

1 Numerical computation and finite numbers

Absolute error $E_x = \tilde{x} - x$

Relative error $R_x = \frac{\tilde{x} - x}{x}$

$x \in \mathbb{R}$, $fl(x) \in \mathcal{F}(\beta, t, L, U) \Leftrightarrow fl(x) = \pm(d_1\beta^{-1} + d_2\beta^{-2} + \dots + d_t\beta^{-t})\beta^p = \pm m\beta^p$ with $0 \leq d_i \leq \beta - 1$, $L \leq p \leq U$ (β "base", t "precision", m "mantissa", p "exponent").

$UFL = \beta^{L-1}$; $OFL = \beta^U(1 - \beta^{-t})$

Machine precision with rounding by chopping: $\epsilon_{mach} = \beta^{1-t}$; rounding to nearest: $\epsilon_{mach} = \frac{1}{2}\beta^{1-t}$.

2 Eigenvectors and eigenvalues

$A \in \mathbb{R}^{n \times n}$; $A\vec{x} = \lambda\vec{x} \Leftrightarrow \vec{x}$ eigenvector and λ eigenvalue of A .

A triangular or symmetric \Rightarrow eigenvalues are on the main diagonal.

Spectrum $\sigma(A) = \{\lambda : \vec{x} \text{ eigenvector of } A\}$. Spectral norm $\rho(A) = \max|\lambda|$

$C \in \mathbb{R}^{n \times n}$ singular $\Leftrightarrow \det(C) = 0$

Similarity transformation: $A, C \in \mathbb{R}^{n \times n}$; C non-singular; A and $C^{-1}AC$ are similar (same spectrum and eigenvalues).

$A \in \mathbb{R}^{m \times n} \Rightarrow A^T A \in \mathbb{R}^{n \times n}$ is positive semi-definite.

$A \in \mathbb{R}^{m \times n}$ with maximum rank ($rk(A) = \min(m, n)$) $\Rightarrow A^T A \in \mathbb{R}^{n \times n}$ is positive definite.

Spectral theorem: $A \in \mathbb{R}^{n \times n}$ symmetric \Rightarrow eigenvalues are real, eigenvectors create an orthogonal basis.

3 Norm

Scalar product: $\vec{x}, \vec{y} \in V = \mathbb{R}^n$, $\langle \vec{x}, \vec{y} \rangle = \sum_{i=1}^n x_i y_i$

$\|\vec{x}\| \geq 0 \forall \vec{x} \in V$; $\|\vec{x}\| = 0 \iff \vec{x} = \vec{0}$; $\|\alpha\vec{x}\| = |\alpha|\|\vec{x}\| \forall \alpha \in \mathbb{R}, \vec{x} \in V$; $\|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\| \forall \vec{x}, \vec{y} \in V$.

p-norm: $p \in [1, \infty[$, $\|\vec{x}\|_p = \sqrt[p]{\sum_{i=1}^n |x_i|^p}$

- 1-norm (a.k.a. Manhattan norm): $\|\vec{x}\|_1 = \sum_{i=1}^n |x_i|$

- 2-norm (a.k.a. Euclidean norm): $\|\vec{x}\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2} = \sqrt{\langle \vec{x}, \vec{x} \rangle}$

- Infinity-norm: $\|\vec{x}\|_\infty = \max |x_i|$

Distance $d(\vec{x}, \vec{y}) = \|\vec{y} - \vec{x}\|$

3.1 Matrix norm

Similar properties of vector norm, plus $\|AB\| \leq \|A\|\|B\| \forall A, B \in \mathbb{R}^{n \times n}$

Frobinus norm: $\|A\|_f = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2}$

p-induced matrix norm:

- 1-norm: $\|A\|_1 = \max_{j=1..n} \sum_{i=1}^m |a_{ij}|$

- 2-norm: $\|A\|_2 = \sqrt{\rho(A^T A)}$

- Infinity-norm: $\|A\|_\infty = \|A^T\|_1 = \max_{i=1..m} \sum_{j=1}^n |a_{ij}|$

