

Sum Rule $P(X = x_i) = \sum_{j=1}^J p(X = x_i, Y = y_j)$			
Product rule	$P(X, Y) = P(Y X)P(X)$		
Independence	$P(X, Y) = P(X)P(Y)$		
Conditional Independence	$P(X, Y Z) = P(X Z)P(Y Z)$		
Bayes' Rule	$P(Y X) = \frac{P(X Y)P(Y)}{P(X)} = \frac{P(X Y)P(Y)}{\sum_{i=1}^k P(X Y_i)P(Y_i)}$		
Cond.	Ind.	$X \perp Y Z \implies P(X, Y Z) = P(X Z)P(Y Z)$	

$\mathbb{E}[X] = \int_{\mathcal{X}} t \cdot f_X(t) dt =: \mu_{\mathcal{X}}$
$\text{Cov}(X, Y) = \mathbb{E}_{x,y} [(X - \mathbb{E}_x[X])(Y - \mathbb{E}_y[Y])]$
$\text{Cov}(X) := \text{Cov}(X, X) = \text{Var}[X]$
X, Y independent $\implies \text{Cov}(X, Y) = 0$
$\mathbf{XX}^T \geq 0$ (symmetric positive semidefinite)
$\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
$\text{Var}[\mathbf{AX}] = \mathbf{A} \text{Var}[\mathbf{X}] \mathbf{A}^T \quad \text{Var}[aX + b] = a^2 \text{Var}[X]$

$\text{Var}\left[\sum_{i=1}^n a_i X_i\right] = \sum_{i=1}^n a_i^2 \text{Var}[X_i] + 2 \sum_{i < j} a_i a_j \text{Cov}(X_i, X_j)$
$\frac{\partial}{\partial t} P(X \leq t) = \frac{\partial}{\partial t} F_X(t) = f_X(t)$ (derivative of c.d.f. is p.d.f)
$f_{\alpha \gamma}(z) = \frac{1}{\alpha} f_{\gamma}(\frac{z}{\alpha})$

T. (Moment Generating Function)
The moment generating function (MGF) $\psi_{\mathcal{X}}(t) = \mathbb{E}[e^{t\mathcal{X}}]$ characterizes the distribution of a random variable \mathcal{X} .
$Be(p) \quad pe^t + (1 - p)$
$\mathcal{N}(\mu, \sigma) \quad \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$
$Bin(n, p) \quad (pe^t + (1 - p))^n$
$Gam(\alpha, \beta) \quad \left(\frac{1}{a - \beta t}\right)^{\alpha}$
for $t < 1/\beta$
$Pois(\lambda) \quad e^{\lambda(e^t - 1)}$

T.If X_1, \dots, X_n are indep. rvs with MGFs $M_{X_i}(t) = \mathbb{E}[e^{tX_i}]$, then the MGF of $Y = \sum_{i=1}^n a_i X_i = M_Y(t) = \prod_{i=1}^n M_{X_i}(a_i t)$.

T.Let X, Y be indep., then the p.d.f. of $Z = X + Y$ is the conv. of the p.d.f. of X and Y : $f_Z(z) = \int_{\mathbb{R}} f_X(t) f_Y(z - t) dt = \int_{\mathbb{R}} f_X(z - t) f_Y(t) dt$

D. (Normal Distribution)
$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$
$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x} - \hat{\boldsymbol{\mu}})(\mathbf{x} - \hat{\boldsymbol{\mu}})^T$
T
$P\left(\left[\hat{\mathbf{a}}_1\right] \right) = \mathcal{N}\left(\left[\hat{\mathbf{a}}_2\right] \middle \left[\hat{\mathbf{u}}_2\right], \left[\hat{\boldsymbol{\Sigma}}_{21}^{11} \quad \hat{\boldsymbol{\Sigma}}_{22}^{12}\right] \right)$
$\mathbf{a}_1, \mathbf{u}_1 \in \mathbb{R}^e, \quad \boldsymbol{\Sigma}_{11} \in \mathbb{R}^{e \times e} \text{ p.s.d.}$
$\mathbf{a}_2, \mathbf{u}_2 \in \mathbb{R}^f, \quad \boldsymbol{\Sigma}_{22} \in \mathbb{R}^{f \times f} \text{ p.s.d.}$
$\mathbf{a}_2, \mathbf{u}_2 \in \mathbb{R}^{f \times e}, \quad \boldsymbol{\Sigma}_{21} \in \mathbb{R}^{f \times e} \text{ p.s.d.}$

$P(\mathbf{a}_2 \mathbf{a}_1 = \mathbf{z}) = \mathcal{N}\left(\mathbf{a}_2 \middle \mathbf{u}_2 + \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{z} - \mathbf{u}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}\right)$
--

T. (Chebyshev) Let X be a rv with $\mathbb{E}[X] = \mu$ and variance $\text{Var}[X] = \sigma^2 < \infty$. Then for any $\epsilon > 0$, we have $P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$.

