão da função objetivo, um valor negativo de uma variável x talvez possa levar a um "melhor valor" para essa função, não é possível no mundo real construir alto-falantes que tenham dimensões negativas.

Outros tipos de restrição, embora não estejam relacionados com a impossibilidade de implementarmos a solução encontrada, igualmente dizem que tal solução não é admissível, se violar a restrição. Um exemplo disso encontra-se no projeto de automóveis: se queremos projetar o veículo de mínimo custo, não podemos entretanto construir um que cause emissão de gases poluentes acima dos limites estabelecidos em lei. Todos os veículos que emitirem poluentes acima de tais limites não serão considerados soluções admissíveis, por mais barata que seja sua construção. O problema de otimização, colocado dessa forma, passa a ser o de encontrar o projeto do veículo mais barato possível *dentre todos os que atenderem à restrição da emissão de poluentes ser menor ou igual ao limite admissível*.

Os dois exemplos anteriormente citados se enquadram na situação da *restrição de desigualdade*, isto é, são representáveis pela expressão:

$$g_i(\boldsymbol{x}) \leq 0, i = 1, \ldots, p \tag{1.4}$$

Em relação à convenção de que as funções de restrição devam ser menores ou iguais a zero, cabem comentários similares àqueles apresentados a respeito da convenção de estarmos minimizando, sempre, a função objetivo. Para as restrições de desigualdade, caso ocorram situações em que se deseja garantir que certa função seja maior que ou igual a zero, basta garantir que essa função multiplicada por -1 seja menor que ou igual a zero. Caso seja necessário ainda que certa função seja menor ou igual a um número diferente de zero, basta fazer com que essa função menos esse número seja menor que ou igual a zero. Dessa forma, ao construirmos as técnicas de otimização, levaremos sempre em consideração o formato convencionado da desigualdade, assim simplificando a teoria.

Deve-se observar que agora a função $g_i(\cdot)$ é, ela própria, vetorial, retornando múltiplos valores, o que quer dizer que na realidade essa expressão sintética, vetorial, contém um conjunto de expressões escalares, cada uma das quais representa uma restrição diferente. Matematicamente, dizemos que $g_i(\cdot) : \mathbb{R}^n \mapsto \mathbb{R}^p$, o que significa que para cada vetor de variáveis de otimização $\boldsymbol{x} \in \mathbb{R}^n$ que for utilizado como argumento da função $g_i(\cdot)$, esta retorna um conjunto de p valores reais como

resultado, ou seja, a expressão (1.4) é o mesmo que:

$$\begin{aligned} g_1(\boldsymbol{x}) &\leq 0 \\ g_2(\boldsymbol{x}) &\leq 0 \\ &\vdots \\ g_p(\boldsymbol{x}) &\leq 0 \end{aligned} \tag{1.5}$$

sendo cada uma das p funções $g_i(\cdot)$ uma função escalar, que retorna um único valor real. Em problemas práticos, usualmente será necessário lidar com diversas restrições simultaneamente. No exemplo do projeto do automóvel, além de atender ao limite legal de emissão de poluentes, provavelmente será necessária também a preocupação com o consumo de combustível (que não pode ultrapassar um máximo aceitável), com a potência do motor (que não deve ser menor que um mínimo aceitável), etc. O veículo a ser projetado não pode violar nenhuma dessas restrições para ser considerado uma solução aceitável.

Resta ainda falar das *restrições de igualdade*, descritas pela expressão:

$$h_j(\boldsymbol{x}) = 0, j = 1, \ldots, q \tag{1.6}$$

Esse tipo de restrição ocorre quando é necessário que certas variáveis assumam precisamente certos valores. Por exemplo, se estamos projetando uma peça que deve se encaixar precisamente num certo espaço disponível num equipamento, do qual a peça faz parte, queremos que a peça tenha exatamente o tamanho especificado, nem mais nem menos. A peça pode até ser constituída de diversos sub-componentes, cujos tamanhos poderemos escolher, desde que a soma de todos os tamanhos tenha o tamanho total especificado. Também essa expressão é vetorial: $h_j(\cdot) : \mathbb{R}^n \mapsto \mathbb{R}^q$, ou seja, a função vetorial representa na realidade q diferentes equações.

Para concluir este tópico, definimos a seguinte nomenclatura, relacionada com as restrições:

Região factível: Conjunto dos pontos do espaço $\mathbb{R}^n$ que satisfazem, simultaneamente, a todas as restrições (tanto de desigualdade quanto de igualdade). Às vezes a *região factível* é chamada de *conjunto factível*, ou de *conjunto viável*.

Região infactível: Conjunto dos pontos do espaço $\mathbb{R}^n$ que deixam de satisfazer (ou seja, *violam*) pelo menos uma das restrições do problema.

Ponto factível: Ponto pertencente à *região factível*.

Ponto infactível: Ponto pertencente à *região infactível*.

Restrição violada: Cada uma das componentes do vetor $g_i(\boldsymbol{x})$ que apresentar valor positivo, ou cada uma das componentes do vetor $h_j(\boldsymbol{x})$ que apresentar valor não-nulo será chamada de *restrição violada* no ponto $\boldsymbol{x}$.

1.1.2 As Regras do Jogo

O problema da *Otimização* fica em parte definido pela expressão (1.1). Para delinear o que vem a ser o campo de conhecimento da *Otimização Não-Linear*, enunciamos agora um conjunto de *regras* que dizem *como* é abordado esse problema: qual é a informação de que podemos fazer uso durante o processo de otimização, e qual é o custo dessa informação. Iremos supor, ao longo deste livro, que:

Regras de Acesso à Informação

- Não conhecemos expressões matemáticas explícitas que representem a função objetivo $f(\cdot)$ e as funções de restrição $g_i(\cdot)$ e $h_j(\cdot)$.
- Temos entretanto a possibilidade de descobrir quanto valem as funções objetivo e de restrição em qualquer ponto do espaço de variáveis de otimização. Essa é a única informação que conseguiremos adquirir, ao longo do processo de otimização, para nos guiar em direção à solução desejada.

O leitor poderia perguntar: por quê introduzimos essa premissa aparentemente arbitrária? O que impede que tenhamos em mãos um modelo matemático de um sistema qualquer, formulado em termos de expressões matemáticas explícitas, que seriam nossas funções objetivo e de restrições? Bem, nada impede isso, pelo contrário, muitas vezes é isso que ocorre. Entretanto, nessas situações, quando temos expressões explícitas simples representando o sistema, podemos fazer (e usualmente fazemos) uso de técnicas da chamada *Análise Matemática* para determinar o mínimo da função objetivo, empregando ferramentas que não estão no escopo daquilo que usualmente chamamos *Otimização*. Um procedimento simples que frequentemente empregamos nesses casos, por exemplo, é o de derivar a função objetivo, e determinar os pontos em que o gradiente se anula. Quando é possível fazer isso, os pontos de mínimo da função são determinados de maneira direta e exata.

Há entretanto situações em que a utilização desse tipo de procedimento é muito difícil, e em muitos casos impossível.

Voltemos ao exemplo do auto-falante. Não é possível descrever ou calcular $\mathbf{B}$ no entreferro (variável x_9) por meio de expressões simples, envolvendo por exemplo funções trigonométricas ou polinomiais. O cálculo de $\mathbf{B}$ envolve normalmente um sistema de equações diferenciais parciais, cuja solução é provavelmente muito difícil, ou mesmo impossível, de ser determinada analiticamente.

Nesse exemplo, seria necessário escrever um algoritmo para efetuar o cálculo numérico da solução desse sistema de equações. Cada vez que fizéssemos a avaliação da função de restrição $g_1(\cdot)$ para um determinado vetor de variáveis de otimização (que significa um determinado auto-falante), teríamos de executar o algoritmo e, com base no resultado do mesmo, fazer o cálculo da função $g_1(\cdot)$. O mesmo raciocínio se aplicaria a função objetivo, se tivermos uma grandeza em $f(\cdot)$ cujo cálculo envolva a resolução de um sistema de equações diferenciais parciais.

Ora, uma função que inclui um algoritmo não pode ser, em geral, explicitamente representada por uma expressão matemática simples, nem pode ser por exemplo derivada ou integrada de maneira explícita. A natureza da função objetivo, ou das funções de restrição, agora deixa de ser a de uma expressão conhecida, que podemos manipular utilizando todas as manipulações matemáticas usuais.

A metáfora mais adequada para compreendermos sua natureza é a de uma *caixa preta*[3], na qual podemos entrar com um vetor $\boldsymbol{x}$, obtendo como resposta o valor de $f(\boldsymbol{x})$ associado a esse vetor[4]. Essa é a única informação disponível para ser utilizada pelos métodos de *Otimização*.

No exemplo do auto-falante, o cálculo de $\mathbf{B}$ é obtido a partir de um programa de cálculo de campo magnético em que se passa como entrada um vetor $\boldsymbol{x}$ e se obtém como resposta o valor para $g_1(\cdot)$ associado a esse valor. Isso será discutido no final do capítulo.

Assim, as regras acima enunciadas simplesmente significam que a teoria da *Otimização* é desenvolvida para o contexto dos problemas em que não temos acesso a uma expressão explícita da função objetivo e das funções de restrição. Obviamente, nos casos de problemas em que conhecemos expressões explícitas de todas as funções, as técnicas da *Otimização* continuam sendo aplicáveis, com a ressalva de que possivelmente haveria maneiras mais simples ou mais precisas para a determinação das soluções[5].

Por fim, há ainda a questão de quão difícil, ou quão demorada, é a obtenção da informação dos valores da função objetivo e das funções de restrição: muitas vezes, para calcularmos o valor da função objetivo em um único ponto (ou seja, para um único vetor $\boldsymbol{x}$) um bom computador de última geração pode demorar horas ou dias. Esse é o caso, por exemplo, de um modelo detalhado da estrutura da asa de um avião; a engenharia, a economia, as ciências naturais, estão repletas de situações assim. Dessa forma, não seria prático prescrever métodos de otimização que dependessem de calcular essa função objetivo alguns milhares ou centenas de milhares de vezes: talvez não seja viável avaliar essas funções mais que algumas dezenas ou centenas de vezes. Uma outra regra então se justifica:

Regra de Custo da Informação

- Os métodos de otimização serão comparados entre si de acordo com os critérios:
  - Número de avaliações da função objetivo e das funções de restrição que são requeridas para determinação da solução. Quanto menos avaliações forem necessárias, melhor será considerado o

[3] O conceito de *caixa preta*, nas ciências, diz respeito a objetos cujas entradas e saídas podem ser observadas, mas cujo interior é inacessível.

[4] O leitor deve notar que, embora não saibamos qual é a expressão analítica de uma função que corresponde à caixa preta, tal função *existe*. Se o leitor se lembrar de como a Matemática define funções, verá que essa caixa preta atende a todos os requisitos para ser uma função.

[5] Se houver, entretanto, um número muito grande de restrições ou variáveis no problema, é possível que as técnicas de Otimização ainda sejam as mais adequadas para a determinação do ponto de ótimo, mesmo havendo expressões analíticas para as funções objetivo e de restrições.

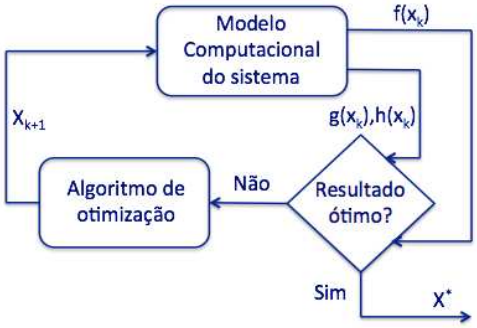

Figura 1.3: Diagrama do processo de otimização. A rotina de otimização fornece o vetor de variáveis de otimização, $\boldsymbol{x}$, para as rotinas que avaliam a função objetivo e de restrições. Essas rotinas devolvem os valores de $f(\boldsymbol{x})$, $g_i(\boldsymbol{x})$ e $h_j(\boldsymbol{x})$ para a rotina de otimização. A rotina de otimização, com essas avaliações, calcula um novo vetor de variáveis de otimização a ser avaliado, e assim por diante, até que seja encontrada uma aproximação da solução ótima $\boldsymbol{x}^*$.

método.

– Precisão e robustez. Quanto mais a solução fornecida pelo método se aproximar da solução exata do problema, melhor será considerado o método[6].

Agora sabemos o que estaremos fazendo ao longo deste livro: iremos construir *algoritmos*, que serão as implementações práticas dos *métodos de otimização*, cujo objetivo é determinar as soluções do problema (1.1). Esses algoritmos irão chamar sub-rotinas que executam a avaliação das funções objetivo e de restrições, devendo entretanto fazer a chamada dessas sub-rotinas o menor número de vezes que for possível. O diagrama da Figura 1.3 ilustra essa ideia.

1.2 Otimização Sem Restrições

Para começar a estudar a interpretação geométrica dos problemas de otimização, iniciaremos analisando a situação mais simples, do problema de minimização de uma função objetivo sem nenhuma restrição:

$$\boldsymbol{x}^* = \arg\ \min\ f(\boldsymbol{x}) \tag{1.7}$$

[6]O termo *precisão* designa a capacidade de um algoritmo de, estando próximo da solução exata do problema, aproximar ainda mais tal solução exata. O termo *robustez* por sua vez designa a capacidade do algoritmo de, estando distante da solução exata do problema, atingir as proximidades dessa solução. Assim, frequentemente um algoritmo é mais preciso e ao mesmo tempo menos robusto que outro, e vice-versa.

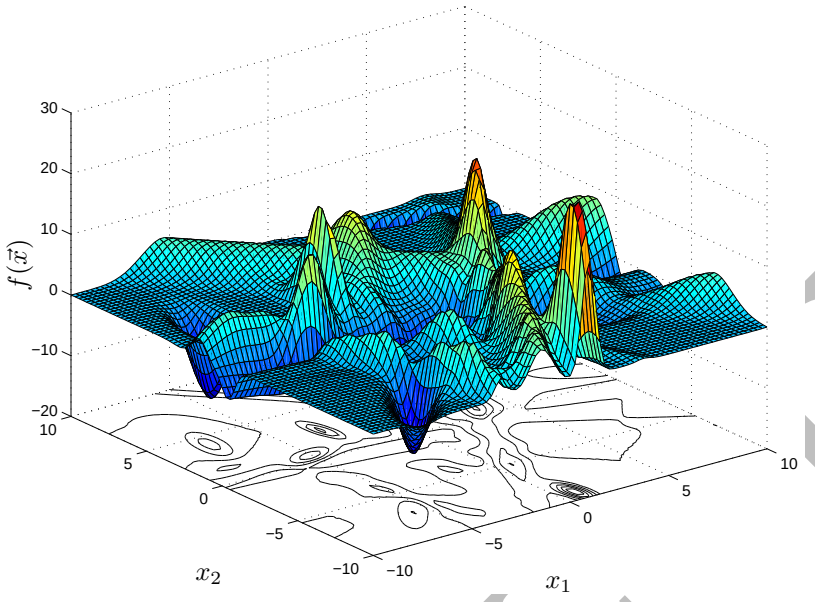

Figura 1.4: Superfície que representa o gráfico de uma função não-linear de duas variáveis reais. Essa superfície poderia representar uma função $f(\boldsymbol{x})$ cujo mínimo devesse ser determinado por um método de otimização. No "chão" do gráfico, encontram-se representadas as *curvas de nível* da função.

Para viabilizar a representação gráfica do problema, estaremos supondo a partir deste ponto que o vetor $\boldsymbol{x}$ possui apenas duas coordenadas, pertencendo, portanto, ao espaço $\mathbb{R}^2$. Evidentemente, na maioria das situações de interesse prático o número de coordenadas desse vetor é maior que dois; entretanto, duas variáveis são suficientes para discutirmos a maior parte das questões conceituais que se encontram por detrás da concepção dos métodos de otimização.