Condition number $K_p(A) = \|A\|_p \|A^{-1}\|_p$. A singular $\Leftrightarrow K_p(A) = \infty$. $K_2(A) = \frac{\lambda_{max}}{\lambda_{min}}$

4 Matrix decompositions / factorizations

4.1 LU decomposition

$A \in \mathbb{R}^{n \times n}$ non-singular ($\det(A) \neq 0$) with all principal minors non-singular $\Rightarrow A = LU$ with $L \in \mathbb{R}^{n \times n}$ lower triangular and $U \in \mathbb{R}^{n \times n}$ upper triangular.

4.2 Cholesky factorization

$A \in \mathbb{R}^{n \times n}$ positive definite $\Rightarrow A = LL^T$ with $L \in \mathbb{R}^{n \times n}$ lower triangular.

4.3 Singular Value Decomposition (SVD)

$A \in \mathbb{R}^{m \times n}$, $r = \text{rk}(A) \in [0, \min(m, n)] \Rightarrow A = U\Sigma V^T$ with

- $U \in \mathbb{R}^{m \times m}$ orthogonal.
- $V \in \mathbb{R}^{n \times n}$ orthogonal.
- $\Sigma \in \mathbb{R}^{m \times n}$ with $\Sigma_{ii} = \sigma_i$ ("singular value") and $i \neq j \Rightarrow \Sigma_{ij} = 0$.

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > \sigma_{r+1} = \sigma_{r+2} = \dots = \sigma_n = 0.$$

$\sigma_i = \sqrt{\lambda_i(A^T A)}$ where $\lambda_i(A)$ is the i -th eigenvalue of A by value.

$$\sigma_1 = \sqrt{\rho(A^T A)} = \|A\|_2. \quad \|A^{-1}\|_2 = \frac{1}{\sigma_r}. \quad K_2(A) = \frac{\sigma_1}{\sigma_r}.$$

4.3.1 Rank-k-approximation

$A = U\Sigma V^T = \sum_{i=1}^r \sigma_i u_i v_i^T = \sum_{i=1}^r \sigma_i A_i$ with $u_i \in \mathbb{R}^m$ column of U and $v_i \in \mathbb{R}^n$ column of V . $\hat{A}_k = \sum_{i=1}^k \sigma_i u_i v_i^T = \sum_{i=1}^k \sigma_i A_i$ with $k < r$ is the rank-k-approximation of A .

5 Vector calculus

Chain rule: $g(f(x))' = (g \circ f)'(x) = g'(f(x))f'(x)$

$f: \mathbb{R}^n \rightarrow \mathbb{R}$; Partial derivative $\frac{\partial f}{\partial x_i}(\vec{x}) = f_{x_i}(\vec{x}) = \lim_{h \rightarrow 0} \frac{f(x_1, \dots, x_i + h, \dots, x_n) - f(\vec{x})}{h}$

Gradient $\nabla f(\vec{x}) = \left(\frac{\partial f}{\partial x_1}(\vec{x}), \dots, \frac{\partial f}{\partial x_n}(\vec{x}) \right)$

Second order partial derivative $\frac{\partial^2 f}{\partial x_i \partial x_j}(\vec{x}) = f_{x_i x_j}(\vec{x}) = \frac{\partial}{\partial x_j} \frac{\partial f}{\partial x_i}(\vec{x})$

Hessian $\nabla^2 f(\vec{x}) = H_f(\vec{x}) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(\vec{x}) \right)_{i,j=1,\dots,n} = \begin{pmatrix} f_{x_1 x_1}(\vec{x}) & \dots & f_{x_n x_1}(\vec{x}) \\ \vdots & \ddots & \vdots \\ f_{x_1 x_n}(\vec{x}) & \dots & f_{x_n x_n}(\vec{x}) \end{pmatrix}$

