Revisão: Álgebra linear, probabilidade, estatística e otimização

Álgebra linear

- Sistemas lineares
- Autovalores e autovetores
- Matrizes positivas definidas
- Decomposições: espectral, SVD, QR

Sistemas lineares

- A equação $\mathbf{Ax} = \mathbf{b}$, com $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{x} \in \mathbb{R}^{n \times 1}$ e $\mathbf{b} \in \mathbb{R}^{m \times 1}$, encapsula o sistema linear

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1n}x_n &= b_1 \\ A_{21}x_1 + A_{22}x_2 + \cdots + A_{2n}x_n &= b_2 \\ \vdots \quad + \quad \vdots \quad + \ddots + \quad \vdots \quad &= \quad \vdots \\ A_{m1}x_1 + A_{m2}x_2 + \cdots + A_{mn}x_n &= b_m \end{aligned}$$

- Trabalhar com vetores e matrizes é a base do Numpy, e simplifica demais a notação
- O sistema pode ter infinitas soluções, nenhuma solução ou exatamente uma solução
- Quando $m = n$ e as colunas são linearmente independentes, existe inversa $\mathbf{A}^{-1}$ e solução única
- Mas resolver $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ é numericamente instável e lento: decomposições ajudam

Autovalores e autovetores

- Um autovetor de uma matriz $\mathbf{A} \in \mathbb{R}^{n \times n}$ é um vetor $\mathbf{v} \neq 0$ tal que existe $\lambda \in \mathbb{R}$ para o qual

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}.$$

  O valor λ é o autovalor associado ao autovetor $\mathbf{v}$

- Se $\mathbf{v}$ é autovetor, então $c\mathbf{v}$ também é; assumiremos normalização: $\|\mathbf{v}\|_2 = (\sum_{i=1}^{n} v_i^2)^{1/2} = 1$

- Se $\mathbf{A}$ tem n autovetores ortogonais $\mathbf{v}_1, \ldots, \mathbf{v}_n$ associados a $\lambda_1, \ldots, \lambda_n$, então

$$\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i \Longrightarrow \mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T = \begin{bmatrix} | & | & | \\ \mathbf{v}_1 & \cdots & \mathbf{v}_n \\ | & | & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \begin{bmatrix} - & \mathbf{v}_1^T & - \\ & \vdots & \\ - & \mathbf{v}_n^T & - \end{bmatrix}$$

- Essa é a decomposição espectral; note que $\mathbf{Q}$ é ortogonal ($\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$) e $\mathbf{\Lambda}$ é diagonal

- Em geral, vamos assumir que os autovalores estão dispostos em ordem decrescente

- Toda matriz real e simétrica ($\mathbf{A} = \mathbf{A}^T$) admite essa decomposição

Matrizes positivas definidas

- Uma matriz real e simétrica com todos seus autovalores sendo positivos é dita positiva definida
- Matrizes positiva definidas são equivalentes a

$$\mathbf{x}^T\mathbf{A}\mathbf{x} > 0 \quad \text{para todo } \mathbf{x} \neq 0$$

- Isso segue da decomposição espectral:

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T\mathbf{x} = (\mathbf{\Lambda}^{1/2}\mathbf{Q}^T\mathbf{x})^T(\mathbf{\Lambda}^{1/2}\mathbf{Q}^T\mathbf{x}) = \mathbf{w}^T\mathbf{w} = \sum_{i=1}^{n} w_i^2 > 0$$

- Se autovalores são não-negativos, a matriz é positiva semi-definida e desigualdades não são estritas
- Toda matriz $\mathbf{A}^T\mathbf{A}$ é positiva semi-definida (pois $\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = \|\mathbf{A}\mathbf{x}\|^2 \geq 0$)

Decomposições: SVD

- Toda matriz real $A \in \mathbb{R}^{m \times n}$ admite uma decomposição SVD:

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T,$$

onde $\mathbf{U} \in \mathbb{R}^{m \times m}$ e $\mathbf{V} \in \mathbb{R}^{n \times n}$ são matrizes ortogonais, e $\mathbf{D} \in \mathbb{R}^{m \times n}$ é uma matriz diagonal