Embora estejamos supondo que a função objetivo $f(\cdot)$ não seja conhecida num contexto prático de otimização, essa função é sempre um objeto matemático muito bem definido. Assim, mesmo não sendo possível traçar explicitamente o gráfico da função objetivo, sabemos que isso é impossível devido às *regras da otimização*, anteriormente estabelecidas. Podemos afirmar que a superfície correspondente à função existe, e é desta superfície que estaremos colhendo *amostras* durante o processo de otimização, a cada vez que estivermos avaliando a função objetivo. A Figura 1.4 mostra uma superfície que corresponde ao gráfico de uma função não-linear de duas variáveis reais. Tal função poderia ser a função objetivo de um problema de otimização.

Uma representação que contém aproximadamente a mesma informação que a da Figura tridimensional 1.4, mas que utiliza apenas recursos gráficos bidimensionais é a das *curvas de nível* da função. A Figura 1.5 mostra as curvas de nível da mesma função representada na Figura 1.4. Essa representação, mais fácil de ser manipulada que a representação tridimensional, é normalmente mais útil que esta para ilustrar conceitos relacionados aos métodos de otimização.

Uma metáfora que pode ajudar a compreender o que é o processo de otimização pode ser apresentada da seguinte forma: imaginemos (aqui a imaginação é o mais importante) um ser matemático, o *Otimizador*. Ele vai ser lançado de pára-quedas

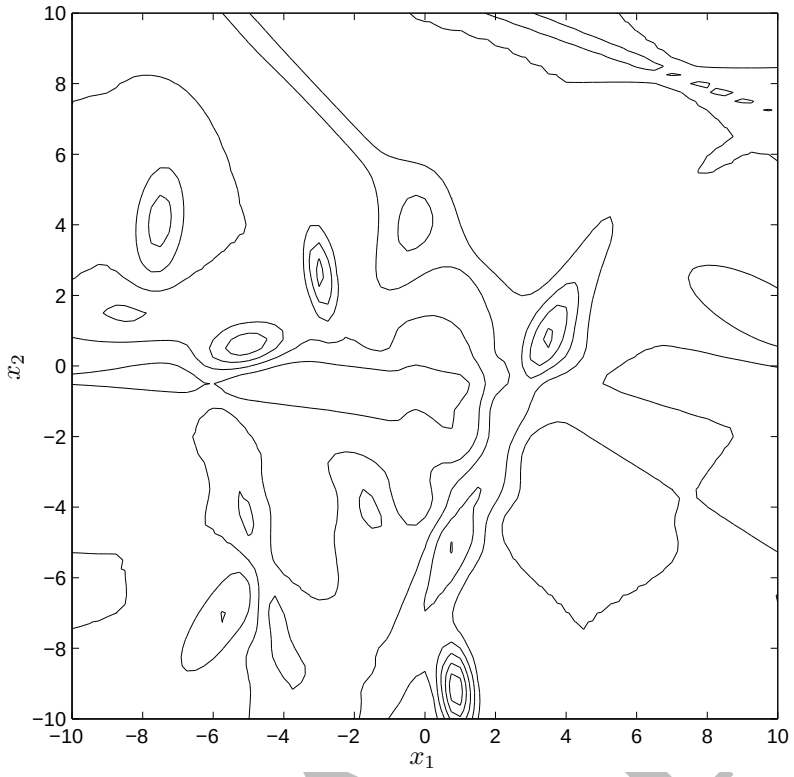

Figura 1.5: Gráfico de *curvas de nível* da mesma função não-linear de duas variáveis reais, $f(\boldsymbol{x})$, que encontra-se representada na figura 1.4.

em um ponto qualquer sobre a superfície da figura 1.4, e deverá caminhar sobre essa superfície, em busca do ponto mais baixo da mesma, o *ponto de mínimo*. O Otimizador, entretanto, deverá caminhar com uma venda cobrindo seus olhos, sem poder "olhar" para a superfície; a única informação que ele pode utilizar a respeito da superfície é a altura do ponto no qual ele estiver "pisando". Ele pode, entretanto, se "lembrar" das alturas dos pontos em que ele já tiver pisado anteriormente, fazendo uso dessa informação já adquirida para tomar a decisão de "para onde caminhar". Seu objetivo, além de chegar no ponto de mínima altura sobre a superfície, é fazer isso tendo utilizado o menor número possível de "passos". Essa situação imaginária ilustra bem o que é o problema de otimização. Construir os chamados *métodos de otimização* corresponde, dentro de nossa metáfora, a formular as estratégias a serem utilizadas pelo Otimizador em sua busca pelo ponto de mínimo.

Algumas características da função objetivo (ou seja, da superfície que está associada a essa função) definem que tipos de estratégias seriam efetivas para a otimização dessa função. Por exemplo, a função ser *diferenciável* implica na possibilidade de se tentar sua otimização fazendo uso do cálculo, pelo menos aproximado, de seu gradiente, que pode ser estimado numericamente a partir de amostras de valores da função. Se a função for *unimodal*, ou seja, se tiver um único ponto de mínimo, as estratégias para a determinação desse mínimo serão bem diferentes daquelas que seriam empregadas caso a função fosse *multimodal*, ou seja, caso tivesse

vários mínimos locais[7].

Com o objetivo de subsidiar a escolha de métodos adequados para a otimização de funções, podemos definir a seguinte classificação das funções:

(i) Modalidade: unimodal ou multimodal

(ii) Diferenciabilidade: diferenciável ou não-diferenciável

(iii) Convexidade: convexa, quasi-convexa, não-convexa

(iv) Linearidade: linear ou não-linear

(v) Escala: uni-escala ou multi-escala

Passamos a mostrar agora algumas superfícies "típicas", que exibem de maneira clara essas propriedades que "fazem a diferença" (o significado dessa classificação deve ficar claro à medida em que essa discussão for apresentada). Com esses exemplos de superfícies, discutiremos de maneira qualitativa possíveis estratégias para a otimização de funções com tais características. Essas estratégias serão depois desdobradas, nos capítulos posteriores, os quais serão dedicados a discutir em detalhe os *métodos de otimização* correspondentes a essas estratégias.

1.2.1 Estratégias de Direção de Busca

Vamos considerar em primeiro lugar a função cujo gráfico é mostrado na figura 1.6, e cujas curvas de nível estão representadas na Figura 1.7.

Para construir essa função, nós utilizamos um esquema bastante simples: o de uma *função quadrática*. A "receita" para a montagem do gráfico da figura 1.6 é dada por:

$$\begin{gathered} f(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{x}_0)' Q (\boldsymbol{x} - \boldsymbol{x}_0) \\ Q = \begin{bmatrix} 2 & 0.3 \\ 0.3 & 1 \end{bmatrix} \qquad \boldsymbol{x}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{gathered} \tag{1.8}$$

Claramente, o gráfico dessa função deve ser um parabolóide com mínimo no ponto $\boldsymbol{x}_0$. O Otimizador, entretanto, como já concordamos, não sabe disso: ele deve descobrir qual é o ponto de mínimo da função objetivo utilizando apenas "amostras" de valores dessa função. Uma estratégia razoável de procedimento para o Otimizador seria:

Método do Gradiente

Passo 1: O Otimizador, localizado inicialmente em um ponto aleatório sobre o mapa da função, toma amostras da função próximas de onde ele se encontra atualmente. Com essas amostras, ele descobre em qual direção a função decresce mais rapidamente, pelo menos sob

[7] Falamos de *mínimos locais* para designar pontos que são de mínimo para uma vizinhança ao seu redor, e de *mínimos globais* para designar o ponto em que a função objetivo atinge seu mínimo valor em todo o domínio considerado.

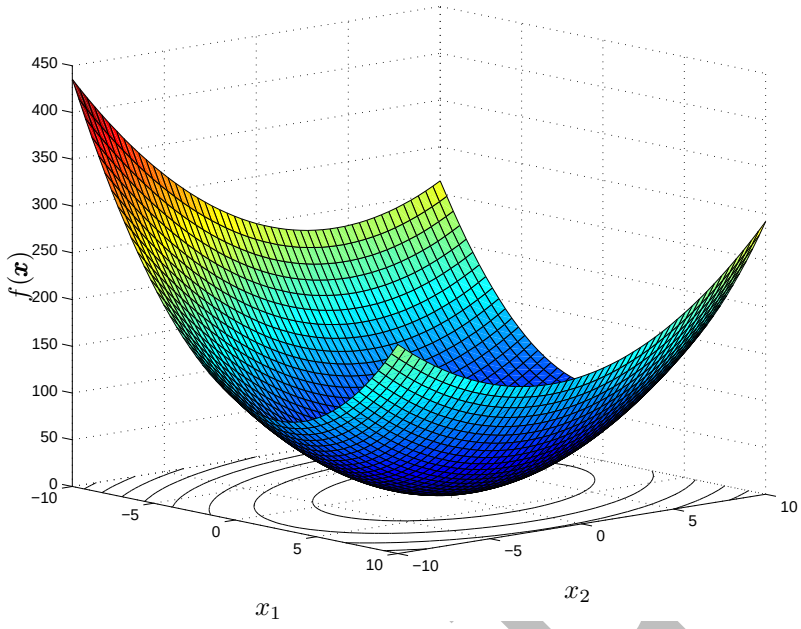

Figura 1.6: Superfície que representa o gráfico de uma função quadrática $f(\boldsymbol{x})$ de duas variáveis reais. No "chão" do gráfico, encontram-se representadas as *curvas de nível* da função.

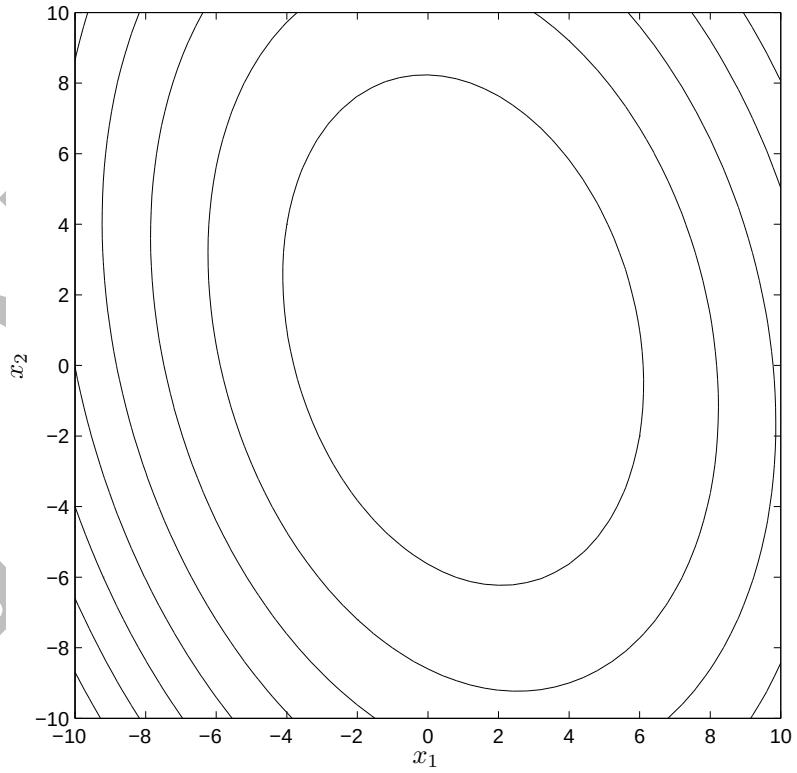

Figura 1.7: Gráfico de *curvas de nível* da mesma função quadrática de duas variáveis reais, $f(\boldsymbol{x})$, que encontra-se representada na Figura 1.6.

o ponto de vista da informação localmente disponível para ele. Em terminologia matemática, o Otimizador calcula uma aproximação numérica do *gradiente* da função no ponto atual, que é o oposto da direção em que a função decresce mais rapidamente.

Passo 2: O Otimizador caminha em linha reta, na direção contrária ao gradiente da função, continuando a andar enquanto estiver sentindo que a função está decrescendo, parando de andar, portanto, assim que percebe que a função volta a crescer nessa direção.

Passo 3: O Otimizador decide agora se ele pára, ou seja, se ele considera que já se encontra suficientemente próximo do ponto de mínimo da função, ou se ele continua a busca, retornando ao Passo 1, para escolher nova direção de caminhada.

O *método do gradiente*, assim esboçado, é um dos métodos de otimização mais primitivos, tendo sido proposto nos primórdios da teoria de otimização, estando hoje obsoleto. Esse método é, entretanto, o protótipo mais simples de toda uma família de métodos, os *métodos de direção de busca*, que incluem importantes métodos hoje utilizados, que sempre têm a estrutura assim descrita:

Métodos de Direção de Busca

Passo 1: O Otimizador toma amostras da função nas proximidades de onde ele se encontra atualmente. Com essas amostras, ele descobre em qual direção a função decresce mais rapidamente, pelo menos sob o ponto de vista da informação localmente disponível para ele. Em terminologia matemática, o Otimizador calcula uma aproximação numérica do *gradiente* da função no ponto atual, que é o oposto da direção em que a função decresce mais rapidamente.

Passo 2: Levando em consideração o gradiente calculado no ponto atual, assim como todo o histórico de gradientes anteriormente calculados e de valores de função objetivo amostrados em pontos que o Otimizador visitou anteriormente, ele tenta "adivinhar" qual seria a direção mais provável em que o mínimo da função devesse estar.

Passo 3: O Otimizador caminha em linha reta, na direção em que ele supõe que o mínimo esteja, continuando a andar enquanto estiver sentindo que a função está decrescendo, parando de andar, portanto, assim que percebe que a função volta a crescer nessa direção.

Passo 4: O Otimizador decide agora se ele pára, ou seja, se ele considera que já se encontra suficientemente próximo do ponto de mínimo da função, ou se ele continua a busca, retornando ao Passo 1, para escolher nova direção de caminhada.

Qualquer estratégia de "direção de busca" irá funcionar para determinar o mínimo da função mostrada na Figura 1.6, pois esta função é bastante simples. Para esses métodos funcionarem, os requisitos que encontram-se implícitos sobre a função são:

- A função é *unimodal*, ou seja, tem um único mínimo global, no interior de uma única bacia de atração[8]. Dessa forma, o Otimizador não precisa se preocupar com a possível existência de outros mínimos diferentes daquele que ele localizar.

- A função é *diferenciável*, ou seja, não só é possível calcular, de forma significativa, aproximações do gradiente da função em qualquer ponto do espaço, como, principalmente, o gradiente da função contém informação significativa sobre a forma como a função varia nas vizinhanças do ponto em que tiver sido calculado. Dessa forma, o Otimizador consegue encontrar direções para as quais possa caminhar, nas quais ele consegue observar a diminuição do valor da função objetivo.

Consideremos agora a função mostrada na Figura 1.8, que tem suas curvas de nível mostradas na Figura 1.9. Essa função, muito menos simples que a função quadrática anteriormente considerada, continua sendo adequadamente otimizada por métodos de direção de busca: ela é unimodal (possui um único mínimo, o ponto $\boldsymbol{x} = \begin{bmatrix} 1 & 1 \end{bmatrix}'$, no interior de uma única bacia de atração), e é diferenciável (possui gradiente bem definido em todos os pontos).

Essa função já é capaz de "confundir" um Otimizador que utilizar simplesmente uma estratégia de gradiente: quando o Otimizador chega no fundo do "vale" existente na topografia da função, e tem de encontrar o ponto mais baixo desse vale, o padrão de mudança da direção do gradiente torna o método do gradiente muito ineficiente. Outros métodos de direção de busca, no entanto, não encontram dificuldades para minimizar esta função.