5.1 Useful identities for computing gradients

$\vec{x}, \vec{a}, \vec{b} \in \mathbb{R}^n$; $X \in \mathbb{R}^{n \times n}$;

$$\frac{\partial \vec{f}(X)^T}{\partial X} = \left(\frac{\partial \vec{f}(X)}{\partial X} \right)^T;$$

$$\frac{\partial \vec{f}(X)^{-1}}{\partial X} = -\vec{f}(X)^{-1} \frac{\partial \vec{f}(X)}{\partial X} \vec{f}(X)^{-1}$$

$$\frac{\partial \vec{x}^T \vec{a}}{\partial \vec{x}} = \frac{\partial \vec{a}^T \vec{x}}{\partial \vec{x}} = \vec{a}^T;$$

$$\frac{\partial}{\partial X} \vec{a}^T X \vec{b} \vec{a} = \vec{a}^T \vec{b};$$

$$\frac{\partial (\vec{a}^T X \vec{a})}{\partial \vec{a}} = \vec{a}^T (X + X^T);$$

$$\frac{\partial \|\vec{a} - X\vec{b}\|_2^2}{\partial X} = \frac{\partial \|X\vec{b} - \vec{a}\|_2^2}{\partial X} = 2(X\vec{b} - \vec{a})^T X = 2X^T(X\vec{b} - \vec{a}) = 2(X^T X\vec{b} - X^T \vec{a});$$

6 Linear systems

$A\vec{x} = \vec{b}$ with $A \in \mathbb{R}^{m \times n}$, $\vec{x} \in \mathbb{R}^n$, $\vec{b} \in \mathbb{R}^m$. $m = n \Leftrightarrow$ Square linear system.

A^{-1} exists $\Leftrightarrow \vec{x} = A^{-1}\vec{b}$

6.1 Least squares problem

\vec{x}^* solution of $A\vec{x} = \vec{b} \Rightarrow \vec{x}^* \cong \arg \min_{\vec{x} \in \mathbb{R}} \|A\vec{x} - \vec{b}\|^2$ (strictly convex, only one minimum which is global).

7 Optimization

$\max_{\vec{x} \in \mathbb{R}} f(\vec{x}) = -\min_{\vec{x} \in \mathbb{R}} -f(\vec{x})$; $\arg \max_{\vec{x} \in \mathbb{R}} f(\vec{x}) = \arg \min_{\vec{x} \in \mathbb{R}} -f(\vec{x})$; We search $\arg \min_{\vec{x} \in \mathbb{R}} f(\vec{x})$

7.1 Iterative methods

$\alpha_k \in \mathbb{R}$ step length; $\vec{p}_k \in \mathbb{R}^n$ descent direction for f in \vec{x}_k ($\vec{p}_k^T \cdot \nabla f(\vec{x}_k) < 0$).
while($k < k_{Max} \wedge \|\nabla f(\vec{x}_k)\| < tol_f \wedge \|\vec{x}_k - \vec{x}_{k-1}\| \geq tol_x$) $\vec{x}_{k+1} = \vec{x}_k + \alpha_k \vec{p}_k$
 Convergence speed:

- Q-linear: $\exists r \in]0, 1[, \vec{x}^*, k^* : \|\vec{x}_{k+1} - \vec{x}^*\| \leq r \|\vec{x}_k - \vec{x}^*\| \forall k > k^*$
- Q-quadratic $\exists M > 0, \vec{x}^*, k^* : \|\vec{x}_{k+1} - \vec{x}^*\| \leq M \|\vec{x}_k - \vec{x}^*\|^2 \forall k > k^*$

7.1.1 Gradient Descent method

Q-linear, uses only first order gradient: $\vec{x}_{k+1} = \vec{x}_k - \alpha_k \nabla f(\vec{x}_k)$

7.1.2 Gradient Descent with momentum

$$\vec{x}_{k+1} = \vec{x}_k - \alpha_k \nabla f(\vec{x}_k) + \beta_k (\vec{x}_k - \vec{x}_{k-1})$$