- Os elementos da diagonal de $\mathbf{D}$, chamados de valores singulares e usualmente denotados por $\sigma_1, \ldots, \sigma_n$, são sempre não-negativos (e assumidos decrescentes)

- As colunas de $\mathbf{U}$ são os vetores singulares esquerdos e as de $\mathbf{V}$ os vetores singulares direitos

- A decomposição SVD diz muito sobre a matriz $\mathbf{A}$
  - Vale que $\text{posto}(\mathbf{A}) = \#\{i : \sigma_i > 0\}$
  - Se $r = \text{posto}(\mathbf{A})$, então $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$
  - $\mathbf{u}_1, \ldots, \mathbf{u}_r$ é uma base ortonormal para espaço coluna de $\mathbf{A}$
  - $\mathbf{v}_{r+1}, .., \mathbf{v}_n$ é uma base ortonormal para espaço nulo de $\mathbf{A}$

Decomposições: SVD

- A decomposição SVD é consequência da decomposição espectral: seja $\mathbf{v}_i$ autovetor de $\mathbf{A}^T\mathbf{A}$:

$$\mathbf{A}^T\mathbf{A}\mathbf{v}_i = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T\mathbf{v}_i = \sum_{j=1}^{n} \lambda_i \mathbf{v}_j \mathbf{v}_j^T \mathbf{v}_i = \lambda_i \mathbf{v}_i$$

- Supondo $\lambda_i > 0$ (depois lidamos com $\lambda_i = 0$), defina

$$\mathbf{u}_i = \frac{\mathbf{A}\mathbf{v}_i}{\sqrt{\lambda_i}} \implies \mathbf{U} = \mathbf{A}\mathbf{V}\mathbf{D}^{-1} \implies \mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

- Por construção, $\mathbf{D} = \text{diag}(\sqrt{\lambda_i})$ é diagonal, e $\mathbf{V} = \mathbf{Q}$ é ortogonal

- A matriz $\mathbf{U}$ também é ortogonal pois são os autovetores de $\mathbf{A}\mathbf{A}^T$:

$$\mathbf{A}\mathbf{A}^T\mathbf{u}_i = \lambda_i^{-1/2}\mathbf{A}\mathbf{A}^T\mathbf{A}\mathbf{v}_i = \sqrt{\lambda_i}\mathbf{A}\mathbf{v}_i = \lambda_i\mathbf{u}_i$$
$$\mathbf{u}_i^T\mathbf{u}_i = \lambda_i^{-1}\mathbf{v}_i^T\mathbf{A}^T\mathbf{A}\mathbf{v}_i = \mathbf{v}_i^T\mathbf{v}_i = 1$$

- Se $\lambda_i = 0$, basta ignorá-lo na construção de $\mathbf{U}$ e $\mathbf{V}$ e depois completar cada base

Decomposições: QR

- Toda matriz $\mathbf{A} \in \mathbb{R}^{m \times n}$ admite uma decomposição QR:

$$\mathbf{A} = \mathbf{QR},$$

onde $\mathbf{Q}$ é ortogonal e $\mathbf{R}$ é triangular superior

- Intuição: ortogonalização das colunas de $\mathbf{A}$ (via Gram-Schmidt):
  - $\tilde{\mathbf{q}}_1 = \mathbf{a}_1, \mathbf{q}_1 = \tilde{\mathbf{q}}_1 / \|\tilde{\mathbf{q}}_1\|$
  - $\tilde{\mathbf{q}}_2 = \mathbf{a}_2 - (\mathbf{q}_1^T \mathbf{a}_2)\mathbf{q}_1, \mathbf{q}_2 = \tilde{\mathbf{q}}_2 / \|\tilde{\mathbf{q}}_2\|$
  - $\tilde{\mathbf{q}}_3 = \mathbf{a}_3 - (\mathbf{q}_1^T \mathbf{a}_3)\mathbf{q}_1 - (\mathbf{q}_2^T \mathbf{a}_3)\mathbf{q}_2, \mathbf{q}_3 = \tilde{\mathbf{q}}_3 / \|\tilde{\mathbf{q}}_3\|$
  - $\cdots$