1.2.2 Estratégias de Exclusão de Regiões

Consideremos agora a função $f(\boldsymbol{x})$, ainda unimodal, porém agora não mais diferenciável, cujo gráfico está mostrado na Figura 1.10, e cujas curvas de nível estão representadas na Figura 1.11. Este tipo de função em geral traz dificuldades para as estratégias de otimização do tipo *direções de busca*.

Ao contrário do que pode parecer à primeira vista, a dificuldade não está na impossibilidade de calcularmos o gradiente da função: na imensa maioria das vezes, uma função não diferenciável de interesse prático é *diferenciável em quase todo ponto*. Esse é o caso da função representada na Figura 1.10: seu gradiente deixa de existir apenas em algumas regiões específicas, que estão situadas em algumas linhas sobre o mapa da função. Em todos os outros pontos, o gradiente é bem definido e pode ser calculado. Assim, se um Otimizador estivesse otimizando uma função não

[8]Uma *bacia de atração* é a região ao redor de um mínimo local na qual as curvas de nível da função são fechadas, ou seja, a região na qual um método de direção de busca irá convergir para tal mínimo.

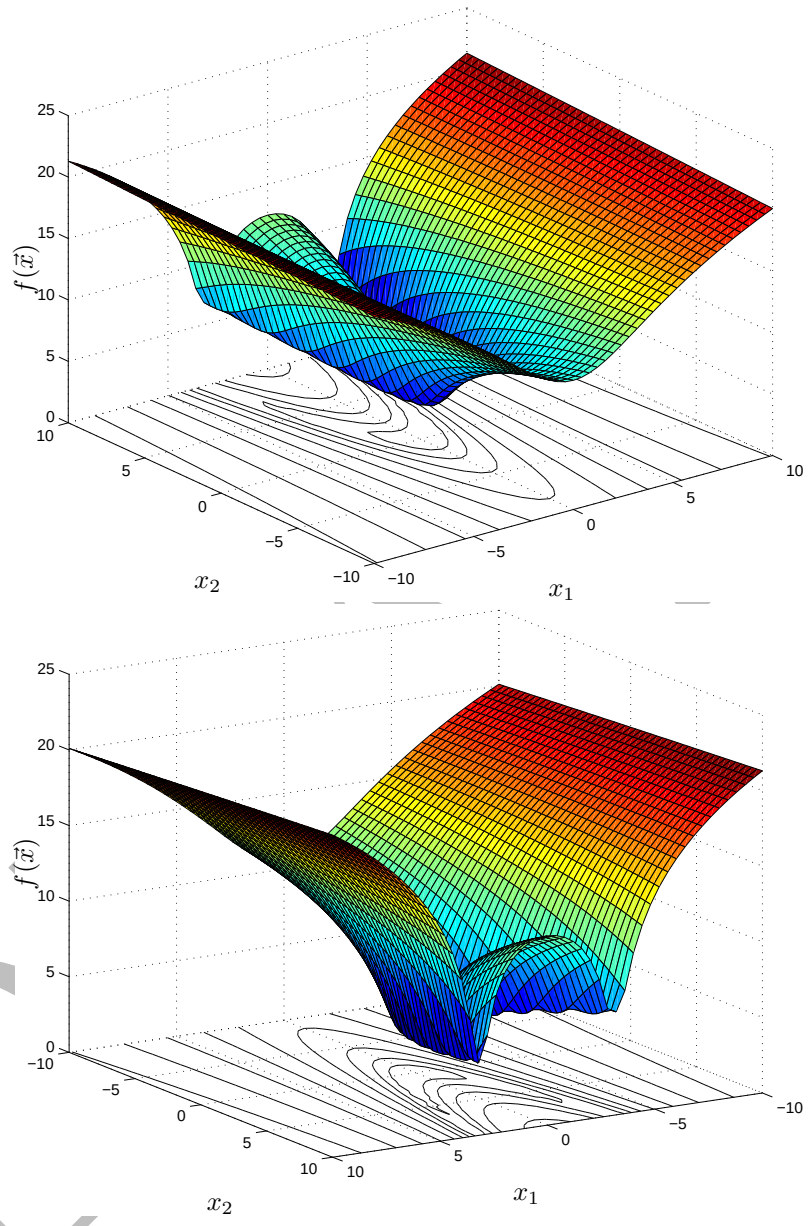

Figura 1.8: Superfície que representa o gráfico de uma função unimodal diferenciável $f(\boldsymbol{x})$ de duas variáveis reais, mostrada em duas vistas diferentes. No "chão" dos gráficos, encontram-se representadas as *curvas de nível* da função.

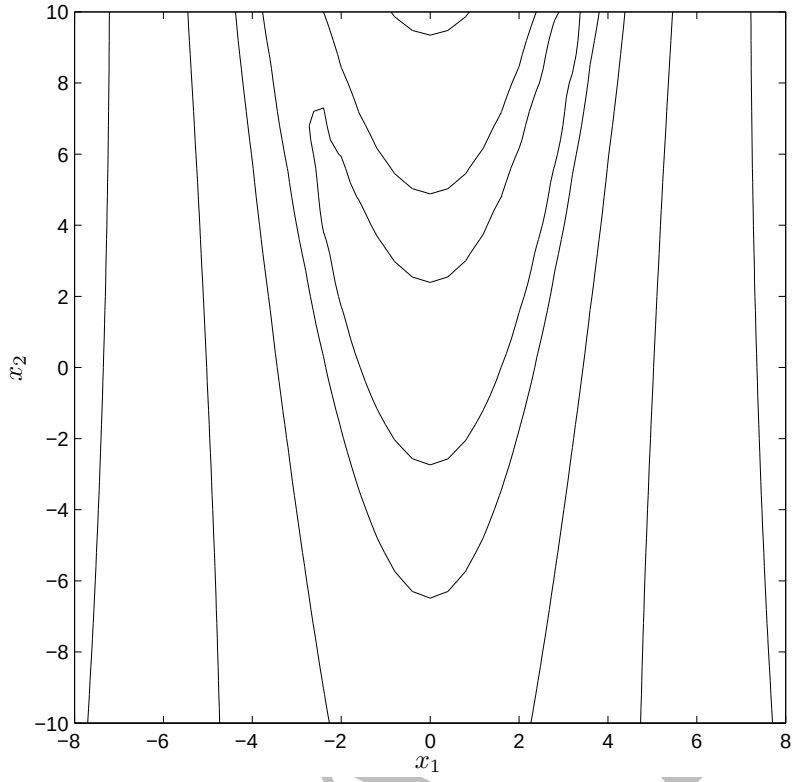

Figura 1.9: Gráfico de *curvas de nível* da mesma função unimodal diferenciável de duas variáveis reais, $f(\boldsymbol{x})$, que encontra-se representada na figura 1.8.

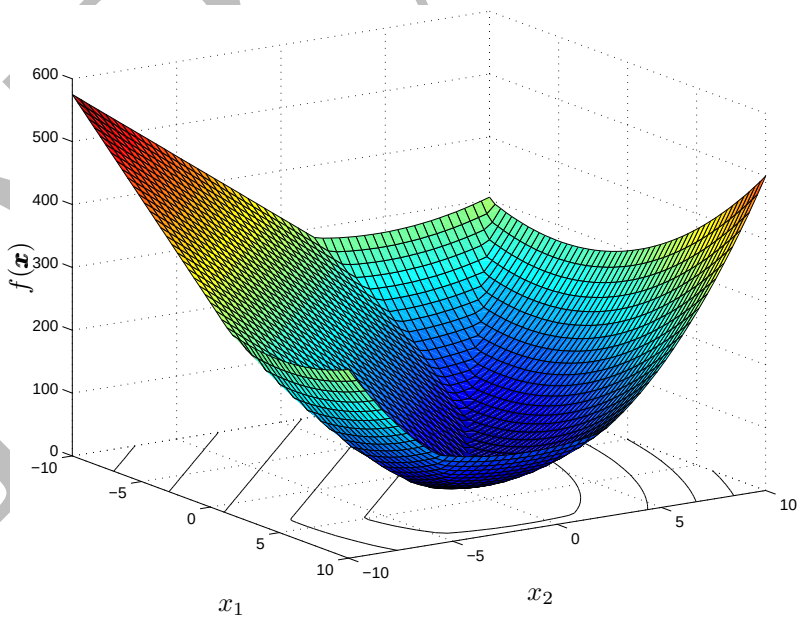

Figura 1.10: Superfície que representa o gráfico de uma função não diferenciável $f(\boldsymbol{x})$ de duas variáveis. No "chão" do gráfico, encontram-se representadas as *curvas de nível* da função.

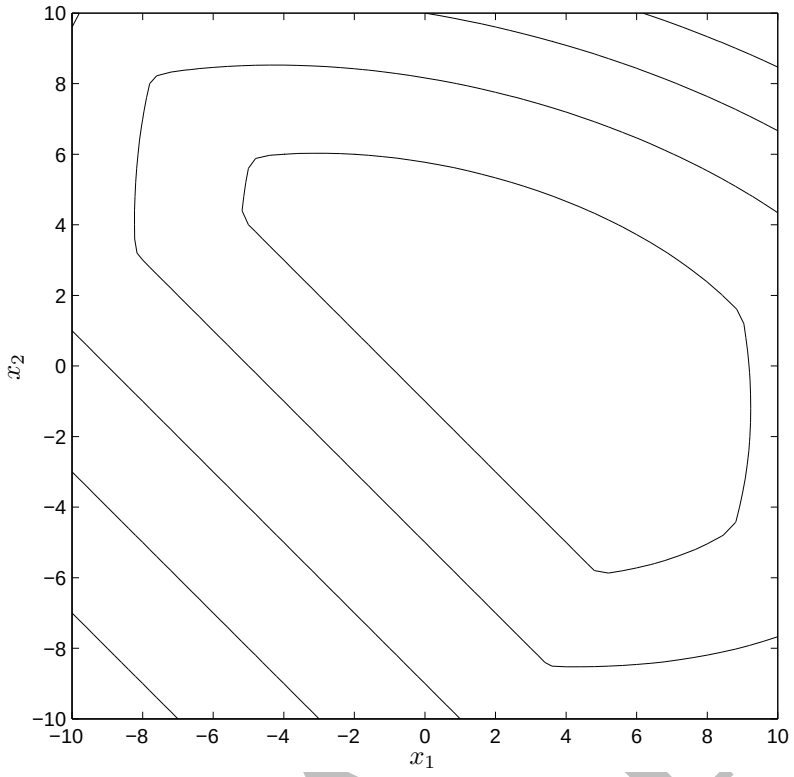

Figura 1.11: Gráfico de *curvas de nível* da mesma função não diferenciável de duas variáveis reais, $f(\boldsymbol{x})$, que encontra-se representada na Figura 1.10.

diferenciável e encontrasse um ponto no qual fosse impossível calcular o gradiente, bastaria ele se deslocar um pouco do ponto, para outro ponto próximo: lá o gradiente poderia ser calculado, e o processo de otimização poderia prosseguir.

O problema com as funções não diferenciáveis, quando submetidas a métodos de direção de busca, é que o cálculo da direção de busca, na qual o Otimizador deve caminhar, é feito a partir da informação obtida pelo cálculo do gradiente (o gradiente atual e o gradiente em pontos anteriores). O Otimizador, ao caminhar nessa direção, espera que a direção tenha validade não apenas pontual: ele espera poder caminhar uma certa distância sobre essa direção, até que a função objetivo pare de decrescer, e ele tenha de mudar de direção. Ora, se a função objetivo muda de comportamento repentinamente nos locais onde a função é não-diferenciável, a informação da direção de busca, obtida com o uso de gradientes pode ser inteiramente inadequada para representar o comportamento da função, mesmo a pequenas distâncias do ponto atual. A otimização por esses métodos pode assim se tornar inviável. Tal dificuldade, por outro lado, não é associada a um ou outro caso específico de método de direção de busca: ela é intrínseca a toda a família dos métodos de direção de busca. A dificuldade é ilustrado na Figura 1.12.

Funções não-diferenciáveis estão longe de ser raras, dentro dos modelos de sistemas que temos interesse em otimizar. Por essa razão, justifica-se a formulação de uma família de métodos diferente, que não esteja sujeita a tal dificuldade: os *métodos de exclusão de regiões*. Para formular a nova estratégia, como estamos abrindo mão da premissa de *diferenciabilidade* da função objetivo, introduzimos em

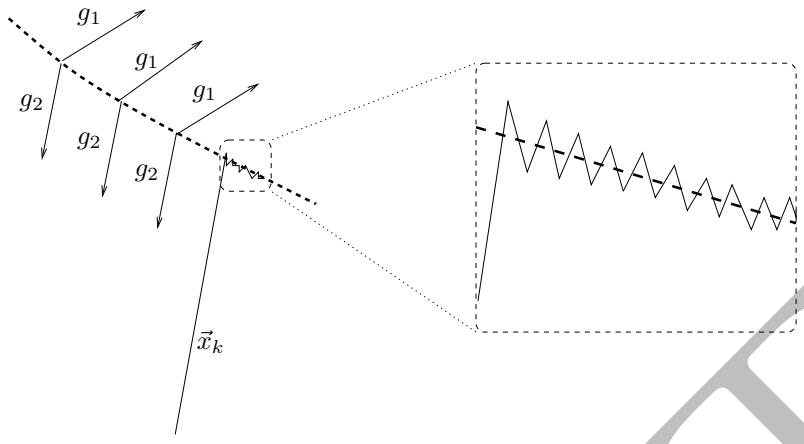

Figura 1.12: Não-diferenciabilidade atratora, representada pela linha tracejada. Acima dessa não-diferenciabilidade, os gradientes da função são representados por g_1, e abaixo por g_2. Exatamente na não-diferenciabilidade, o gradiente da função muda subitamente (ou seja, o gradiente é descontínuo sobre essa linha). A Figura mostra ainda a trajetória de um Otimizador que utiliza uma estratégia de direções de busca, percorrendo uma sequência de pontos $\boldsymbol{x}_k$. Quando atinge a não-diferenciabilidade atratora, o Otimizador passa a se mover segundo passos muito pequenos. Uma ampliação desse movimento é mostrada na Figura à direita.

lugar desta a premissa de *convexidade* dessa função[9].

A propriedade associada à convexidade que iremos utilizar na nova estratégia de otimização pode ser entendida da seguinte forma:

- Uma curva de nível de uma função convexa sempre delimita uma região convexa em seu interior.
- O vetor gradiente, por sua vez, é sempre perpendicular à curva de nível que passa pelo ponto onde o vetor foi calculado.
- Assim, a reta perpendicular ao vetor gradiente que passa no ponto onde esse vetor foi calculado é tangente à curva de nível.
- Devido à convexidade da região no interior da curva de nível, esta região sempre fica inteiramente localizada em apenas um dos lados dessa reta tangente (essa reta não corta a região no interior da curva de nível), ou seja, do lado oposto àquele para onde aponta o vetor gradiente.

Isso significa que, se calcularmos o gradiente de uma função convexa num ponto, podemos ter certeza que o ponto de mínimo dessa função, que se localiza necessariamente no interior da curva de nível fechada que passa nesse ponto, está no semi-plano oposto ao do vetor gradiente, delimitado pela reta perpendicular ao vetor gradiente. Esse conceito é ilustrado na Figura 1.13.

O procedimento do Otimizador agora é descrito por:

[9]É claro que às vezes as funções a serem otimizadas serão convexas e às vezes não serão. Se não forem, os métodos de exclusão de regiões poderão falhar.