2- Analysis
Log-Trick (Identity): $\nabla_{\theta} [p_{\theta}(\mathbf{x})] = p_{\theta}(\mathbf{x}) \nabla_{\theta} [\log(p_{\theta}(\mathbf{x}))]$
T. (Cauchy-Schwarz)
$\forall \mathbf{u}, \mathbf{v} \in V$: $\langle \mathbf{u}, \mathbf{v} \rangle \leq \langle \mathbf{u}, \mathbf{v} \rangle \leq \ \mathbf{u}\ \ \mathbf{v}\ $.
$\forall \mathbf{u}, \mathbf{v} \in V$: $0 \leq \langle \mathbf{u}, \mathbf{v} \rangle \leq \ \mathbf{u}\ \ \mathbf{v}\ $.
Special case: $(\sum x_i y_i)^2 \leq (\sum x_i^2)(\sum y_i^2)$.
Special case: $\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2] \mathbb{E}[Y^2]$.

D. (Convex Set) A set $S \subseteq \mathbb{R}^d$ is called *convex* if $\forall \mathbf{x}, \mathbf{x}' \in S$, $\forall \lambda \in [0, 1]$: $\lambda \mathbf{x} + (1 - \lambda) \mathbf{x}' \in S$.

Com. Any point on the line between two points is within the set. \mathbb{R}^d is convex.

D. (Convex Function) A function $f: S \rightarrow \mathbb{R}$ defined on a *convex set* $S \subseteq \mathbb{R}^d$ is called *convex* if $\forall \mathbf{x}, \mathbf{x}' \in S, \lambda \in [0, 1]$:

$$f(\lambda \mathbf{x} + (1 - \lambda) \mathbf{x}') \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{x}')$$

Com. A function is strictly convex if the line segment between any two points on the graph of the function lies strictly above the graph. This guarantees that there is a unique global minimum.

T. (Properties of Convex Functions)

- $f(y) \geq f(x) + \nabla f(x)^T (y - x)$
- $f''(x) \geq 0$
- Local minima are global minima, strictly convex functions have a unique global minimum
- If f, g are convex then $\alpha f + \beta g$ is convex for $\alpha, \beta \geq 0$
- If f, g are convex then $\max(f, g)$ is convex
- If f is convex and g is convex and non-decreasing then $g \circ f$ is convex

D. (Strongly Convex Function) A function f is μ -strongly convex if it curves up at least as much as a quadratic function with curvature $\mu > 0$. For all x, y :

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|^2$$

Relation to Optimization:

- Guarantee:** Ensures a unique global minimum exists.
- Convergence:** Gradient Descent on strongly convex (and Lipschitz smooth) functions guarantees a **linear convergence rate** ($O(e^k)$ for some $c < 1$).

- Condition Number:** The convergence slows down on the condition number $\kappa = L/\mu$. If κ is large (poor conditioning), convergence slows down.

D. (Condition Number) The condition number $\kappa(\mathcal{A})$ measures the sensitivity of a function's output to small perturbations in the input. For a symmetric positive semi-definite matrix (like the Hessian H of a loss function), it is the ratio of the largest to the smallest eigenvalue:

$$\kappa(H) = \frac{|\lambda_{\max}|}{|\lambda_{\min}|} \geq 1$$

Implications for Optimization:

- Well-conditioned** ($\kappa \approx 1$): The contours of the loss function are nearly spherical. Gradient Descent converges quickly and directly toward the minimum.
- Ill-conditioned** ($\kappa \gg 1$): The contours form narrow, elongated ellipses (steep valleys). Gradient Descent tends to oscillate ("zigzag") across the narrow valley rather than moving down the slope, leading to very slow convergence.

Com. Momentum and Adaptive Learning Rate methods (like Adam) are specifically designed to mitigate the issues caused by high condition numbers.