7.1.3 Stochastic Gradient Descent

$L(\theta) = \sum_{n=1}^N L_n(\theta)$; $\vec{x}_{k+1} = \vec{x}_k - \alpha_k \nabla L(\theta)$ with:

- Ordinary Gradient Descent: $\nabla L(\theta) = \sum_{n=1}^N \nabla L_n(\theta)$
- Random item: $\forall k \ i_k \in \{0, 1, \dots, N\}$; $\nabla L(\theta) \approx \nabla L_{i_k}(\theta)$
- Mini-batch: $p < n$; $\forall k \ i_{1k}, i_{2k}, \dots, i_{pk} \in \{0, 1, \dots, N\}$; $\nabla L(\theta) \approx \sum_{j=1}^p \nabla L_{i_{jk}}(\theta)$

7.1.4 Newton method

Q-quadratic, uses also higher order info: $H_f(\vec{x}_k) \vec{p}_k = -\nabla^T f(\vec{x}_k)$ (linear system with solution \vec{p}_k)

8 Statistics

Ω sample space, $A \subseteq \Omega$ event space, $P : A \rightarrow [0, 1]$ probability, $P(\Omega) = 1$.

8.1 Discrete random variables

$X : A \rightarrow T \subset \mathbb{R}$ discrete random variable (Target/Support space T finite or numerable); $x \in T$.

Probability Mass Function $f_X(x) = P(X = x)$. $\sum_{x \in T} f_X(x) = P(T) = 1$.

Mean PMF $\mu = E(f_X) = \sum_{x \in T} x f_X(x)$. Variance $\sigma^2 = \sum_{x \in T} (x - \mu)^2 f_X(x)$. Standard deviation $\sigma = \sqrt{\sigma^2}$.

Uniform distribution: $f_X(x) = \frac{1}{N}$.

Poisson dist.: $f_X(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$ (λ mean of events in unit). $\mu = \lambda$. $\sigma = \lambda$.

8.2 Continuous random variables

$X : A \rightarrow T \subseteq \mathbb{R}$ continuous random variable; $x \in T$.

$f_X : T \rightarrow \mathbb{R}$ Probability Density Function. $P(a \leq x \leq b) = \int_a^b f_X(x) dx$. $\int_T f_X(x) dx = P(T) = 1$

Mean PDF $\mu = E(f_X) = \int_T x f_X(x) dx$. Variance $\sigma^2 = \int_T (x - \mu)^2 f_X(x) dx$. Standard deviation $\sigma = \sqrt{\sigma^2}$.

Gaussian/Normal distribution: $f_X(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$

Cumulative Distribution Function (for both discrete and continuous) $F_X : T \rightarrow [0, 1] : F_X(x) = P(X \leq x)$

8.3 Multivariate probability

$X : A \rightarrow T_X; Y : A \rightarrow T_Y; T_{XY} = T_X \times T_Y.$

Joint probability $P(X = x, Y = y) = P(X = x \wedge Y = y).$

Marginal probability $P(X = x) = \begin{cases} \sum_{y \in T_Y} P(X = x, Y = y) & \text{if } Y \text{ discrete} \\ \int_{T_Y} P(X = x, Y = y) dy & \text{if } Y \text{ continuous} \end{cases}.$

Conditional probability $P(\text{Effect} \mid \text{Cause}) = \frac{P(\text{Effect} \wedge \text{Cause})}{P(\text{Cause})}$

Bayes theorem: $P(\text{Cause} \mid \text{Effect}) = \frac{P(\text{Effect} \mid \text{Cause})P(\text{Cause})}{P(\text{Effect})}$

8.4 Statistical and conditional independence

$Cov(x, y) = 0 \Leftrightarrow P \models (A \perp B) \Leftrightarrow P(A \mid B) = P(A) \Leftrightarrow P(B \mid A) = P(B) \Leftrightarrow P(A, B) = P(A)P(B)$
 $P \models (A \perp B \mid C) \Leftrightarrow P(A \mid B, C) = P(A, C) \Leftrightarrow P(B \mid A, C) = P(B, C) \Leftrightarrow P(A, B \mid C) = P(A \mid C)P(B \mid C)$

