1 Error measurement

Absolute error $E_x = \tilde{x} - x$
Relative error $R_x = \dfrac{\tilde{x} - x}{x}$

2 Matrices

$A \in \mathbb{R}^{nxn}; \vec{x}$ right eigenvector $\Leftrightarrow A\vec{x} = \lambda\vec{x}; \vec{x}$ left eigenvector $\Leftrightarrow A\vec{x} = \lambda\vec{x}$. λ are eigenvalues.
A triangular or symmetric $\Rightarrow$ eigenvalues are on the main diagonal.
Spectrum $\sigma(A) = \{\vec{x} : \vec{x}$ eigenvector of $A\}$. Spectral norm $\rho(A) = max(\lambda)$
$C \in \mathbb{R}^{nxn}$ singular $\Leftrightarrow \det(C) = 0$
Similarity transformation: $A, C \in \mathbb{R}^{nxn}$, C non-singular $\Rightarrow A$ and $C^{-1}AC$ are similar (same spectrum and eigenvalues).
$A \in \mathbb{R}^{mxn} \Rightarrow A^TA \in \mathbb{R}^{nxn}$ is positive semi-definite.
$A \in \mathbb{R}^{mxn}$ with maximum rank $(rk(A) = min(m, n)) \Rightarrow A^TA \in \mathbb{R}^{nxn}$ is positive definite.
Spectral theorem: $A \in \mathbb{R}^{nxn}$ symmetric $\Rightarrow$ eigenvalues are real, eigenvectors create an orthogonal basis.

3 Norm

Scalar product: $\vec{x}, \vec{y} \in V = \mathbb{R}^n$, $\langle\vec{x}, \vec{y}\rangle = \sum_{i=1}^{n} x_iy_i$
$\|\vec{x}\| \geq 0 \; \forall\vec{x} \in V$; $\|\vec{x}\| = 0 \iff \vec{x} = \vec{0}$
$\|\alpha\vec{x}\| = |\alpha|\|\vec{x}\| \; \forall\alpha \in \mathbb{R}, \vec{x} \in V$
$\|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\| \; \forall\vec{x}, \vec{y} \in V$.
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): $\|\vec{x}\|_2 = \sqrt{\sum_{i=1}^{n}|x_i|^2}$
- 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 $\|AB\| \leq \|A\|\|B\| \; \forall A, B \in \mathbb{R}^{nxn}$
Frobinius norm: $\|A\|_p = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}$
p-induced matrix norm:

- 1-norm: $\|A\|_1 = \max_{j=1..n} \sum_{i=1}^{m}|a_{ij}|$
- 2-norm: $\|A\|_2 = \sqrt{\rho(A^TA)}$
- Infinity-norm: $\|A\|_\infty = \max_{i=1..m} \sum_{v=1}^{n}|a_{ij}|$

4 Projections

TODO

5 Matrix decompositions / factorizations

5.1 LU decomposition

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

5.2 Cholesky factorization

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

5.3 Singular Value Decomposition (SVD)

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

- $U \in \mathbb{R}^{mxm}$ orthogonal.
- $V \in \mathbb{R}^{nxn}$ orthogonal.
- $\Sigma \in \mathbb{R}^{mxn}$ with $\Sigma_{ii} = \sigma_i$ ("singular value") and $i \neq f \Rightarrow \Sigma_i j = 0$.

$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \sigma_{r+2} = \ldots = \sigma_n = 0$.
$\sigma_i = \sqrt{\lambda_i(A^T A)}$ where $\lambda_i(A)$ is the i-th eigenvalue of A by value.
$\sigma_1 = \sqrt{\rho(A^T A)} = \|A\|_2$. $\|A^{-1}\|_2 = \dfrac{1}{\sigma_r}$. $K_2(A) = \dfrac{\sigma_1}{\sigma_r}$.

5.3.1 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.

6 Vector calculus

$f : \mathbb{R}^n \to \mathbb{R}^m$
Partial derivative $\frac{\partial f}{\partial x_i}(\vec{x}) = f_{x_i}(\vec{x}) = D_{x_i} f(\vec{x}) = \lim_{h \to 0} \frac{f(x_1, \ldots, x_i + h, \ldots, x_n) - f(\vec{x})}{h}$
Gradient $\nabla f(\vec{x}) = \left(\frac{\partial f}{\partial x_1}(\vec{x}), ..., \frac{\partial f}{\partial x_n}(\vec{x}) \right)$
Second order partial derivative $\frac{\partial^2 f}{\partial x_i \partial x_j}(\vec{x}) = f_{x_i x_j}(\vec{x}) = D_{x_i x_j} f(\vec{x}) = \frac{\partial}{\partial x_j} \frac{\partial f}{\partial x_i}(\vec{x})$
Hessian $\nabla^2 f(\vec{x}) = H_f(\vec{x}) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(\vec{x}) \right)_{i,j=1,\ldots,n} = \begin{pmatrix} f_{x_1 x_1}(\vec{x}) & \cdots & f_{x_n x_1}(\vec{x}) \\ \vdots & \ddots & \vdots \\ f_{x_1 x_n}(\vec{x}) & \cdots & f_{x_n x_n}(\vec{x}) \end{pmatrix}$

7 Linear systems

7.1 Least squares problem

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

8 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 $\arg\min_{\vec{x} \in \mathbb{R}} f(\vec{x})$

8.1 Iterative methods

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

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

8.1.1 Gradient Descent method

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

8.1.2 Gradient Descent with momentum

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

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

8.1.4 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 $\vec{p}_k$)

9 Statistics

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

9.1 Discrete random variables

$X : A \to T \subset \mathbb{R}$ 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). $\mu = \lambda$. $\sigma = \lambda$.

9.2 Continuous random variables

$X : A \to T \subseteq \mathbb{R}$ continuous random variable; $x \in T$.
$f_X : T \to \mathbb{R}$ 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) $F_X : T \to [0, 1] : F_X(x) = P(X \leq x)$

9.3 Multivariate probability

$X : A \to T_X$; $Y : A \to T_Y$; $T_{XY} = T_X \times T_Y$.
Joint probability $P(X = x, Y = y) = P(X = x \wedge Y = y)$.

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

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

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