T. (Taylor-Lagrange Formula)
$f(x) = \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k + \int_{x_0}^x \frac{f^{(n+1)}(x-t)}{n!} dt$
T. (Jensen) f convex/ concave , $\forall i$: $\lambda_i \geq 0, \sum_{i=1}^n \lambda_i = 1$
$f(\sum_{i=1}^n \lambda_i \mathbf{x}_i) \leq / \geq \sum_{i=1}^n \lambda_i f(\mathbf{x}_i)$
Special case: $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$.
D. (Lipschitz Continuous Function)
Given two <i>metric spaces</i> (X, d_X) and (Y, d_Y) , a function $f: X \rightarrow Y$ is called <i>Lipschitz continuous</i> , if there exists a real constant $L \in \mathbb{R}_0^+$ (<i>Lipschitz constant</i>), such that
$\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n: \quad \ f(\mathbf{x}_1) - f(\mathbf{x}_2)\ \leq L \cdot \ \mathbf{x}_1 - \mathbf{x}_2\ $.

Com. If the objective function is L -smooth a step size of $\alpha = 1/L$ guarantees convergence.

D. (Lagrangian Formulation) of arg max $_{x,y} f(x, y)$ s.t. $g(x, y) = c$: $\mathcal{L}(x, y, \gamma) = f(x, y) - \gamma(g(x, y) - c)$

D. (PL Condition) A differentiable function $f(x)$ with global minimum f^* satisfies the μ -Polyak-Lojasiewicz (PL) condition if there exists a constant $\mu > 0$ such that for all x :

$$\frac{1}{2} \|\nabla f(x)\|^2 \geq \mu(f(x) - f^*)$$

Significance:

- Gradient Dominance:** It implies that the gradient magnitude dominates the suboptimality. If the gradient is small, the function value must be close to the optimal f^* .
- Convergence without Convexity:** The PL condition is weaker than strong convexity (it does not require convexity at all). However, it is sufficient to guarantee a **linear convergence rate** for Gradient Descent.
- In Deep Learning:** Over-parameterized neural networks often satisfy the PL condition in the neighborhood of a minimum, explaining fast convergence despite non-convexity.

Convergence Rate: Gradient Descent with step size $\alpha = 1/L$ (where L is the Lipschitz constant) converges as:

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*)$$

3- Linear Algebra
D. (Group) A set G with an operation \circ is a group if it satisfies:
1. Closure: $\forall a, b \in G, a \circ b \in G$.
2. Associativity: $\forall a, b, c \in G, (a \circ b) \circ c = a \circ (b \circ c)$.
3. Identity Element: $\exists e \in G$ such that $\forall a \in G, a \circ e = a = e \circ a$.
4. Inverse Element: $\forall a \in G, \exists b \in G$ such that $a \circ b = b \circ a = e$.
Kernels are positive semi-definite matrices.
D. (Positive Semi-Definite Matrix)

A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is PSD if for all non-zero vectors $\mathbf{x} \in \mathbb{R}^n$: $\mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0$.

Properties:

- All eigenvalues $\lambda_i \geq 0$.
- The trace $\text{Tr}(\mathbf{A}) \geq 0$ and determinant $\det(\mathbf{A}) \geq 0$.
- Cholesky Decomposition exists: $\mathbf{A} = \mathbf{LL}^T$.

T. (Sylvester Criterion) A $d \times d$ matrix is positive semi-definite if and only if all the upper left $k \times k$ for $k = 1, \dots, d$ have a positive determinant.

Negative definite: $\det < 0$ for all odd-sized minors, and $\det > 0$ for all even-sized minors

Otherwise: indefinite.

D. (Trace) of $\mathbf{A} \in \mathbb{R}^{n \times n}$ is $\text{Tr}(\mathbf{A}) = \sum_{i=1}^n a_{ii}$.

Properties:

- $\text{Tr}(\mathbf{A}) = \sum_i \lambda_i$ (sum of eigenvalues).
- Cyclic property: $\text{Tr}(\mathbf{ABC}) = \text{Tr}(\mathbf{BCA}) = \text{Tr}(\mathbf{CAB})$.
- Linear: $\text{Tr}(\mathbf{A} + \mathbf{B}) = \text{Tr}(\mathbf{A}) + \text{Tr}(\mathbf{B})$ and $\text{Tr}(c\mathbf{A}) = c\text{Tr}(\mathbf{A})$.
- $\text{Tr}(\mathbf{A}) = \text{Tr}(\mathbf{A}^T)$.