- Chamando $r_{ki} = \mathbf{q}_k^T \mathbf{a}_i$, obtemos

$$\begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{bmatrix} = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_n \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{nn} \end{bmatrix}$$

Probabilidade e estatística

- Variáveis aleatórias, médias e variância
- Distribuições conjuntas e condicionais
- Distribuições importantes
- Estimação: método da máxima verossimilhança
- Inferência: teste de hipótese

Variáveis aleatórias

- A nossa modelagem sobre o mundo pressupõe incertezas (*e.g.,* medidas, amostras finitas)
- Uma variável aleatória X é uma função resultado de algum processo aleatório
- Por isso, ela tem uma função de distribuição associada, $F(x) = \mathbb{P}[X \leq x]$
- Se a variável for contínua, tem densidade $f(x) = F'(x)$; se for discreta tem massa $p(x) = \mathbb{P}[X = x]$
- Frequentemente, estamos interessados em dois números associados à variável aleatória
  - Média: $\mu = \mathbb{E}[X] = \int xf(x)dx \quad (= \sum xp(x))$
  - Variância: $\sigma^2 = \mathbb{E}[(X - \mathbb{E}[X])^2] = \int (x-\mu)^2 f(x)dx \quad (= \sum (x-\mu)^2 p(x))$
- Propriedades: $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$, e $\mathbb{V}[aX + b] = a^2\mathbb{V}[X]$
- Um vetor aleatório é uma coleção de variáveis aleatórias $\mathbf{X} = (X_1, X_2, \ldots, X_n)$

Distribuições conjuntas e condicionais

- Duas variáveis aleatórias têm distribuições conjunta $F(x, y) = \mathbb{P}[X \leq x, Y \leq y]$
- Elas podem ter densidade $f(x, y) = \partial^2 F(x, y)/\partial x \partial y$ ou massa $p(x, y) = \mathbb{P}[X = x, Y = y]$
- Além de média e variância, elas podem ter covariância: $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$
  - Se $\mathbf{X} = (X_1, X_2)$, e $\mathrm{Cov}(\mathbf{X}) = \mathbf{\Sigma}$, então $\mathrm{Cov}(\mathbf{AX}) = \mathbf{A}\,\mathrm{Cov}(\mathbf{X})\mathbf{A}^T = \mathbf{A\Sigma A}^T$
- Se as variáveis são independentes, $f(x, y) = f_X(x) f_Y(y)$ ou $\mathbb{P}[X = x, Y = y] = \mathbb{P}[X = x]\mathbb{P}[Y = y]$
  - Se X e Y são independentes, $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])]\mathbb{E}[(Y - \mathbb{E}[Y])] = 0$
- Podemos falar na distribuição condicional: $f_{Y|X}(y|x) = f(x, y)/f(x)$ ou $p_{Y|X}(y|x) = p(x, y)/p(x)$
- Esperança condicional: $\mathbb{E}[Y|X] = \int y f_{y|x} dy$ ou $\mathbb{E}[Y|X] = \sum y p(y|x)$
- Lei da esperança total: $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$
  - Segue da lei da probabilidade total: $\mathbb{P}[A] = \mathbb{P}[A|B]\mathbb{P}[B] + \mathbb{P}[A|B^c]\mathbb{P}[B^c]$

Distribuições importantes: discretas

- Bernoulli: $X \sim \text{Bern}(p)$ se $X \in \{0, 1\}$ e $\mathbb{P}[X = x] = p^x(1-p)^{1-x}$
  - Exemplo: jogar de uma moeda com probabilidade p de dar caras
  - $\mathbb{E}[X] = p$, $\mathbb{V}[X] = p(1-p)$

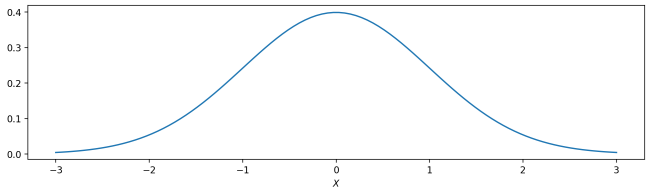
- Binomial: $X \sim \text{Bin}(n, p)$ se $X \in \{0, \ldots, n\}$ e $\mathbb{P}[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$
  - Exemplo: $X = \sum_{i=1}^{n} X_i$ com $X_i \overset{\text{iid}}{\sim} \text{Bern}(p)$, número de jogos ganhos em n jogos
  - $\mathbb{E}[X] = np$, $\mathbb{V}[X] = np(1-p)$