Métodos de Exclusão de Regiões

Passo 1: O Otimizador adquire informação em alguns pontos próximos do atual, e faz uma estimativa do gradiente da função objetivo nesse ponto (se ele estiver exatamente sobre um ponto em que a função é não-diferenciável, admitamos, para simplificar, que ele se movimenta para algum ponto próximo do atual, em que a função é diferenciável).

Passo 2: Com base no gradiente, o Otimizador descobre qual é a reta tangente à curva de nível que passa pelo ponto atual, e descarta todo o semi-plano que se encontra do lado dessa reta para o qual o vetor gradiente aponta (o Otimizador tem certeza de que o mínimo da função *não está* nesse semi-plano).

Passo 3: O Otimizador se move para algum ponto no interior da região que ainda não está descartada, de preferência para um ponto aproximadamente "no meio" dessa região[10].

Passo 4: O Otimizador decide se existem indícios suficientes de que o novo ponto já esteja suficientemente próximo do mínimo da função, caso em que o processo termina, ou se a otimização deve continuar. Nesse último caso, retorna ao Passo 1.

Deve-se observar que agora a convergência da sequência de pontos para o ponto de mínimo da função objetivo ocorre em virtude da diminuição sistemática que é feita, a cada iteração do método, da região em que esse ponto de mínimo pode estar localizado. Com o avançar das iterações, a região fica cada vez menor, e o novo ponto, que é escolhido dentro dessa região, tende a ficar cada vez mais próximo do ponto de mínimo. Não há a possibilidade, agora, de uma não-diferenciabilidade impedir a convergência do método.

Uma sequência de iterações de um *método de exclusão de região* é ilustrada na Figura 1.13.

1.2.3 Estratégias de Populações

Grande parte das funções objetivo que queremos otimizar na prática, infelizmente, não é unimodal. Por consequência, tanto as estratégias de direção de busca quanto as estratégias de exclusão de regiões irão falhar em sua otimização[11]. Uma função desse tipo é mostrada na Figura 1.14, e suas curvas de nível são mostradas na Figura 1.15.

De fato, essa função possui diversas bacias de atração diferentes, associadas a diferentes mínimos locais. Na tentativa de se fazer a otimização desta função por meio de um mecanismo de direção de busca, por exemplo, o resultado sempre será o

[10] A maneira exata de escolher o novo ponto varia de método para método.

[11] Deve-se lembrar que se uma função não é unimodal, ela também não pode ser convexa.

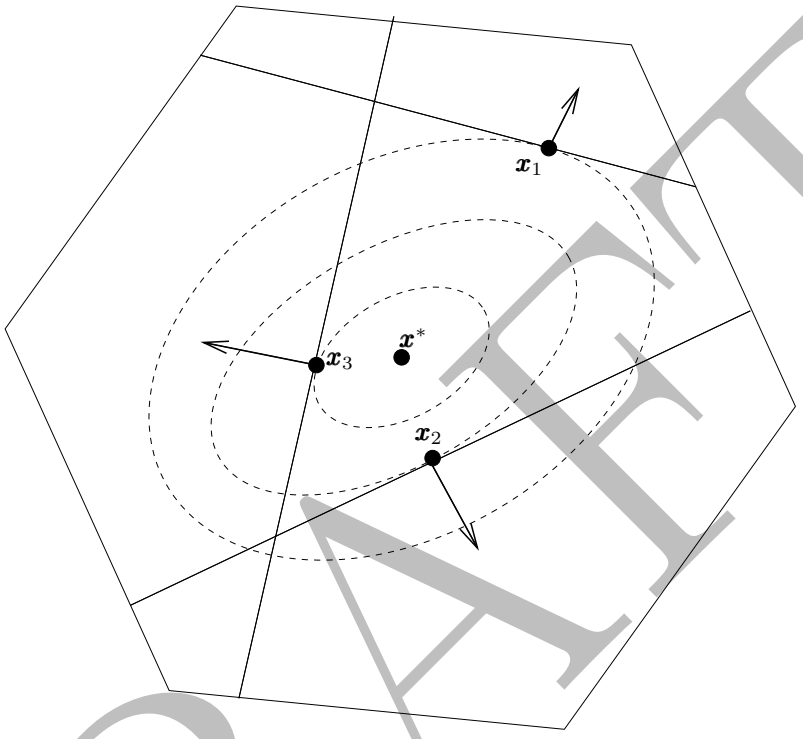

Figura 1.13: Iterações de um método de exclusão de regiões, mostradas sobre as curvas de nível de uma função cujo mínimo exato é $\boldsymbol{x}^*$. Suponha-se que, *a priori*, se sabe que o mínimo da função se encontra na região delimitada pelo hexágono. Após avaliar o gradiente da função em $\boldsymbol{x}_1$, o Otimizador pode concluir que o mínimo $\boldsymbol{x}^*$, cuja localização ainda não é conhecida, encontra-se abaixo da reta perpendicular a esse gradiente, que passa nesse ponto. Um novo ponto $\boldsymbol{x}_2$ é escolhido no interior da região restante. O gradiente nesse ponto também é calculado, trazendo a informação de que o ponto $\boldsymbol{x}^*$ não se encontra abaixo da reta perpendicular ao gradiente que passa nesse ponto. A seguir um novo ponto $\boldsymbol{x}_3$ é escolhido, e o processo se repete, levando à conclusão de que $\boldsymbol{x}^*$ não se encontra à esquerda da reta que passa por esse ponto. Observa-se que a cada passo vai diminuindo a região onde é possível que $\boldsymbol{x}$ se encontre. O processo termina quando a região "possível" é suficientemente pequena.

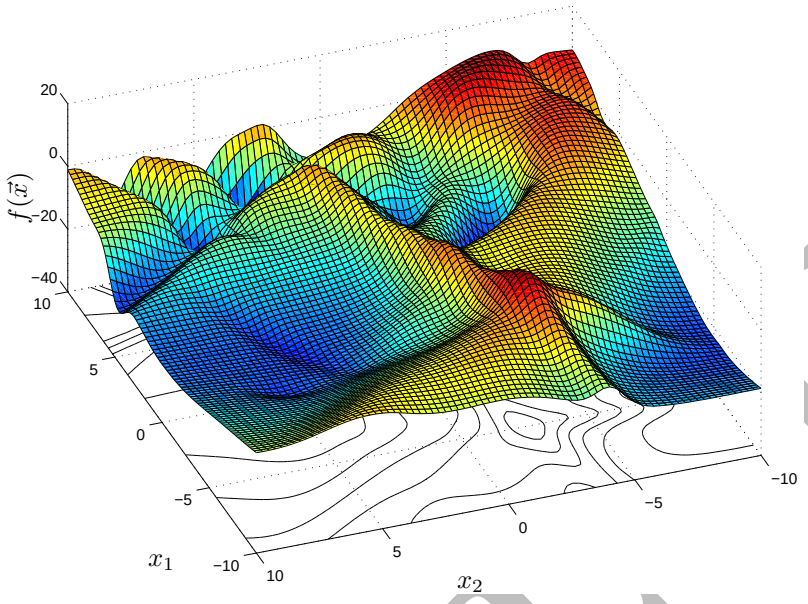

Figura 1.14: Superfície que representa o gráfico de uma função multimodal $f(\boldsymbol{x})$ de duas variáveis. No "chão" do gráfico, encontram-se representadas as *curvas de nível* da função.

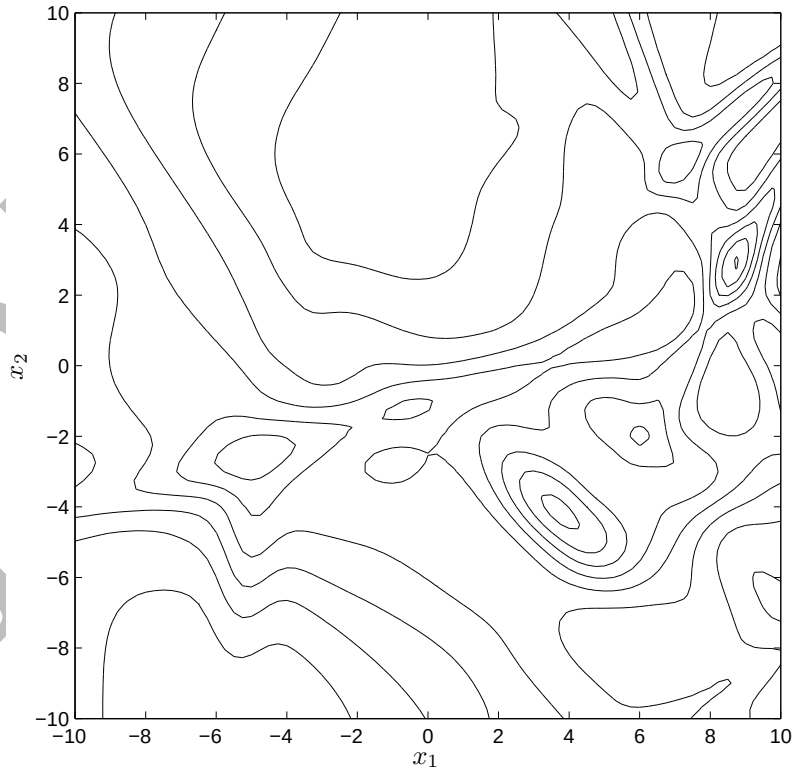

Figura 1.15: Gráfico de *curvas de nível* da mesma função multimodal de duas variáveis reais, $f(\boldsymbol{x})$, que encontra-se representada na Figura 1.14.

ponto de mínimo local associado à bacia de atração onde a busca tiver sido iniciada. Para se atingir o mínimo global com algum grau de certeza, é necessário "investigar" a função em suas diferentes bacias de atração.

A estratégia a ser adotada envolve agora o trabalho não mais de um único Otimizador sozinho: um grupo de Otimizadores será agora chamado a cooperar, para tentar descobrir a localização do ponto de mínimo da função. Essa estratégia é descrita a seguir:

Métodos de Populações

Passo 1: Um grupo de Otimizadores encontra-se espalhado pela região onde acredita-se que se encontre o ponto de mínimo da função. Cada um dos Otimizadores avalia a função objetivo no ponto onde ele se encontra.

Passo 2: Os Otimizadores se comunicam, e trocam informações a respeito dos valores da função objetivo em cada ponto.

Passo 3: Um pequeno sub-grupo do grupo de Otimizadores, que estiver nas melhores localizações fica parado. Os demais Otimizadores se movimentam, com movimentos que simultaneamente: (i) os façam se aproximarem dos otimizadores melhor localizados; e (ii) os façam explorarem outras regiões, diferentes daquelas já visitadas anteriormente pelo grupo de Otimizadores.

Passo 4: Cada um dos Otimizadores avalia a função objetivo no ponto para onde foi.

Passo 5: Os otimizadores decidem se o processo de otimização já produziu melhoria suficiente na função objetivo, caso em que o processo se interrompe; do contrário, eles retornam ao Passo 2.

Há diferentes maneiras de realizar cada um dos passos do esquema descrito acima. Cada combinação dessas diferentes fórmulas leva a um método específico diferente.

Esse tipo de estratégia pode ser pensado como um mecanismo útil para localizar não exatamente o mínimo global da função objetivo, mas sim a bacia de atração na qual este se encontra. Como usualmente os esquemas de "populações" requerem um número muito maior de avaliações da função objetivo até atingirem o ponto de mínimo da função objetivo, estas técnicas são muito "caras" comparado aos esquemas de direções de busca ou de exclusão de regiões. Assim sendo, a ideia é que o esquema de populações apenas conduza o Otimizador às proximidades do ponto de mínimo global. Uma vez dentro da bacia de atração do mínimo global, o Otimizador passa a adotar uma estratégia por exemplo de direção de busca, que o leva muito mais rapidamente ao mínimo da função. Esse raciocínio funcionaria corretamente, por exemplo, na otimização da função ilustrada na Figura 1.14. A Figura 1.16 mostra sucessivas aproximações do ponto de mínimo global da função, que terminam por

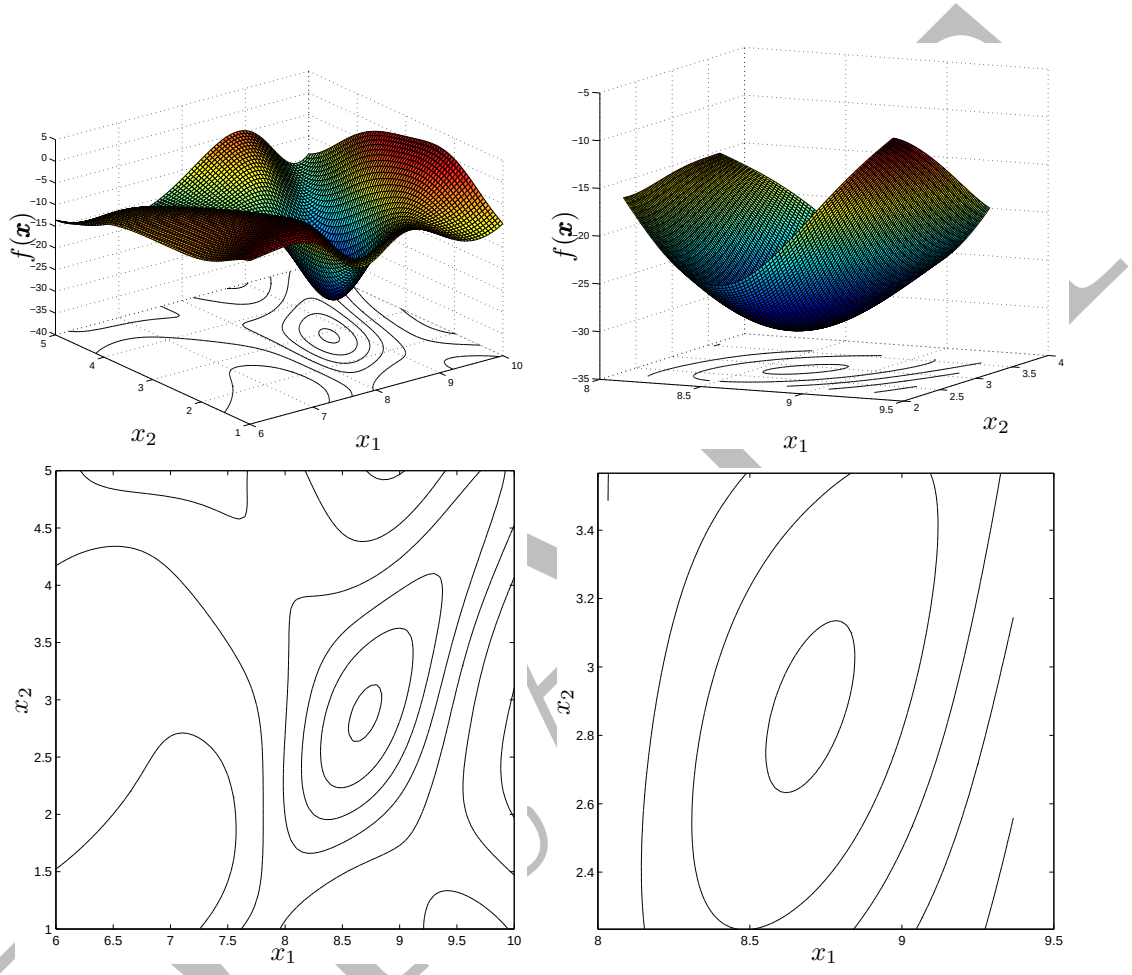

Figura 1.16: Superfície que representa o gráfico da mesma função multimodal $f(\boldsymbol{x})$ de duas variáveis mostrada na Figura 1.14, em sucessivas aproximações da região onde se encontra seu mínimo global. Acima, estão representados os gráficos da superfície, e abaixo as correspondentes curvas de nível na mesma região. Deve-se observar que, na região mais próxima ao mínimo, a função tem a "aparência" de uma função unimodal.