D. (Frobenius Norm ($\|\cdot\|_F$)) The square root of the sum of the absolute squares of its elements.

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} |A_{ij}|^2}$$

Properties:

- Relation to Trace: $\|\mathbf{A}\|_F = \sqrt{\text{Tr}(\mathbf{A}^T \mathbf{A})}$.
- Invariant under orthogonal rotations: $\|\mathbf{Q}\mathbf{A}\|_F = \|\mathbf{A}\|_F$ for orthogonal \mathbf{Q} .
- Relation to Singular Values: $\|\mathbf{A}\|_F = \sqrt{\sum_i \sigma_i^2}$.

4- Derivatives
- 4.1 — Numerator and Denominator Convent—tion
Jacobian Layout Convention Com. We use the <i>numerator-layout</i>
For a vector-valued function $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ we define

$$\left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right)_{ij} := \frac{\partial f_i}{\partial x_j}, \quad \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \in \mathbb{R}^{n \times m}.$$

Hence, gradients of scalar-valued functions are row vectors, and the chain rule takes the form

$$\frac{\partial}{\partial \mathbf{x}} [\mathbf{f}(\mathbf{g}(\mathbf{x}))] = \frac{\partial \mathbf{f}}{\partial \mathbf{g}} \frac{\partial \mathbf{g}}{\partial \mathbf{x}}$$

Remark. Sometimes the *denominator-layout* is used, where the Jacobian is defined as $(\mathbf{J}_f)_{ij} = \partial f_j / \partial x_i \in \mathbb{R}^{n \times m}$. The two conventions are related by transposition.

- 4.2 — Scalar-by-Vector
<i>Denominator Convention</i>
$\frac{\partial}{\partial \mathbf{x}} [u(\mathbf{x})v(\mathbf{x})] = u(\mathbf{x}) \frac{\partial v(\mathbf{x})}{\partial \mathbf{x}} + v(\mathbf{x}) \frac{\partial u(\mathbf{x})}{\partial \mathbf{x}}$
$\frac{\partial}{\partial \mathbf{x}} [u(v(\mathbf{x}))] = \frac{\partial u(v)}{\partial v} \frac{\partial v(\mathbf{x})}{\partial \mathbf{x}}$
$\frac{\partial}{\partial \mathbf{x}} [\mathbf{f}(\mathbf{x})^T \mathbf{g}(\mathbf{x})] = \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} \mathbf{g}(\mathbf{x}) + \frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}} \mathbf{f}(\mathbf{x}) = \mathbf{J}_f \mathbf{g}(\mathbf{x}) + \mathbf{J}_g \mathbf{f}(\mathbf{x})$
$\frac{\partial}{\partial \mathbf{x}} [\mathbf{f}(\mathbf{x})^T \mathbf{A} \mathbf{g}(\mathbf{x})] = \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} \mathbf{A} \mathbf{g}(\mathbf{x}) + \frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}} \mathbf{A}^T \mathbf{f}(\mathbf{x})$
$\frac{\partial}{\partial \mathbf{x}} [\mathbf{a}^T \mathbf{x}] = \frac{\partial}{\partial \mathbf{x}} [\mathbf{x}^T \mathbf{a}] = \mathbf{a}$
$\frac{\partial}{\partial \mathbf{x}} [\mathbf{x}^T \mathbf{x}] = 2\mathbf{x}$
$\frac{\partial}{\partial \mathbf{x}} [\mathbf{b}^T \mathbf{A} \mathbf{x}] = \mathbf{A}^T \mathbf{b}$
$\frac{\partial}{\partial \mathbf{x}} [\mathbf{x}^T \mathbf{A} \mathbf{x}] = (\mathbf{A} + \mathbf{A}^T) \mathbf{x}$

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}} [(\mathbf{A}\mathbf{x} + \mathbf{b})^T \mathbf{C}(\mathbf{D}\mathbf{x} + \mathbf{e})] &= \mathbf{D}^T \mathbf{C}^T (\mathbf{A}\mathbf{x} + \mathbf{b}) + \mathbf{A}^T \mathbf{C}(\mathbf{D}\mathbf{x} + \mathbf{e}) \\ \frac{\partial}{\partial \mathbf{x}} [\|\mathbf{f}(\mathbf{x})\|_2^2] &= \frac{\partial}{\partial \mathbf{x}} [\mathbf{f}(\mathbf{x})^T \mathbf{f}(\mathbf{x})] = 2 \frac{\partial}{\partial \mathbf{x}} [\mathbf{f}(\mathbf{x})] \mathbf{f}(\mathbf{x}) = 2 \mathbf{J}_f \mathbf{f}(\mathbf{x}) \end{aligned}$$