- Multinomial: $\mathbf{X} \sim \text{Mult}(p_1, p_2, \ldots, p_k)$ se $\mathbf{X} = (X_1, \ldots, X_k)$ com $X_i \in \{0, \ldots, n\}$ e $\sum_{i=1}^{n} X_i = n$; a função de distribuição é $\mathbb{P}[\mathbf{X} = (x_1, \ldots, x_k)] = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$
  - Exemplo: número de jogos com $0, 1, 2, \ldots, k$ gols em n jogos
  - $\mathbb{E}[\mathbf{X}] = n\mathbf{p}$, $\mathbb{V}[\mathbf{X}] = n(\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T)$

Distribuições importantes: contínuas

- Normal: $X \sim N(\mu, \sigma^2)$ se $X \in (-\infty, \infty)$ e

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

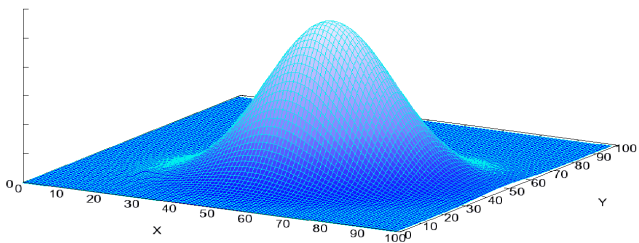


- Exemplo: alturas, mensurações com erros
- Por definição, $\mathbb{E}[X] = \mu$ e $\mathbb{V}[X] = \sigma^2$
- CLT: se $Y_1, \ldots, Y_n$ são iid com média μ_Y e variância σ_Y^2, então $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \to N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$

Distribuições importantes: contínuas

- Normal multivariada: se $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ com $\mathbf{X} \in \mathbb{R}^d$, $\boldsymbol{\mu} \in \mathbb{R}^d$ e $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$, e

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$



- Generalização natural da Normal para mais dimensões: altura e peso, por exemplo
- Por definição, $\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}$ e $\mathbb{V}[\mathbf{X}] = \boldsymbol{\Sigma}$
- A matriz $\boldsymbol{\Sigma}$ é simétrica ($\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$) e positiva semidefinida ($\mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T]$)

Estimação: máxima verossimilhança

- Suponha que $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, mas μ e σ^2 são desconhecidos. Como estimá-los?

- Método da máxima verossimilhança: a probabilidade de observar os dados $x_1, \ldots, x_n$ é

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$$

- Vamos encontrar os valores de μ e σ^2 que maximizam a probabilidade de observar a amostra:

$$(\hat{\mu}, \hat{\sigma}^2) = \underset{\tilde{\mu}, \tilde{\sigma}^2}{\operatorname{argmax}} \log f(x_1, \ldots, x_n) = \underset{\tilde{\mu}, \tilde{\sigma}^2}{\operatorname{argmax}} -\frac{n}{2} \log(2\pi\tilde{\sigma}^2) - \sum_{i=1}^{n} \frac{(x_i - \tilde{\mu})^2}{2\tilde{\sigma}^2}$$

- Tomando a derivada em $\tilde{\mu}$ e $\tilde{\sigma}^2$ e colocando-as igual a zero, obtemos

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i, \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

- Isso sugere estimadores para os parâmetros μ e σ^2; note que eles são bastante razoáveis

Exemplo: viés e variância de estimador

- Vamos considerar um exemplo em que $Y = f(X) + \varepsilon = \mu + \varepsilon$, onde $\varepsilon \sim N(0, \sigma^2)$, ou seja, não temos covariadas X, só Y observado. Queremos estimar o valor de um novo Y_* não-observado
- Pela hipótese acima, $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ e queremos estimar o novo Y_* via $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$
- Em primeiro lugar, por definição, $\mathbb{E}[Y_1] = \mu$ (então $\text{bias}(Y_1) = 0$) e $\mathbb{V}[Y_1] = \sigma^2$. Daí,

  $$\text{MSE}(Y_1, Y_*) = \mathbb{E}[(Y_1 - Y_*)^2] = \mathbb{E}[(Y_1 - (\mu + \varepsilon_*))^2] = \mathbb{E}[(Y_1 - \mu)^2] + \mathbb{E}[\varepsilon_*^2] = \mathbb{V}[Y_1] + \mathbb{V}[\varepsilon_*] = 2\sigma^2$$