9 Learning

N observations $\vec{x}_n \in \mathbb{R}^D$ and labels $y_n \in \mathbb{R}; X = [\vec{x}_1, \dots, \vec{x}_N]^T; \vec{y} = [y_1, \dots, y_N]^T$; parameters $\vec{\theta} \in \mathbb{R}^D$

9.1 Empirical Risk Minimization

Linear model $f(\cdot, \vec{\theta}) : \mathbb{R}^D \rightarrow \mathbb{R} : f(\vec{x}) = \vec{\theta}^T \vec{x} + \theta_0 = \theta_0 + \sum_{d=1}^D \theta_d x_{n,d}$

We search $\vec{\theta}^* : f(\vec{x}_n, \vec{\theta}^*) = \hat{y}_n \approx y_n \forall n = 1, 2, \dots, N$

Loss function $l(y, \hat{y})$; Empirical risk $R_{emp}(f, X, \vec{y}, \vec{\theta}) = \frac{1}{N} \sum_{n=1}^N l(y_n, f(\vec{x}_n, \vec{\theta}))$

$\vec{\theta}^* = \min_{\vec{\theta} \in \mathbb{R}^D} R_{emp}(f, X, \vec{y}, \vec{\theta}) = \min_{\vec{\theta} \in \mathbb{R}^D} \frac{1}{N} \sum_{n=1}^N (y_n - \vec{x}_n^T \vec{\theta})^2 = \min_{\vec{\theta} \in \mathbb{R}^D} \frac{1}{N} \|\vec{y} - X\vec{\theta}\|^2$

9.2 Maximum Likelihood Estimation (ML)

Family of probability densities $p(\vec{x} \mid \vec{\theta})$; Loss $\mathcal{L}_x(\vec{\theta}) = -\log p(\vec{x} \mid \vec{\theta})$; $\theta^* = \min_{\vec{\theta}} \mathcal{L}_x(\vec{\theta})$

$p(\vec{y} \mid X, \vec{\theta}) = \prod_{n=1}^N p(y_n \mid \vec{x}_n, \vec{\theta}) \Rightarrow \mathcal{L}_x(\vec{\theta}) = -\log \left(\prod_{n=1}^N p(y_n \mid \vec{x}_n, \vec{\theta}) \right) = -\sum_{n=1}^N \log p(y_n \mid \vec{x}_n, \vec{\theta})$

$p(y_n \mid \vec{x}_n, \vec{\theta}) \sim \mathcal{N}(y_n - \vec{\theta}^T \vec{x}_n, \sigma^2) \Rightarrow \vec{\theta}^* = \min_{\vec{\theta}} \frac{1}{2\sigma^2} \|\vec{y} - X\vec{\theta}\|_2^2$

9.3 Maximum A Posteriori Estimation (MAP)

$p(\vec{\theta} \mid \vec{x}) = \frac{p(\vec{x} \mid \vec{\theta})p(\vec{\theta})}{p(\vec{x})}$; $\vec{\theta}^* = \min_{\vec{\theta}} -\log(p(\vec{\theta} \mid \vec{x})) = \min_{\vec{\theta}} -(\log(p(\vec{x} \mid \vec{\theta})) + \log(p(\vec{\theta})))$

$p(\vec{y} \mid X, \vec{\theta}) = \prod_{n=1}^N p(y_n \mid \vec{x}_n, \vec{\theta}) \Rightarrow TODO$

$p(y_n \mid \vec{x}_n, \vec{\theta}) \sim \mathcal{N}(y_n - \vec{\theta}^T \vec{x}_n, \sigma^2) \Rightarrow \vec{\theta}^* = \min_{\vec{\theta}} \frac{1}{2\sigma^2} \|\vec{y} - X\vec{\theta}\|_2^2 + \|\vec{\theta}\|_2^2$