- 4.3 — Vector-by-Vector

$\mathbf{A}, \mathbf{C}, \mathbf{D}, \mathbf{a}, \mathbf{b}, \mathbf{e}$ not a function of \mathbf{x} ,

$\mathbf{f} = \mathbf{f}(\mathbf{x}), \mathbf{g} = \mathbf{g}(\mathbf{x}), \mathbf{h} = \mathbf{h}(\mathbf{x}), u = u(\mathbf{x}), v = v(\mathbf{x})$

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}} [u(\mathbf{x})\mathbf{f}(\mathbf{x})] &= u(\mathbf{x}) \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} + \mathbf{f}(\mathbf{x}) \frac{\partial u(\mathbf{x})}{\partial \mathbf{x}} \\ \frac{\partial}{\partial \mathbf{x}} [\mathbf{x} \circ \mathbf{a}] &= \text{diag}(\mathbf{a}) \\ \frac{\partial}{\partial \mathbf{x}} [\mathbf{a}] &= \mathbf{0} \\ \frac{\partial}{\partial \mathbf{x}} [\mathbf{x}] &= \mathbf{I} \\ \frac{\partial}{\partial \mathbf{x}} [\mathbf{A}\mathbf{x}] &= \mathbf{A} \\ \frac{\partial}{\partial \mathbf{x}} [\mathbf{x}^T \mathbf{A}] &= \mathbf{A}^T \\ \frac{\partial}{\partial \mathbf{x}} [\mathbf{a}\mathbf{f}(\mathbf{x})] &= a \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} = a \mathbf{J}_f \\ \frac{\partial}{\partial \mathbf{x}} [\mathbf{A}\mathbf{f}(\mathbf{x})] &= \mathbf{A} \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} \end{aligned} \quad \left| \quad \begin{aligned} \frac{\partial}{\partial \mathbf{x}} [\mathbf{f}(\mathbf{g}(\mathbf{x}))] &= \frac{\partial \mathbf{f}}{\partial \mathbf{g}} \frac{\partial \mathbf{g}}{\partial \mathbf{x}} = \mathbf{J}_f(\mathbf{g}) \mathbf{J}_g(\mathbf{x}) \\ \frac{\partial}{\partial \mathbf{x}} [\mathbf{f}(\mathbf{h}(\mathbf{x}))] &= \frac{\partial \mathbf{f}}{\partial \mathbf{g}} (\mathbf{g}(\mathbf{h})) \frac{\partial \mathbf{h}}{\partial \mathbf{x}} \\ \frac{\partial}{\partial \mathbf{x}} [\mathbf{a}^T \mathbf{x}^T \mathbf{b}] &= \mathbf{b}^T \\ \frac{\partial}{\partial \mathbf{x}} [\mathbf{a}^T \mathbf{x}^T \mathbf{b}] &= \mathbf{X}(\mathbf{a}\mathbf{b}^T + \mathbf{b}\mathbf{a}^T) \end{aligned} \right.$$

- 4.4 — Scalar-by-Matrix

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}} [\mathbf{a}^T \mathbf{X} \mathbf{b}] &= \mathbf{a}\mathbf{b}^T \\ \frac{\partial}{\partial \mathbf{x}} [\mathbf{a}^T \mathbf{X}^T \mathbf{b}] &= \mathbf{a}\mathbf{b}^T \\ \frac{\partial}{\partial \mathbf{x}} [\mathbf{a}^T \mathbf{X}^T \mathbf{X} \mathbf{b}] &= \mathbf{X}(\mathbf{a}\mathbf{b}^T + \mathbf{b}\mathbf{a}^T) \end{aligned} \quad \left| \quad \begin{aligned} \frac{\partial}{\partial \mathbf{x}} [\text{Tr}(\mathbf{X})] &= \mathbf{I} \\ \frac{\partial}{\partial \mathbf{x}} [\text{Tr}(\mathbf{AXB})] &= \mathbf{A}^T \mathbf{B}^T \\ \frac{\partial}{\partial \mathbf{x}} [\text{Tr}(\mathbf{AX}^T \mathbf{B})] &= \mathbf{B}\mathbf{A} \end{aligned} \right.$$