"se parecer" com uma função convexa e unimodal, nas proximidades do ponto de mínimo. Na região correspondente à última aproximação mostrada na Figura, um método de direções de busca ou de exclusão de regiões funcionaria. O método de populações então poderia ser paralisado assim que houvesse indícios suficientes de que determinado ponto se encontra no interior da bacia de atração do mínimo global, sendo iniciado um outro método de otimização nesse ponto.

Essa lógica de mudança de um método de população para outro tipo de método nem sempre funciona. Um exemplo de situação em que tal esquema não funcionaria é a função representada na Figura 1.17. Nessa figura, vemos um exemplo de função em que ocorre o fenômeno das *múltiplas escalas*. Essa função, olhada a uma "grande distância", parece ter algumas bacias de atração. Olhada "de perto", ela revela uma estrutura muito mais complexa, com a presença de dezenas de pequenas "sub-bacias" onde parecia estar cada uma das bacias de atração inicialmente aparentes. Um método de direção de busca que fosse iniciado no interior dessa "grande bacia" aparente iria quase certamente falhar na busca do mínimo global, ficando provavelmente detido em algum dos múltiplos mínimos locais existentes nessa região. Funções desse tipo vão requerer a utilização de um esquema *de população* para realizar sua otimização, do princípio ao fim, sem a possibilidade de mudança para outro tipo de método.

1.3 Otimização com Restrições de Desigualdade

A próxima situação a ser estudada aqui é aquela em que, na formulação do problema de otimização, aparecem as chamadas *restrições de desigualdade*:

$$\begin{aligned} \boldsymbol{x}^* = &\ \arg\ \min\ f(\boldsymbol{x}) \\ \text{sujeito a:}\ &\{g_i(\boldsymbol{x}) \leq 0, i, \ldots, p \end{aligned} \tag{1.9}$$

Essa descrição do problema significa, conforme já foi visto, que o ponto de ótimo $\boldsymbol{x}^*$ a ser determinado deve satisfazer às p desigualdades:

$$\begin{aligned} &g_1(\boldsymbol{x}^*) \leq 0 \\ &g_2(\boldsymbol{x}^*) \leq 0 \\ &\vdots \\ &g_p(\boldsymbol{x}^*) \leq 0 \end{aligned} \tag{1.10}$$

1.3.1 Interpretação geométrica de uma restrição de desigualdade

Examinemos primeiro o que significa uma dessas desigualdades apenas, por exemplo a primeira:

$$g_1(\boldsymbol{x}) \leq 0 \tag{1.11}$$

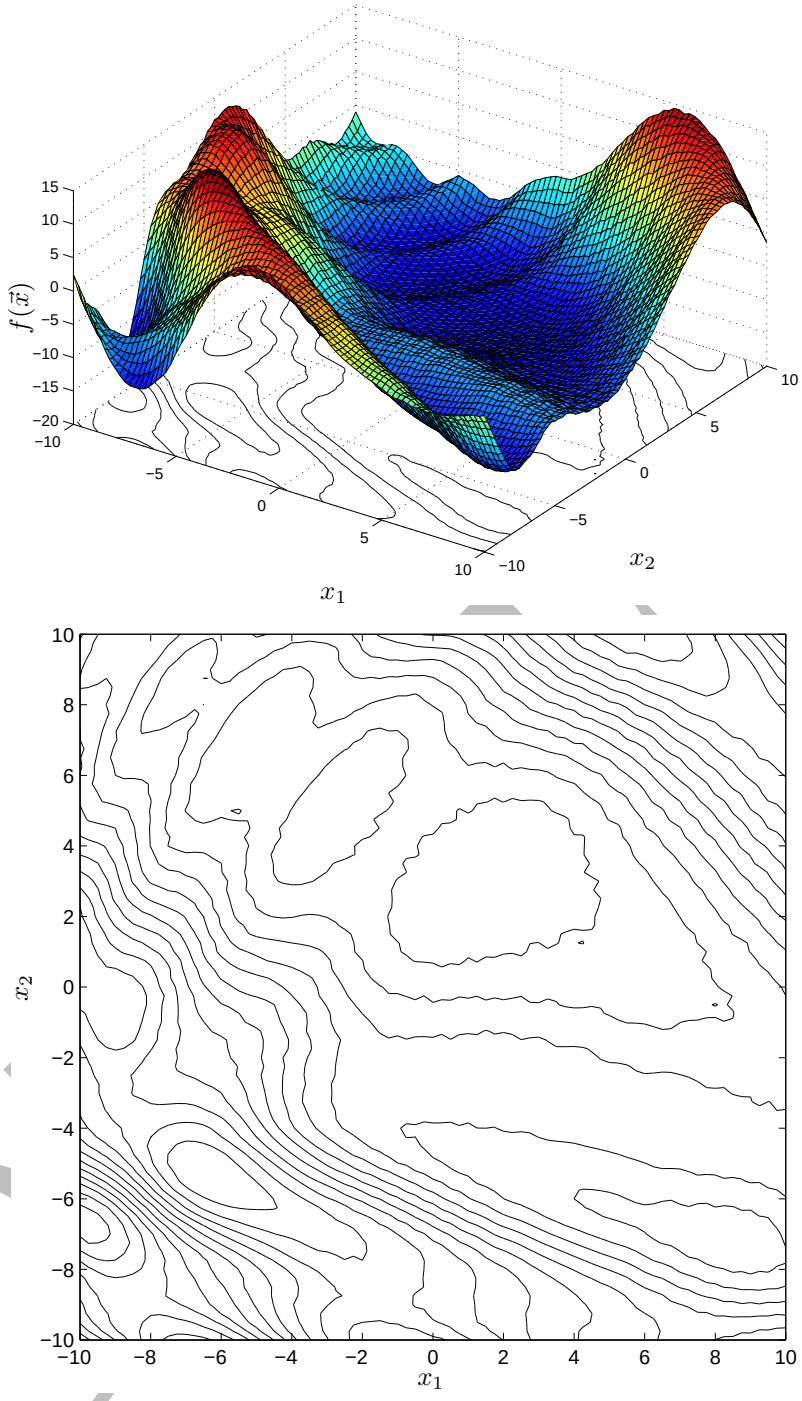

Figura 1.17: Superfície que representa o gráfico de uma função multimodal $f(\boldsymbol{x})$ de duas variáveis que apresenta a característica de *múltiplas escalas*. Sucessivas aproximações da região onde se encontra seu mínimo global irão revelar sucessivas estruturas de menor escala, que possuem múltiplas bacias de atração dentro de cada bacia de atração maior. Acima, estão representados os gráficos da superfície, e abaixo as correspondentes curvas de nível na mesma região. Deve-se observar pelo primeiro par de gráficos, que onde esperaríamos encontrar uma única bacia de atração, encontramos, no exame mais detalhado, uma estrutura com múltiplas pequenas "sub-bacias".

Admitamos que a função $g_1(\cdot)$ seja contínua. Se isso for verdade, essa função nunca muda "bruscamente" de valor. Por exemplo, para passar de um valor negativo para um valor positivo, necessariamente ela tem de passar pelo valor zero. Isso significa que, considerando todo o espaço $\mathbb{R}^n$ dos $\boldsymbol{x}$, se houver um subconjunto $\mathcal{P}_1 \subset \mathbb{R}^n$ para cujos pontos $\boldsymbol{x}$ a função $g_1(\cdot)$ fica positiva, e outro subconjunto $\mathcal{N}_1 \subset \mathbb{R}^n$ para o qual a função $g_1(\cdot)$ fica negativa, então tem de haver um conjunto $\mathcal{G}_1 \subset \mathbb{R}^n$ para o qual a função se anula, e que separa $\mathcal{P}_1$ de $\mathcal{N}_1$.

Matematicamente, definimos o conjunto $\mathcal{P}_1$ da seguinte forma:

$$\mathcal{P}_1 \triangleq \{\boldsymbol{x} \mid g_1(\boldsymbol{x}) > 0\} \tag{1.12}$$

Em palavras, essa expressão deve ser lida como: *O conjunto $\mathcal{P}_1$ é definido como ($\triangleq$) o conjunto dos pontos $\boldsymbol{x}$ tais que ($\mid$) a função $g_1(\cdot)$ avaliada nesses pontos seja maior que zero.* De forma similar, são definidos os conjuntos $\mathcal{G}_1$ e $\mathcal{N}_1$:

$$\begin{aligned} \mathcal{G}_1 &\triangleq \{\boldsymbol{x} \mid g_1(\boldsymbol{x}) = 0\} \\ \mathcal{N}_1 &\triangleq \{\boldsymbol{x} \mid g_1(\boldsymbol{x}) < 0\} \end{aligned} \tag{1.13}$$

A Figura 1.18 ilustra tais conjuntos, para um espaço de duas dimensões.

Quando inserimos, no problema de otimização, a exigência de que $g_1(\boldsymbol{x}^*) \leq 0$, queremos dizer que iremos aceitar como soluções do problema de otimização apenas pontos que sejam pertencentes ao conjunto $\mathcal{N}_1$ ou ao conjunto $\mathcal{G}_1$. Não serão admissíveis pontos pertencentes ao conjunto $\mathcal{P}_1$, que será assim denominado *conjunto infactível*, ou *região infactível*. Diz-se então que o *conjunto factível*, ou a *região factível* $\mathcal{F}_1$ é a união de $\mathcal{G}_1$ e $\mathcal{N}_1$:

$$\mathcal{F}_1 = \mathcal{G}_1 \cup \mathcal{N}_1 \tag{1.14}$$

Se aplicarmos agora um dos métodos de otimização irrestrita, discutidos nas seções anteriores, para a minimização da função $f(\boldsymbol{x})$, há duas possibilidades para a localização do ponto de mínimo: ele tem de estar em $\mathcal{P}_1$ ou em $\mathcal{F}_1$. Se a última hipótese ocorrer, a solução do problema será o ponto de mínimo encontrado. No entanto, se o mínimo *irrestrito* (ou seja, o mínimo encontrado sem levar em consideração a restrição $g_1(\boldsymbol{x}^*) \leq 0$) estiver na região *infactível* $\mathcal{P}_1$, alguma modificação deverá ser introduzida no mecanismo de otimização, para que seja possível localizar o ponto de ótimo $\boldsymbol{x}^*$ que minimiza a função objetivo $f(\cdot)$ nos pontos pertencentes ao conjunto factível $\mathcal{F}_1$.

Esse é, basicamente, o problema da *otimização restrita com restrições de desigualdade*: determinar o ponto $\boldsymbol{x}^* \in \mathcal{F}$ (ou seja, pertencente à região factível) que minimiza a função $f(\cdot)$ nessa região (ou seja, que produz o menor valor dessa função, quando comparado com os valores da função em todos os demais pontos da região factível).

1.3.2 Interpretação geométrica de várias restrições de desigualdade

Antes de discutirmos como modificar os mecanismos de otimização para lidar com problemas de *otimização restrita*, vamos procurar entender o que significa o sistema

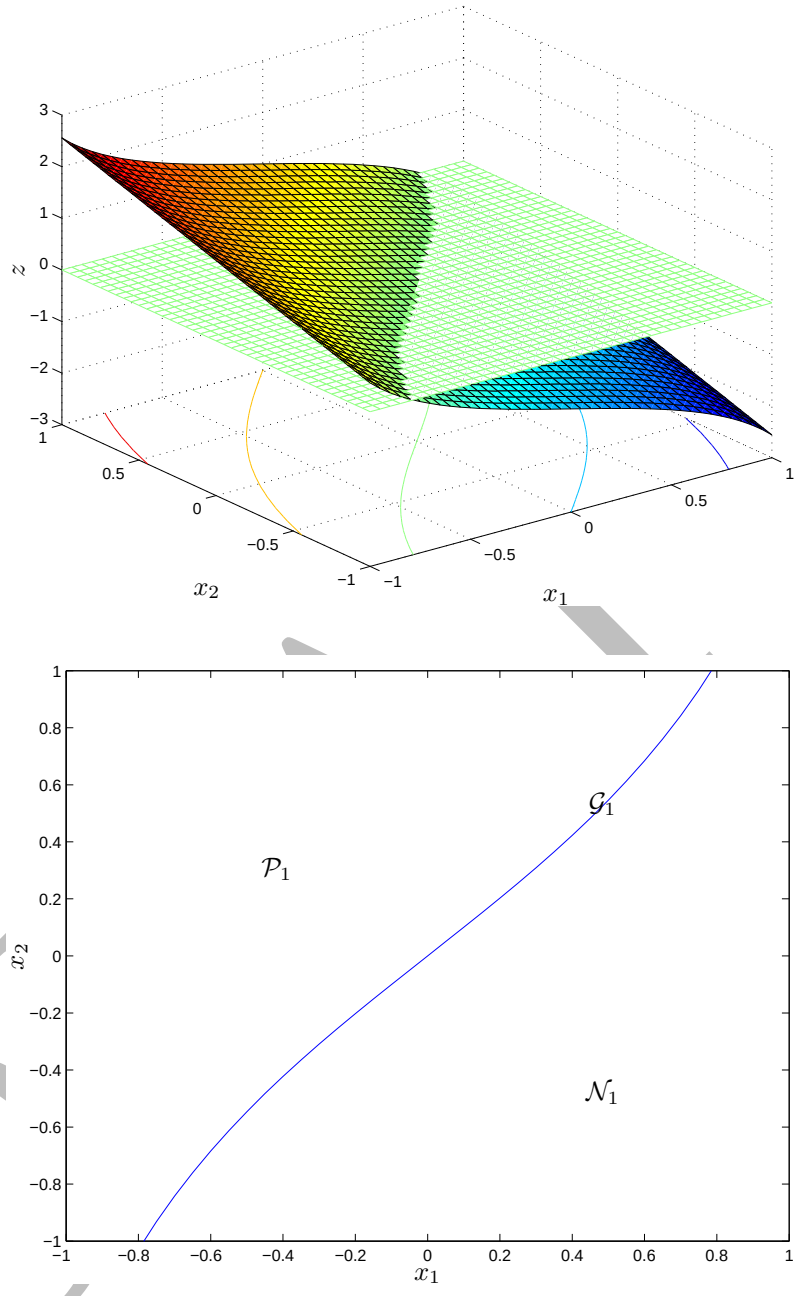

Figura 1.18: Na figura superior, é mostrada a superfície $z = g_1(\boldsymbol{x})$ com suas curvas de nível e sua interseção com o plano $z = 0$. Na figura inferior, é mostrado o plano x, onde se apresenta apenas a curva de nível $g_1(\boldsymbol{x}) = 0$. Nesse plano, a região $\mathcal{N}_1$ corresponde aos pontos em que a função $g_1(\cdot)$ é negativa; a região $\mathcal{P}_1$ corresponde aos pontos em que a função $g_1(\cdot)$ é positiva; e a fronteira que separa essas regiões, $\mathcal{G}_1$, corresponde aos pontos em que a função $g_1(\cdot)$ se anula.