- Mas em algum sentido o estimador $\overline{Y}$ deve ser melhor. De fato, $\mathbb{E}[\overline{Y}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i] = \mu$ e $\mathbb{V}[\overline{Y}] = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}[Y_i] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$. Daí

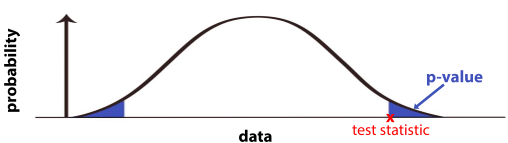
  $$\text{MSE}(\overline{Y}, Y_*) = \mathbb{E}[(\overline{Y} - Y_*)^2] = \mathbb{E}[(\overline{Y} - \mu)^2] + \mathbb{E}[\varepsilon_*^2] = \mathbb{V}[\overline{Y}] + \mathbb{V}[\varepsilon_*] = \frac{\sigma^2}{n} + \sigma^2 = \frac{n+1}{n}\sigma^2,$$

  e para $n > 1$, $\overline{Y}$ tem erro médio quadrático menor

Inferência: teste de hipótese

- Se observamos uma amostra $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ e $\overline{y} = 0.02$, é possível dizer que $\mu = 0$?
- Aqui, vamos assumir que σ^2 é um valor conhecido, e desconhecemos apenas μ
- Duas hipóteses, $H_0 : \mu = 0$ e $H_1 : \mu \neq 0$. Se H_0 vale, qual é a probabilidade de observar $\overline{y} = 0.02$?
- Vamos usar o estimador de antes, $\overline{Y} \sim N(\mu, \sigma^2/n)$. Daí, temos

$$\frac{\overline{Y} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1) \overset{H_0}{\Longrightarrow} T = \frac{\overline{Y}}{\sqrt{\sigma^2/n}} \sim N(0,1)$$



- Se $n = 10^4$, $\sigma^2 = 1$, $t = \sqrt{10^4} \cdot 0.02 = 2$, e $P[|Z| > t] = 0.045$, então a nível 95% rejeitamos H_0

Otimização

- Convexidade
- Métodos de primeira ordem
  - descida de gradiente
  - descida de gradiente estocástico
- Otimização com restrições

Convexidade

- Estamos interessados em encontrar o melhor fit para os dados; "melhor" sugere otimização
- Problema: é raro conseguirmos encontrar um mínimo global, mas não mínimo local
- Existe uma classe de funções onde mínimo local implica em mínimo global: funções convexas
- Uma função $f : \mathbb{R}^n \to \mathbb{R}$ é convexa se, dados dois pontos $\boldsymbol{\theta}_1$ e $\boldsymbol{\theta}_2$ e $\alpha \in [0, 1]$,

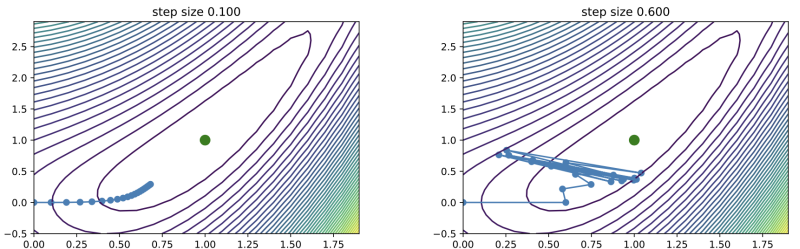
$$f(\alpha\boldsymbol{\theta}_1 + (1-\alpha)\boldsymbol{\theta}_2) \leq \alpha f(\boldsymbol{\theta}_1) + (1-\alpha) f(\boldsymbol{\theta}_2)$$