- 4.5 — Vector-by-Matrix (Generalized Gradient)

5- Information Theory
D. (Entropy) Let \mathcal{X} be a random variable distributed according to $p(\mathcal{X})$. Then the entropy of \mathcal{X}

$$H(\mathcal{X}) = \mathbb{E}[-\log(P(\mathcal{X}))] = -\sum_{\mathbf{x} \in \mathcal{X}} p(\mathbf{x}) \log(p(\mathbf{x})) \geq 0.$$

It describes the expected information content $I(\mathcal{X})$ of \mathcal{X} .

D. (Cross-Entropy) For distributions p and q over a given set is $H(p, q) = -\sum_{\mathbf{x} \in \mathcal{X}} p(\mathbf{x}) \log(q(\mathbf{x})) = \mathbb{E}_{\mathbf{x} \sim p} [-\log(q(\mathbf{x}))] \geq 0$.

$H(\mathcal{X}; p, q) = H(\mathcal{X}) + KL(p, q) \geq 0$, where H uses p .

Com. The minimizer of the cross-entropy is $q := p$, due to the second formulation.

Com. Usually, q is the approximation of the unknown p .

D. (Kullback-Leibler Divergence)

For probability distributions p and q defined on the same probability space, the KL-divergence between p and q is defined as

$$\begin{aligned} KL(p, q) &= -\sum_{\mathbf{x} \in \mathcal{X}} p(\mathbf{x}) \log\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right) \\ &= \sum_{\mathbf{x} \in \mathcal{X}} p(\mathbf{x}) \log\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) \geq 0. \end{aligned}$$

$$\begin{aligned} KL(p, q) &= -\mathbb{E}_{\mathbf{x} \sim p} \left[\log\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right) \right] = \mathbb{E}_{\mathbf{x} \sim p} \left[\log\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) \right] \geq 0. \\ KL(\mathcal{X}; p, q) &= H(p, q) - H(\mathcal{X}), \quad \text{where } H \text{ uses } p. \end{aligned}$$

The KL-divergence is defined only if $\forall \mathbf{x}$: $q(\mathbf{x}) = 0 \implies p(\mathbf{x}) = 0$ (absolute continuity). Whenever $p(\mathbf{x})$ is zero the contribution of the corresponding term is interpreted as zero because $\lim_{x \rightarrow 0+} x \log(x) = 0$.

In ML it is a measure of the amount of information lost, when q (model) is used to approximate p (true).

Com. $KL(p, y) = 0 \iff p \equiv q$.

Com. Note that the KL-divergence is not symmetric!

D. (Jensen-Shannon Divergence)
$JSD(P, Q) = \frac{1}{2}KL(P, M) + \frac{1}{2}KL(Q, M) \in [0, \log(n)]$
$M = \frac{1}{2}(P + Q)$
C. The JSD is symmetric!
Com. The JSD is a symmetrized and smoothed version of the KL-divergence.

6- Activation Functions
Activation functions are non-linear functions that are applied element-wise to the output of a linear function.

D. (Tanh)
$\tanh(\mathbf{x}) = \frac{e^{\mathbf{x}} - e^{-\mathbf{x}}}{e^{\mathbf{x}} + e^{-\mathbf{x}}} \quad \tanh'(\mathbf{x}) = 1 - \tanh^2(\mathbf{x})$
$\text{ReLU}(x) = \max(x, 0) \quad \text{ReLU}'(x) = \begin{cases} 0 & x \leq 0 \\ 1 & x > 0 \end{cases}$

D. (Sigmoid/Logistic)

11.3—Gradient Descent
D. (Gradient Descent (GD)) Iteratively moves parameters θ in the direction of the negative gradient of the loss function $J(\theta)$.

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta} J(\theta_t)$$

Com. In Stochastic GD (SGD), the gradient is approximated using a single sample (or mini-batch) to introduce noise and escape local minima. Although, the gradient is unbiased it adds variance, this can help to escape local minima and saddle points.

Com. Gradient Flow can be seen as the numerical integration of the continuous-time ordinary differential equation (ODE) $\dot{x} = -\nabla f(x)$.