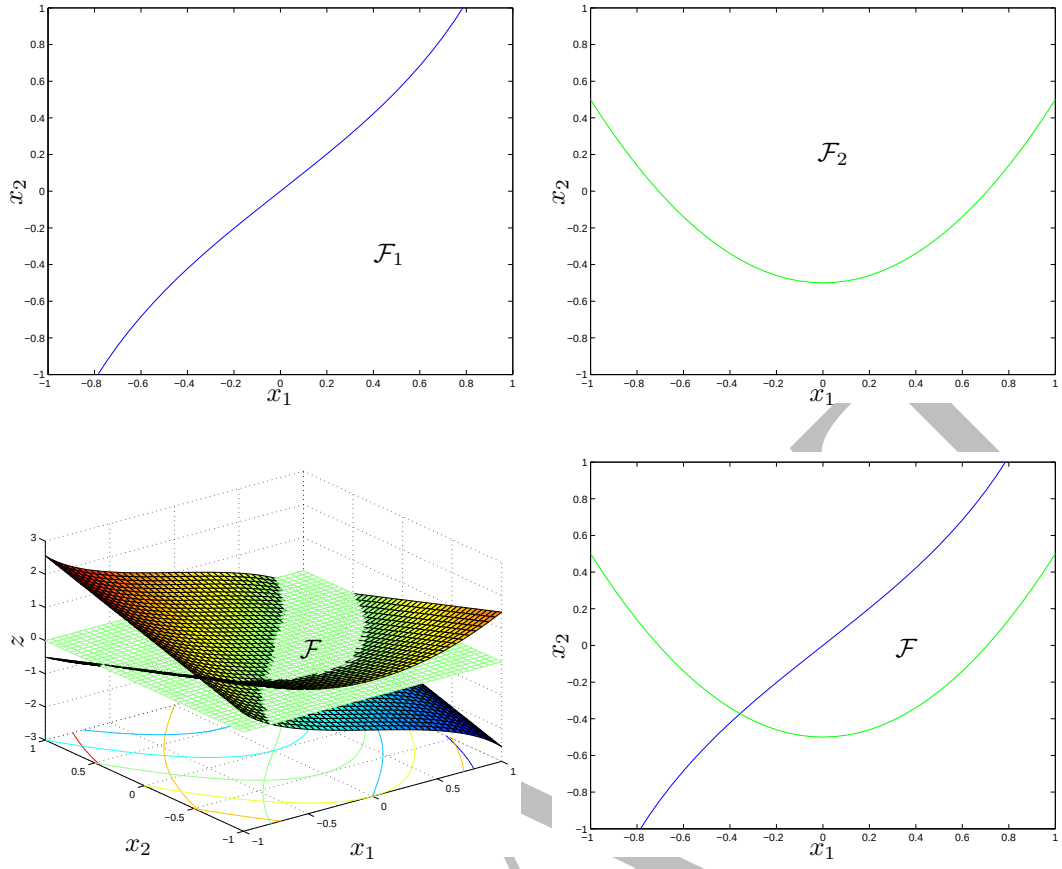

Figura 1.19: A região $\mathcal{F}_1$ corresponde aos pontos em que a função $g_1(\cdot)$ é negativa (Figura superior esquerda). A região $\mathcal{F}_2$ corresponde aos pontos em que a função $g_2(\cdot)$ é negativa (Figura superior direita). A interseção dessas duas regiões, $\mathcal{F}$, corresponde aos pontos em que ambas as funções são negativas, simultaneamente (Figura inferior direita). A Figura inferior esquerda mostra as superfícies $z = g_1(x)$, $z = g_2(x)$, assim como sua interseção com o plano $z = 0$ e suas curvas de nível. Pode-se observar também nesta Figura a região $\mathcal{F}$.

de restrições na forma em que o mesmo usualmente aparece: um conjunto de várias desigualdades que devem ser simultaneamente satisfeitas. Escrevendo novamente o sistema:

$$\begin{aligned} g_1(\boldsymbol{x}^*) &\leq 0 \\ g_2(\boldsymbol{x}^*) &\leq 0 \\ &\vdots \\ g_p(\boldsymbol{x}^*) &\leq 0 \end{aligned} \tag{1.15}$$

A Figura 1.19 mostra a situação para duas restrições: a região factível (ou seja, a região dos pontos que simultaneamente atendem às duas restrições) corresponde à interseção da região cujos pontos atendem à primeira restrição com a região cujos pontos atendem à segunda restrição. Em geral, se $\mathcal{F}_i$ designa a região em que a

função $g_i(\cdot)$ é menor ou igual a zero[12], temos que a *região factível* $\mathcal{F}$ do problema envolvendo todo o conjunto de restrições (1.15) corresponde à interseção de todas essas regiões:

$$\mathcal{F} = \mathcal{F}_1 \cap \mathcal{F}_2 \cap \ldots \cap \mathcal{F}_p \tag{1.16}$$

O *problema de otimização restrita com restrições de desigualdade*, em sua forma geral, trata da questão de determinação do ponto de mínimo $\boldsymbol{x}^*$ de uma função, dentro de uma *região factível* $\mathcal{F}$ definida dessa forma. Nas subseções que se seguem, mostraremos algumas formas do nosso Otimizador lidar com tal problema.

1.3.3 Barreiras e Penalidades

A primeira maneira de tentar adaptar os métodos de otimização, que foram formulados para problemas de otimização irrestrita, para o caso agora em análise, com restrições de desigualdade, é a técnica das *barreiras* e *penalidades*. A ideia é modificar a função-objetivo, acrescentando um termo que, dentro da região factível, afeta pouco a função, mas que nas proximidades da fronteira da região factível (no caso das *barreiras*) ou no exterior da região factível (no caso das *penalidades*) muda bastante a função, "impedindo" ou "penalizando" o Otimizador, ou seja, o algoritmo de otimização, de sair da região factível (método de barreiras) ou de permanecer na região inviável (método de penalidades).

Em termos matemáticos, o problema de otimização original, definido por:

$$\begin{aligned} \boldsymbol{x}^* &= \arg\min_{\boldsymbol{x}} f(\boldsymbol{x}) \\ \text{sujeito a: } & \{g_i(\boldsymbol{x}) \leq 0 \end{aligned} \tag{1.17}$$

é substituído pelo problema:

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x}) + F(\boldsymbol{x}) \tag{1.18}$$

A função $F(\cdot)$ deve ser muito pequena (ou zero) no interior da região factível, de tal forma que $f(\cdot)$ seja muito parecida com $f(\cdot) + F(\cdot)$ em qualquer ponto deste espaço.

No caso de métodos de *barreiras*, a função $F(\cdot)$ deve crescer muito rapidamente quando estamos perto da fronteira da região factível. A ideia é que o Otimizador, ao se aproximar dessa fronteira, verifique um súbito aumento da função $f(\boldsymbol{x})+F(\boldsymbol{x})$ (que é a função que ele está otimizando), de forma que ele não caminha em direção a essa fronteira. O Otimizador, se tiver iniciado a busca no interior da região factível, irá sempre ficar nessa região, portanto[13]. Esse tipo de método é denominado de *barreira* porque a função $F(\cdot)$ cria uma espécie de "barreira", que impede que o Otimizador atinja a fronteira da região factível. A Figura 1.20 ilustra uma função modificada com uma barreira, para uma situação de otimização em uma única variável.

[12]Observe que essa notação, utilizando o índice i, significa o mesmo que uma enumeração de todas as funções e regiões: $\mathcal{F}_1$ correspondendo à região em que $g_1(\cdot) \leq 0$, $\mathcal{F}_2$ correspondendo à região em que $g_2(\cdot) \leq 0$, e assim por diante.

[13]Deve-se tomar o cuidado, ao utilizar um método de barreira, para que o ponto inicial já esteja no interior da região factível.

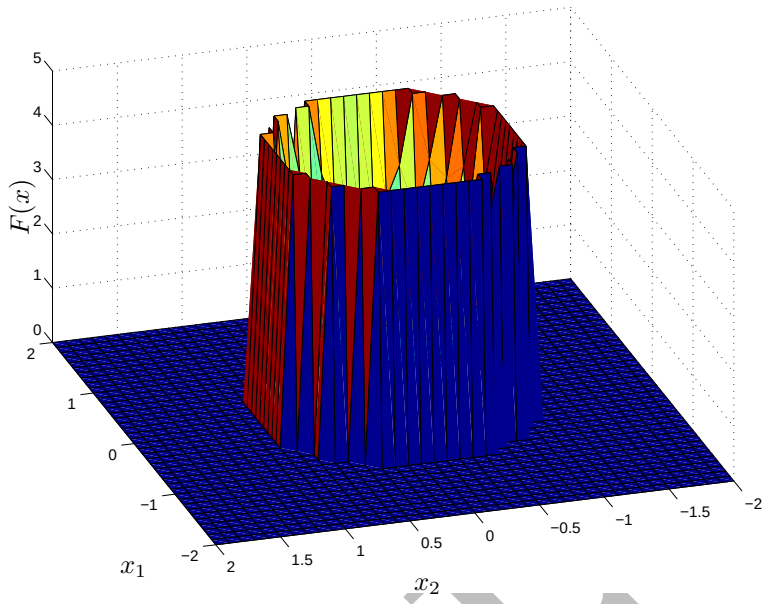

Figura 1.20: Ilustração de uma função de barreira, construída para garantir a restrição de que a otimização deva ocorrer no interior de um círculo de raio igual a 1, que seria a região factível de um problema de otimização. Essa função, somada à função objetivo, teria o papel de "impedir" a saída de um Otimizador do interior desse círculo de raio 1 que corresponde à região factível.

Os métodos de *penalidades*, por outro lado, são obtidos se se faz a função $F(\cdot)$ crescer rapidamente do lado de fora da região factível, para valores que aumentam à medida em que nos afastamos dessa região. A ideia, neste caso, é fazer com que o Otimizador, ao sair da região factível, encontre um crescimento da função $f(\boldsymbol{x})+F(\boldsymbol{x})$ que ele está otimizando, de forma que ele tende a voltar ao interior da região. Esse tipo de método é denominado de *penalidade* porque a função $F(\cdot)$ faz com que o Otimizador seja apenado (ou seja, sofra uma penalidade) caso ultrapasse a fronteira da região factível, sendo tanto maior a penalidade quanto mais o Otimizador se afastar dessa região. A Figura 1.21 ilustra uma função de penalidade.

A Figura 1.22 sobrepõe os gráficos das Figuras 1.20 e 1.21, que mostram uma função barreira e uma função penalidade para o tratamento da mesma restrição.

Deve-se notar que, uma vez que a função objetivo esteja modificada, seja por uma função de barreira, seja por uma de penalidade, a resultante função modificada pode ser otimizada utilizando os mesmos métodos que foram desenvolvidos para o caso da otimização sem restrições. Tipicamente, serão empregados métodos de direções de busca para resolver problemas formulados dessa maneira[14].

[14]Deve-se notar que, em particular, as funções de barreira não seriam funcionais se empregados nem junto com métodos de exclusão de regiões nem junto com métodos de populações. Já as funções de penalidade não causariam essas dificuldades, e poderiam ser empregadas com qualquer sistema de otimização. O leitor é convidado a explicar por quê isso ocorre.

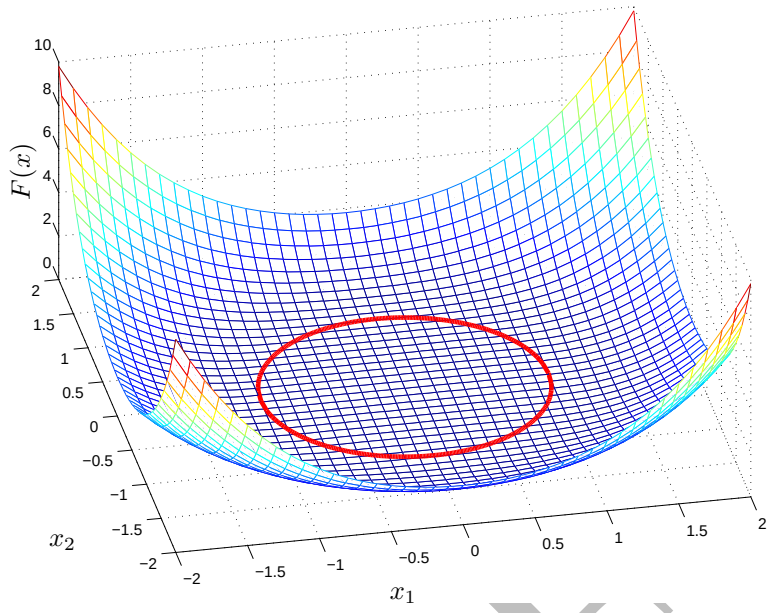

Figura 1.21: Ilustração de uma função de penalidade. A região factível corresponde ao interior do círculo indicado em vermelho. A função de penalidade é igual a zero no interior da região factível, e cresce rapidamente à medida em que o ponto se afasta dessa região.

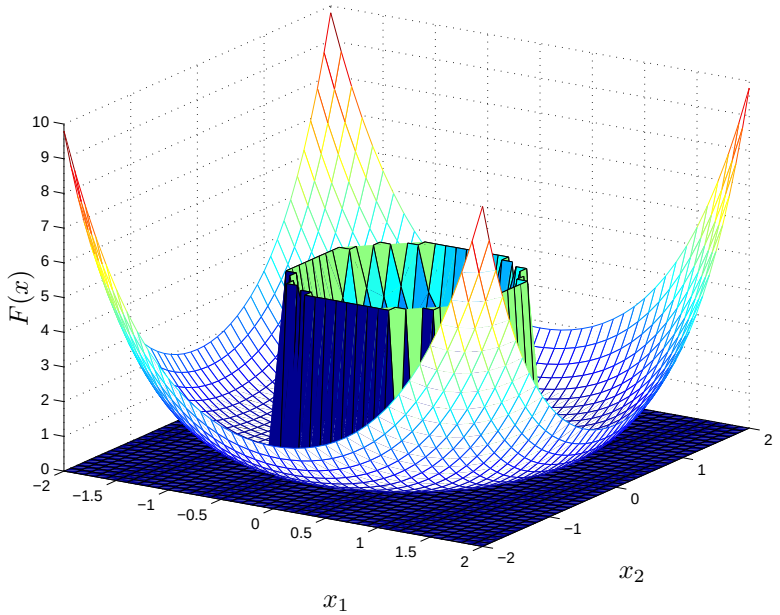

Figura 1.22: Sobreposição dos gráficos das figuras 1.20 e 1.21, de forma a mostrar uma função barreira e uma função penalidade para a mesma restrição. No caso, a restrição define como região factível o interior do círculo de raio 1 centrado na origem.

1.3.4 Composição pelo Máximo

Embora seja possível utilizar as funções de penalidade para lidar com as restrições de problemas de otimização nos casos em que o mecanismo de otimização a ser empregado é do tipo *exclusão de regiões*, há uma forma mais natural de tratar as restrições nesse caso. Considera-se, primeiro, a seguinte função:

$$G(\boldsymbol{x}) = \max(g_1(\boldsymbol{x}), g_2(\boldsymbol{x}), \ldots, g_p(\boldsymbol{x})) \tag{1.19}$$

A função $G(\cdot)$ é a chamada *composição pelo máximo* das funções $g_i(\cdot)$. O leitor é convidado a examinar a curva de nível $G(\boldsymbol{x}) = 0$. Essa curva de nível corresponde exatamente à fronteira da região factível do problema. Cada curva de nível $G(\boldsymbol{x}) = \alpha$, para $\alpha > 0$, corresponde a uma curva (ou hipersuperfície, em dimensões maiores que dois) fechada que é exterior às curvas correspondentes a valores menores de α, e todas têm em seu interior a região factível do problema (a curva correspondente a $\alpha = 0$)[15].

Imagine-se agora a aplicação de uma técnica de otimização por exclusão de regiões à função $G(\cdot)$. Se o Otimizador começar, nesse caso, em um ponto fora da região factível, a primeira exclusão será de um semi-espaço que garantidamente não contém a região factível, ficando para continuar a ser examinado o semi-espaço que contém a região factível. O processo continua até que, certamente, o Otimizador finalmente cai dentro da região factível.

Para fechar o procedimento a ser aplicado, uma vez dentro da região factível do problema, aplica-se um passo convencional de "exclusão de região", utilizando a função objetivo $f(\cdot)$ para determinar a exclusão. O significado desse passo é: após esse corte, o Otimizador permanece com o semi-espaço que contém a parcela da região factível na qual o ponto de ótimo do problema se encontra (ou seja, elimina-se a parcela da região factível em que o ponto de ótimo não se encontra). Essas operações são ilustradas na Figura 1.23.