- Se $f \in C^2$, ela é convexa se e somente se a Hessiana $\nabla^2 f(\theta)$ é positiva semi-definida
- Exemplos: $e^{a\theta}$, θ^a ($a \geq 1$ ou $a \leq 0$), $\theta \log \theta$, $-\log \theta$, $\mathbf{a}^T\boldsymbol{\theta} + b$, $\|\boldsymbol{\theta}\|^p$ ($p \geq 1$), $\boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta} + \mathbf{b}^T\boldsymbol{\theta} + c$ ($\mathbf{C} \succeq 0$)
- Funções convexas podem ser compostas: (i) se f é convexa, $c \cdot f$ também é, onde $c \geq 0$; (ii) se f_1 e f_2 são convexas, $f_1 + f_2$ também; (iii) se $f_1, \ldots, f_p$ são convexas, $\max\{f_1(\boldsymbol{\theta}), \ldots, f_p(\boldsymbol{\theta})\}$ também

Métodos de primeira ordem: descida de gradiente

- Se f é diferenciável, um método iterativo para mínimos locais é caminhar na direção do gradiente:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \rho_t \nabla f(\boldsymbol{\theta}^t)$$



Credit: Kevin Murphy

- Aqui, $\rho_t \in \mathbb{R}$ é chamado de taxa de aprendizado (ou tamanho de passo), e pode variar com t
- Esse método é chamado de descida de gradiente

Métodos de primeira ordem: descida de gradiente estocástico

- Como estamos lidando com dados aleatórios, em geral queremos minimizar $f(\boldsymbol{\theta}) = \mathbb{E}[L(\boldsymbol{\theta}, \mathbf{x})]$, onde $\mathbf{x} \sim p(\mathbf{x})$ é uma variável aleatória representando os dados e $\boldsymbol{\theta}$ representa um parâmetro
- Descida de gradiente: aproximamos $\mathbb{E}[L(\boldsymbol{\theta}, \mathbf{x})] \approx \frac{1}{n} \sum_{i=1}^{n} L(\boldsymbol{\theta}, \mathbf{x}_i)$ com $\{\mathbf{x}_i\}_{i=1}^{n}$ amostra de treino
- Mas isso é muito custoso; vamos usar uma sub-amostra (ou batch $\mathcal{B}_t$) apenas:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \rho_t \left(\frac{1}{|\mathcal{B}_t|} \sum_{i \in \mathcal{B}_t} \nabla L(\boldsymbol{\theta}^t, \mathbf{x}_i) \right),$$

- Esse método é chamado de descida de gradiente estocástico; é fundamental em machine learning
- A escolha de learning rate é fundamental para garantir convergência; *e.g.*, $\rho_t = \rho_0 e^{-\lambda t}$
- Na teoria, a convergência de SGD é mais lenta, mas na prática é mais rápido porque cada passo toma menos tempo. SGD tem outras vantagens, mas ainda não são totalmente entendidas

Otimização com restrições: igualdade

- Às vezes queremos minimizar uma função sujeita a restrições de igualdade:

$$\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \text{ tal que } h_i(\boldsymbol{\theta}) = 0, i \in \mathcal{E}$$

- Repare que num ótimo pode não valer $\nabla L(\boldsymbol{\theta}^*) = 0$
- Mas: (i) $\nabla h_i(\boldsymbol{\theta})$ é perpendicular à curva de nível, pois $h_i(\boldsymbol{\theta} + \boldsymbol{\epsilon}) \approx h_i(\boldsymbol{\theta}) + \boldsymbol{\epsilon}^T \nabla h_i(\boldsymbol{\theta})$; (ii) $\nabla L(\boldsymbol{\theta}^*)$ também é perpendicular, senão poderíamos diminuir a função objetivo ao longo da curva. Daí:

$$\nabla L(\boldsymbol{\theta}^*) = \lambda_i \nabla h_i(\boldsymbol{\theta}^*)$$

- Para encapsular essa condição, e a restrição $h_i(\boldsymbol{\theta}^*) = 0$, basta derivar e igual a zero o Lagrangeano

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\lambda}} L(\boldsymbol{\theta}) + \sum_{i \in \mathcal{E}} \lambda_i h_i(\boldsymbol{\theta})$$

- Se L é convexa e h_i também, o mínimo local é global

Otimização com restrições: exemplo

- Resolva o seguinte problema, assumindo $\mathbf{A}$ simétrica positiva semi-definida:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \mathbf{x}^T\mathbf{A}\mathbf{x} \text{ tal que } \|\mathbf{x}\|_2^2 = 1$$

- Este é um problema convexo com restrição de igualdade. O Lagrangeano é

$$\mathbf{x}^T\mathbf{A}\mathbf{x} + \lambda(\mathbf{x}^T\mathbf{x} - 1)$$

- A condição de primeira ordem é

$$2\mathbf{A}\mathbf{x} + 2\lambda\mathbf{x} = 0,$$

  ou seja, $\mathbf{A}\mathbf{x} = -\lambda\mathbf{x}$

- Isto é, o valor mínimo é um autovalor λ; naturalmente, deve ser o menor deles, λ_1

- O mínimo é atingido pelo autovetor associado ao autovalor λ_1

Otimização com restrições: desigualdade

- Às vezes queremos minimizar uma função sujeita a restrições de desigualdade:

$$\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \text{ tal que } g_j(\boldsymbol{\theta}) \leq 0, j \in \mathcal{I}$$

- Para simplificar, vamos considerar o caso de uma desigualdade apenas. Se $\mu \geq 0$, defina

$$\tilde{L}(\boldsymbol{\theta}) = \max_{\mu \geq 0} L(\boldsymbol{\theta}) + \mu g(\boldsymbol{\theta}) = \begin{cases} \infty & \text{se } g(\boldsymbol{\theta}) > 0 \\ L(\boldsymbol{\theta}) & \text{caso contrário} \end{cases}$$

  e note que resolver esse problema é equivalente ao original

- Ou seja, reformulamos o problema como $\min_{\boldsymbol{\theta}} \max_{\mu \geq 0} L(\boldsymbol{\theta}) + \mu g(\boldsymbol{\theta})$

- Se tivermos várias restrições $g_j(\boldsymbol{\theta}) \leq 0$, $j \in \mathcal{I}$, o problema se torna

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\mu} \geq \mathbf{0}} L(\boldsymbol{\theta}) + \sum_{j \in \mathcal{I}} \mu_j g_j(\boldsymbol{\theta})$$

Otimização com restrições: condições de KKT

- Colocando tudo junto, para resolver um problema convexo (*i.e.*, L, g_j convexas e C^1, h_i afim)

$$\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \text{ tal que } h_i(\boldsymbol{\theta}) = 0, g_j(\boldsymbol{\theta}) \leq 0, i \in \mathcal{E}, j \in \mathcal{I},$$

  olhamos para

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\mu} \geq \mathbf{0}, \boldsymbol{\lambda}} L(\boldsymbol{\theta}) + \sum_{i \in \mathcal{E}} \lambda_i h_i(\boldsymbol{\theta}) + \sum_{j \in \mathcal{I}} \mu_j g_j(\boldsymbol{\theta})$$

- Condições de KKT: suficiência (e necessidade*) para resolver o problema acima:
  - (Estacionariedade) Ótimo é estacionário: $\nabla L(\boldsymbol{\theta}^*) + \sum_{i \in \mathcal{E}} \lambda_i \nabla h_i(\boldsymbol{\theta}^*) + \sum_{j \in \mathcal{I}} \mu_j \nabla g_j(\boldsymbol{\theta}^*) = \mathbf{0}$
  - (Viabilidade) As restrições precisam ser satisfeitas: $g_j(\boldsymbol{\theta}^*) \leq 0, h_i(\boldsymbol{\theta}^*) = 0$ para $i \in \mathcal{E}, j \in \mathcal{I}$
  - (Viabilidade dual) A penalidade das desigualdades é positiva: $\boldsymbol{\mu} \geq \mathbf{0}$
  - (Folga complementar) Restrições inativas zeram o multiplicador: $\mu_j g_j(\boldsymbol{\theta}^*) = 0$ para $j \in \mathcal{I}$

*Sob condições de regularidade, *e.g.*, Slater: existe x tal que $h_i(x) = 0$ e $g_j(x) < 0$