O algoritmo resultante da sequência dessas operações pode oscilar, levando o Otimizador sucessivamente para dentro e para fora da região factível. No entanto, como no caso irrestrito, o volume da região considerada necessariamente diminui a cada passo, sendo que o ponto de ótimo permanece nessa região. O Otimizador, assim, termina arbitrariamente próximo do ótimo.

1.4 Otimização com Restrições de Igualdade

Consideremos agora o problema de otimização com *restrições de igualdade*:

$$\begin{aligned} \boldsymbol{x}^* = &\ \arg\ \min\ f(\boldsymbol{x}) \\ \text{sujeito a:}\ &\{h_j(\boldsymbol{x}) = 0, j = 1 \ldots, q \end{aligned} \tag{1.20}$$

[15] Para fazermos essa afirmativa, na verdade, estamos assumindo que as funções $g_i(\cdot)$ sejam todas *convexas* ou, pelo menos, *quasi-convexas*.

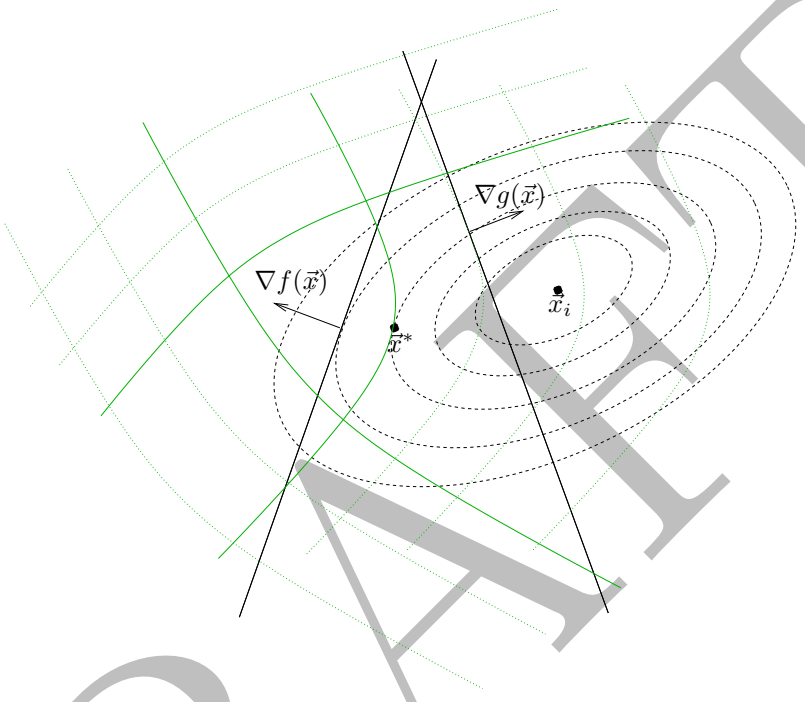

Figura 1.23: Ilustração da aplicação do processo de exclusão de região em um problema de otimização restrita. São mostradas, na figura, as curvas de nível da função objetivo $f(\boldsymbol{x})$, ao redor do mínimo irrestrito $\boldsymbol{x}_i$, e as curvas de nível das restrições $g_i(\boldsymbol{x})$. Estas são mostradas no exterior da região factível, sendo mostradas, em traço mais grosso, as curvas correspondentes a $g_i(\boldsymbol{x}) = 0$ (ou seja, as curvas que definem as fronteiras da região factível). O ponto de ótimo do problema é representado por $\boldsymbol{x}^*$. São mostrados os vetores gradientes da função objetivo, $\nabla f(\boldsymbol{x})$, em um ponto factível, e gradiente de uma restrição violada, $\nabla g(\boldsymbol{x})$, em um ponto infactível. Deve-se observar que as retas normais a ambos os vetores gradiente definem cortes do plano tais que o semi-plano oposto ao vetor gradiente, em ambos os casos, necessariamente contém a solução $\boldsymbol{x}^*$. (No caso do corte feito no ponto infactível, o semi-plano oposto ao gradiente contém de fato toda a região factível).

Essa descrição do problema significa, conforme já foi visto, que o ponto de ótimo $\boldsymbol{x}^*$ a ser determinado deve satisfazer às q equações:

$$\begin{aligned} h_1(\boldsymbol{x}^*) &= 0 \\ h_2(\boldsymbol{x}^*) &= 0 \\ &\vdots \\ h_q(\boldsymbol{x}^*) &= 0 \end{aligned} \tag{1.21}$$

Num espaço de n dimensões, cada uma dessas equações pode ser interpretada como uma descrição de um conjunto de pontos (os pontos $\boldsymbol{x}$ que a satisfazem) que fazem parte de uma superfície de dimensão $n-1$. Por exemplo, num espaço de dimensão 3, uma equação dessas significa uma superfície no sentido convencional, dotada de duas dimensões (algo como uma "folha" curvada). Essa superfície corresponde ao conjunto dos pontos factíveis do problema de otimização, se ele envolver apenas uma restrição de igualdade. No caso de q restrições de igualdade, o conjunto factível corresponde à interseção de todas as superfícies (cada uma associada a uma das restrições de igualdade).

O espaço que estamos considerando, na série de exemplos que vem sendo apresentada neste capítulo, possui apenas duas dimensões. Assim, o lugar geométrico definido por uma equação do tipo:

$$h_1(\boldsymbol{x}) = 0 \tag{1.22}$$

corresponde a um objeto de dimensão um, ou seja, uma linha (possivelmente curva). Este será o conjunto factível de um problema de otimização que tiver (1.22) como restrição. A Figura 1.24 mostra um exemplo dessa situação.

Das técnicas mostradas anteriormente para tratar de problemas de otimização com restrições de desigualdade, duas simplesmente não funcionam para o caso de restrições de igualdade: o método de barreiras e o método de composição pelo máximo. A razão disso é que ambas as técnicas dependem da existência de pontos que sejam *interiores* à região factível do problema para funcionarem, e as regiões factíveis de restrições de igualdade não possuem pontos interiores[16]. A técnica de penalidades, por sua vez, pode ser empregada.

1.5 Otimização Linear

Um caso especial particularmente importante do problema de otimização ocorre quando tanto a função objetivo quanto as funções de restrição são lineares[17]. Esse

[16]Pontos interiores a uma região são pontos que pertencem a essa região e não estão em sua fronteira. Claramente, todos os pontos factíveis de problemas de otimização com restrições de igualdade estão na fronteira da região factível, isto é, possuem algum ponto vizinho fora dessa região.

[17]No caso das restrições, uma terminologia mais precisa iria dizer que são *afins* e não *lineares*.

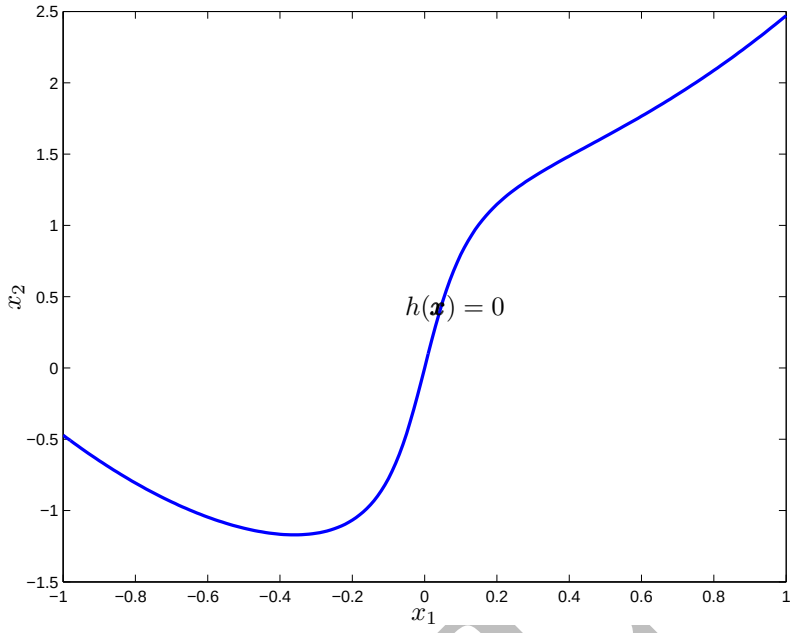

Figura 1.24: A linha corresponde ao lugar geométrico dos pontos que satisfazem $h(\boldsymbol{x}) = 0$. Essa linha é a região factível de um problema de otimização com essa restrição.

é o chamado problema de *otimização linear*:

$$\begin{aligned} \boldsymbol{x}^* &= \arg\min \boldsymbol{c}'\boldsymbol{x} \\ \text{sujeito a: } &\begin{cases} A\boldsymbol{x} \leq \boldsymbol{b} \end{cases} \end{aligned} \tag{1.23}$$

sendo $\boldsymbol{c}$ um vetor de dimensão n (mesmo tamanho que $\boldsymbol{x}$), A uma matriz $\mathbb{R}^{m \times n}$ e $\boldsymbol{b}$ um vetor de dimensão m. Claramente, a função objetivo desse problema é a função linear:

$$f(\boldsymbol{x}) = c_1 x_1 + c_2 x_2 + \ldots + c_n x_n \tag{1.24}$$

e o conjunto de restrições corresponde às m desigualdades:

$$\begin{aligned} &a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \leq b_1 \\ &a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \leq b_2 \\ &\vdots \\ &a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n \leq b_m \end{aligned} \tag{1.25}$$

A otimização linear é particularmente importante por duas razões: Primeiro, um número muito grande de situações práticas é modelado pela formulação linear. Segundo, devido à sua estrutura peculiar, problemas de otimização linear podem ser resolvidos muito mais rapidamente que problemas de otimização não-linear com o mesmo número de variáveis e o mesmo número de restrições. Assim, algoritmos especializados para resolver apenas problemas lineares são capazes de lidar com problemas muito grandes (muito maiores que aqueles que poderiam ser resolvidos no caso não-linear geral).

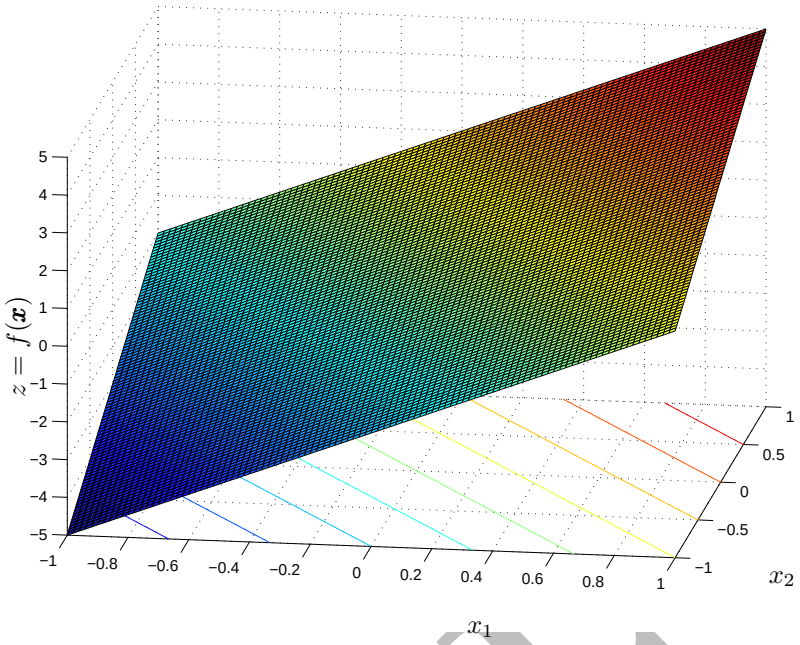

Figura 1.25: Superfície correspondente à função objetivo linear $f(\boldsymbol{x}) = \boldsymbol{c}'\boldsymbol{x}$. Na figura estão representadas também as curvas de nível da função, que são retas paralelas.

Vamos examinar essa estrutura peculiar que torna tão favorável a otimização linear. No caso de duas variáveis de otimização, a superfície representativa da função linear é simplesmente um plano, e suas curvas de nível são retas paralelas. Isso é mostrado na Figura 1.25.

O problema de otimização de uma função linear não faz sentido se não estiver acompanhado de restrições, pois o ponto que minimiza tal função objetivo encontra-se no infinito[18]. Examinemos o que são as restrições do problema de otimização linear. Num espaço de n dimensões, a desigualdade:

$$a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \leq b_1 \tag{1.26}$$

representa um semi-espaço. A fronteira que separa a região factível da infactível corresponde a um hiperplano nesse espaço. No caso de duas dimensões, a desigualdade:

$$a_{11}x_1 + a_{12}x_2 \leq b_1 \tag{1.27}$$

define um semi-plano como região factível, e a fronteira dessa região factível corresponde à reta $a_{11}x_1 + a_{12}x_2 = b_1$. Consideremos agora várias restrições de desigualdade em duas dimensões:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &\leq b_1 \\ a_{21}x_1 + a_{22}x_2 &\leq b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 &\leq b_m \end{aligned} \tag{1.28}$$

Como cada uma dessas restrições de desigualdade define um semi-plano, as várias restrições de desigualdade correspondem à interseção de vários semi-planos, o que define um poliedro. Isso é mostrado na Figura 1.26.

[18]Em outras palavras, não existe nenhum mínimo local irrestrito de uma função objetivo linear.

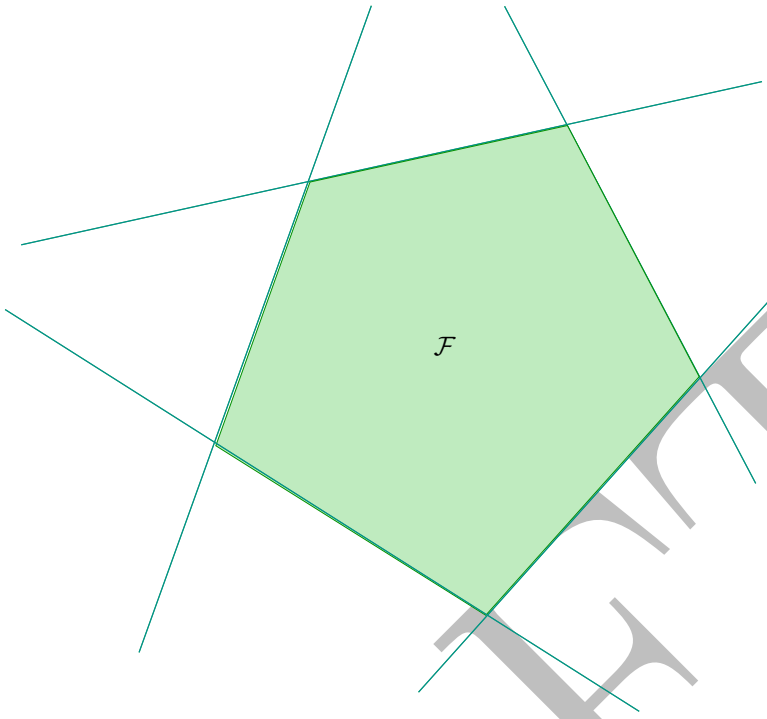

Figura 1.26: Região factível $\mathcal{F}$ correspondente a várias restrições lineares de desigualdade. Cada reta que contém um dos lados do poliedro factível corresponde à fronteira de uma restrição de desigualdade.

Observemos agora, na Figura 1.27, a superposição das curvas de nível de uma função objetivo linear com uma região factível linear. O dado relevante a ser observado é que, num problema linear, o ponto de ótimo *necessariamente se encontra sobre um vértice do poliedro factível.*

O leitor deve se convencer de que seria impossível, num problema linear, que o mínimo da função objetivo estivesse no interior da região factível. Seria também impossível que esse mínimo estivesse em um ponto da fronteira da região factível sem estar em um dos vértices dessa fronteira[19]. Assim, uma possível estratégia para resolver problemas lineares seria fazer o Otimizador percorrer apenas o conjunto dos vértices da região factível, escolhendo dentre esses vértices aquele com menor valor de função objetivo. É possível implementar métodos bastante eficientes de otimização com base em tal estratégia: esses são os chamados métodos Simplex. Esse tipo de lógica, largamente empregada no contexto da otimização linear, é fundamentalmente diferente dos procedimentos que podem ser utilizados na otimização não-linear[20].

[19]No entanto, seria possível que houvesse múltiplos mínimos, incluindo pontos diversos da fronteira, dentre esses necessariamente pelo menos um dos vértices.

[20]Devemos entretanto informar o leitor que, recentemente, outras estratégias de otimização linear, denominadas *métodos de pontos interiores*, vêm ganhando a preferência dos usuários, estratégias essas que têm semelhança com métodos de otimização não-linear.

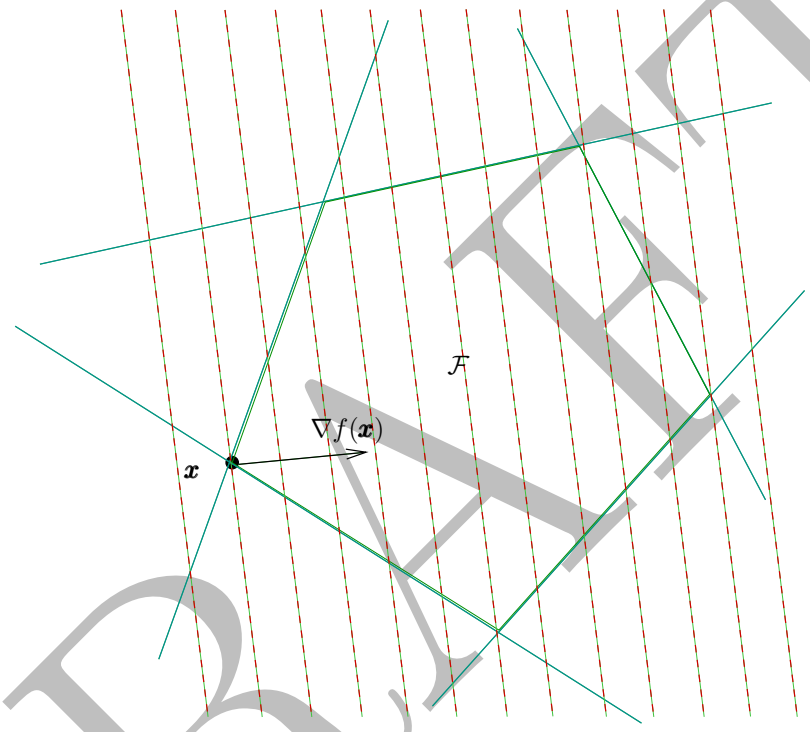

Figura 1.27: O vetor gradiente da função objetivo, $\nabla f(\boldsymbol{x})$, mostrado no ponto $\boldsymbol{x}$, é constante em todo o espaço, pois a função objetivo é linear. As linhas tracejadas correspondem às curvas de nível da função objetivo, sendo que elas correspondem a valores cada vez menores de função objetivo quando se caminha da direita para a esquerda. Dessa forma, o ponto $\boldsymbol{x}$ indicado na figura é o de menor valor de função objetivo dentro da região factível $\mathcal{F}$, correspondendo ao ponto em que a curva de nível de menor valor toca a região factível.

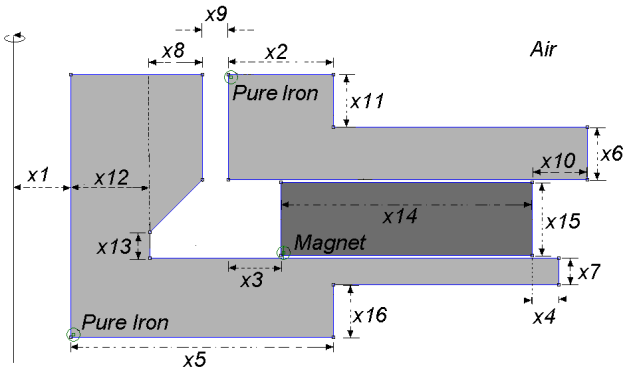

Figura 1.28: Ilustração do auto-falante em 2D com indicação das regiões de "Ferro", "Ímã" e das variáveis x de projeto.

1.6 Estudos de Casos

1.6.1 O projeto de um auto-falante

Nesta subseção discutiremos o projeto de um auto-falante.

Descrição do problema

O modelo, representado na Figura1.28, consiste de três materiais distintos: *Ar*, *Ferro* e *Ímã*. As propriedades físicas de cada um destes materiais são dadas na Tabela 1.1. As curvas de magnetização B-H do ferro e do ímã foram obtidas através da interpolação quadrática de pontos amostrados experimentalmente, conforme ilustrado nas Figuras 1.29 e 1.30. Os pontos utilizados para gerar estas interpolações foram obtidos na biblioteca de materiais do software de análise numérica *FEMM 4.2* (*Finite Element Method Magnetics*), utilizado na construção deste modelo.

Tabela 1.1: Materiais utilizados no modelo do auto-falante.

Denominação	Ar	Ferro	Ímã
Material	Air	Pure Iron	Ceramic 5 magnet
μ_r	1,0	*	*
H_c [A/m]	0,0	0,0	191262
σ [MS/m]	0,0	10,44	0,0

Definição do problema de otimização

O objetivo neste problema consiste na minimização do volume total de material utilizado na construção do auto-falante. Este objetivo é restrito pelo requisito de

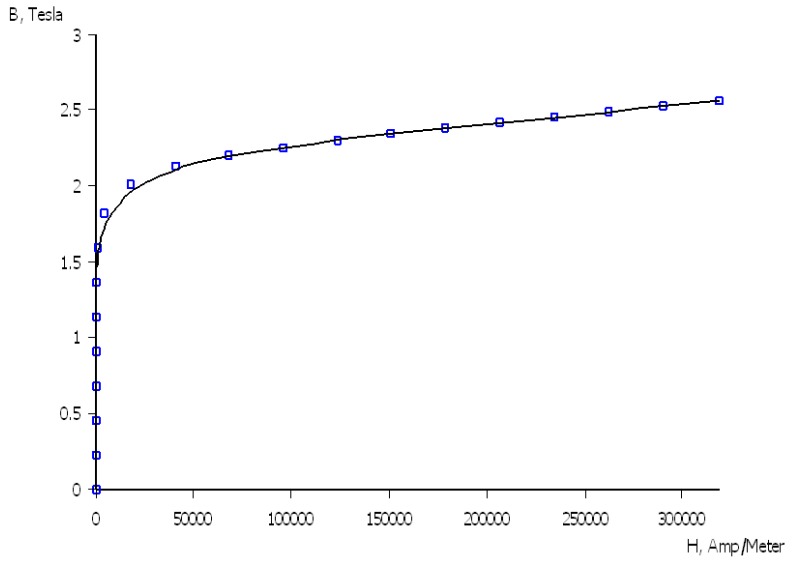

Figura 1.29: Curva de magnetização utilizada para a modelagem do núcleo de ferro.

um valor mínimo da densidade de fluxo magnético na região definida pela variável x_9. Matematicamente, o problema pode é descrito por (1.29):

$$\begin{aligned} &\min f(\boldsymbol{x}) = \text{volume} \\ &\text{sujeito a: } g_1(\boldsymbol{x}) : |\mathbf{B}| \geq \mathbf{B}_{min} \end{aligned} \tag{1.29}$$

com $\mathbf{B}_{min} = 0,5\ T$ e o *volume* representando a soma total do volume das partes do alto-falante.

Os limites recomendados para as variáveis de otimização são dados na Tabela 1.2. Esta tabela também fornece sugestões de valores fixos, a serem utilizados em casos de otimização parcial do modelo ou como ponto de partida para o teste de algoritmos determinísticos.

O cálculo da densidade de fluxo magnético B

O auto-falante descrito nas seções anteriores foi modelado na forma de um script LUA (www.lua.org), que por sua vez é interpretado pelo pacote de elementos finitos FEMM 4.2 (www.femm.info). A implementação atual é capaz de realizar simulações em batelada, retornando um arquivo de saída contendo os valores de densidade de fluxo magnético e volume do dispositivo. Este pacote á capaz ainda de gerar facilmente a visualização de linhas de campo e mapas de densidade de fluxo magnético.

Instruções de Uso

1. Software necessário:
   - Finite Element Method Magnetics v.4.2
   - Matlab
2. Arquivos necessários (www.cpdee.ufmg.br/∼ fcampelo/files/loudspeaker/):

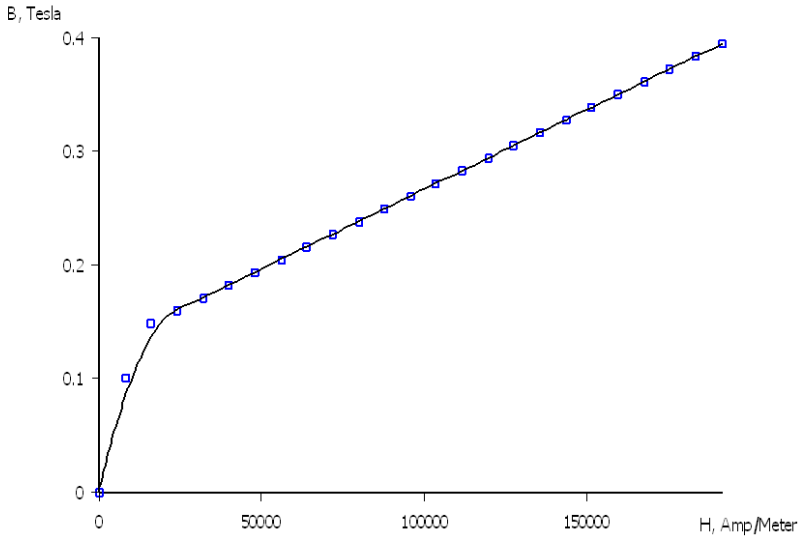

Figura 1.30: Curva de magnetização utilizada para a modelagem do ímã de cerâmica.

- loudspeaker.lua
- CallFEMM_LS.m
- LS_fun.m

3. Opções de problema:

   - Otimização completa (16 variáveis)
   - Otimização parcial (7 variáveis)

4. Forma de utilização:

   - Copie todos os arquivos contidos em (/∼ fcampelo/files/loudspeaker/) para um diretório local (p.ex., "C:\loudspeaker\" – este caminho não deve conter espaços em branco).
   - Nas linhas 33–35 do arquivo *loudspeaker.lua*, insira o caminho escolhido.
   - Nas linhas 5–8 do arquivo *CallFEMM_LS.m*, insira o caminho escolhido.

Para testar se os diretórios estão corretos, proceda da seguinte forma:

1. LUA script:

   - Abra o FEMM 4.2;
   - Selecione *File - Open LUA Script - loudspeaker.lua*
   - Caso o arquivo *loudspeaker.lua* esteja correto, o FEMM deve executar uma simulação de teste (definida pelo arquivo *loudspeaker.in* contido em /∼ fcampelo/files/loudspeaker/) e fechar automaticamente.

2. Rotina Matlab:

Tabela 1.2: Limites do espaço de busca.

Variável	min (mm)	max (mm)	fixo (mm)
x_1	3.0	12.0	5.0
x_2	1.0	4.0	3.0
x_3	1.0	4.0	2.0
x_4	0.0	3.0	1.5
x_5	5.0	15.0	7.0
x_6	2.0	5.0	4.0
x_7	1.0	10.0	2.0
x_8	1.0	3.0	2.0
x_9	0.5	2.0	1.0
x_{10}	0.0	3.0	1.0
x_{11}	1.0	5.0	2.0
x_{12}	2.0	5.0	2.0
x_{13}	0.0	2.0	1.0
x_{14}	5.0	12.0	7.0
x_{15}	2.0	5.0	4.0
x_{16}	1.0	5.0	2.0

- Abra o Matlab e selecione o diretório contendo os arquivos do auto-falante;
- Na janela de comando, digite:
  >> X = [5.0,3.0,1.0,0.0,7.0,6.0,2.0,5.0,0.5,...
  0.0,1.0,0.5,1.0,7.0,4.0,1.0]';
  >> Y = CallFEMM_LS(X)
- Caso o arquivo *CallFEMM_LS.m* esteja correto, o Matlab invocará uma janela do FEMM, que executará uma simulação de teste e retornará o foco para o Matlab.

Além da função *LS_fun.m*, há também as funções *LS_vol.m* e *LS_B.m*, capazes de retornar as componentes de volume e de campo separadamente. As rotinas do Matlab são extensivamente comentadas e facilmente adaptáveis para uma ampla gama de algoritmos de otimização.

Resultados

A Tabela 1.3 mostra os resultados obtidos para o problema com 7 variáveis, que indica um auto-falante com volume total $V = 15.4696\ cm^3$ e densidade de fluxo $\mathbf{B} = 0.4953\ T$.

A Figura 1.31 ilustra uma configuração possível para o problema do auto-falante.

Tabela 1.3: Resultado para o problema com 7 variáveis.

Variável	Resultado (mm)	Fixo (mm)
x_1	–	5.0
x_2	3.5089	–
x_3	–	2.0
x_4	–	1.5
x_5	–	7.0
x_6	2.0053	–
x_7	–	2.0
x_8	–	2.0
x_9	–	1.0
x_{10}	1.1941	–
x_{11}	1.0000	–
x_{12}	–	2.0
x_{13}	–	1.0
x_{14}	11.9946	–
x_{15}	5.0000	–
x_{16}	–	2.0

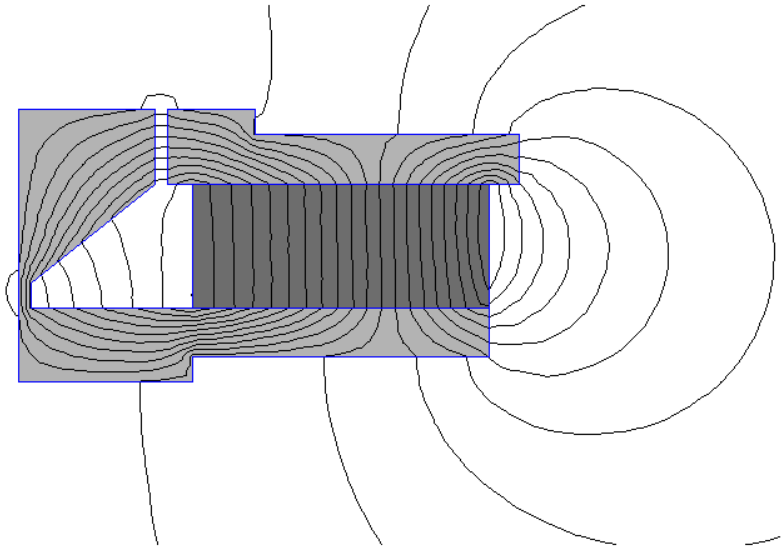

Figura 1.31: Resultado de uma configuração possível do alto-falante com ilustração das linhas equipotenciais de **B**

Referências Bibliográficas

[1] D. G. Luenberger. *Linear and Nonlinear Programming.* Addison-Wesley, 2 edition, 1989.

[2] P. Venkataraman. *Applied Optimization with Matlab Programming.* John Wiley, 1 edition, 2002.