D. (Polyak Averaging (Averaged SGD)) Instead of using the final parameter vector θ_T , this method uses the arithmetic mean of the parameters traversed during training.

$$\bar{\theta}_t = \frac{1}{t} \sum_{i=1}^t \theta_i \quad \text{or} \quad \bar{\theta}_t = (1 - \beta)\bar{\theta}_{t-1} + \beta\theta_t$$

Benefits:

- It effectively increases the effective batch size and reduces the variance of the estimate.
- Allows the use of larger learning rates (longer steps) while still converging to the optimal solution asymptotically.
- Often achieves the optimal convergence rate of $O(1/t)$ for convex problems.

D. (Learning Rate Condition) (Robbins-Monro Conditions), for Stochastic Gradient Descent (SGD) to guarantee convergence to a local minimum (in non-convex cases) or global minimum (in convex cases), the step size schedule α_t must satisfy two conditions:

1.Explore Forever: The steps must sum to infinity to ensure the algorithm can reach the optimum from any starting point, no matter how far.

$$\sum_{t=1}^{\infty} \alpha_t = \infty$$

2.Decay Fast Enough: The squared steps must sum to a finite value to ensure the variance (noise) of the updates tends to zero, preventing the parameters from oscillating forever around the minimum.

$$\sum_{t=1}^{\infty} \alpha_t^2 < \infty$$

Example: A schedule of $\alpha_t = \frac{1}{t}$ satisfies both, whereas $\alpha_t = \frac{1}{\sqrt{t}}$. satisfies the first but not the second.

D. (Momentum) Accelerates SGD by navigating along the relevant direction and softening oscillations in irrelevant directions. It maintains a velocity vector v (exponential moving average of past gradients).

$$\theta^{t+1} = \theta^t - \eta \nabla J(\theta^t) + \beta(\theta^t - \theta^{t-1})$$

where $\beta \in [0, 1]$ is the momentum term (friction).

D. (Nesterov Accelerated Gradient) A "look ahead" version of GD, also called NAG. It computes the gradient at the *approximate future position* of the parameters rather than the current position.

$$\begin{aligned} \theta^{t+1} &= \theta^t + \beta(\theta^t - \theta^{t-1}) \\ \theta^{t+1} &= \theta^{t+1} - \eta \nabla f(\theta^{t+1}) \end{aligned}$$

D. (Adaptive Learning Rate Methods) These methods adjust the learning rate for each parameter individually, scaling them based on the history of gradient magnitudes.

D. (RMSProp) Designed to resolve the diminishing learning rates of Adagrad. It uses a decaying average of squared gradients.

$$\begin{aligned} E[g^2]_t &= \beta E[g^2]_{t-1} + (1 - \beta)(\nabla J(\theta_t))^2 \\ \theta_{t+1} &= \theta_t - \frac{\alpha}{\sqrt{E[g^2]_t + \epsilon}} \nabla J(\theta_t) \end{aligned}$$

D. (Adam (Adaptive Moment Estimation)) Combines Momentum (first moment m_t) and RMSProp (second moment v_t). It also includes bias correction terms \hat{m}_t, \hat{v}_t to account for initialization at zero.

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1 - \beta_1) \nabla J(\theta_t) \\ v_t &= \beta_2 v_{t-1} + (1 - \beta_2) (\nabla J(\theta_t))^2 \\ \theta_{t+1} &= \theta_t - \frac{\alpha}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t \end{aligned}$$

12. Convolutional Neural Networks

— 12.1—Convolutional Layers

D. (Transform) A transform T is a mapping from one function space \mathcal{F} to another function space \mathcal{F}' . So $T: \mathcal{F} \rightarrow \mathcal{F}'$.

D. (Linear Transform) A transform T is linear, if for all functions f, g and scalars α, β , $T(\alpha f + \beta g) = \alpha(Tf) + \beta(Tg)$.

D. (Integral Transform) An *integral transform* is any transform T of the following form

$$(Tf)(u) = \int_{t_1}^{t_2} K(t, u) f(t) \, dt.$$

Com. The Fourier transform is an example of an integral transform.

📌 Any integral transform is a linear transform.

D. (Convolution) Given two functions $f, h: \mathbb{R} \rightarrow \mathbb{R}$, their convolution is defined as

$$(f * h)(u) := \int_{-\infty}^{\infty} h(t) f(u - t) \, dt = \int_{-\infty}^{\infty} h(u - t) f(t) \, dt$$

Com. Whether the convolution exists depends on the properties of f and h (the integral might diverge).

However, a typical use is $f = \text{signal}$, and $h = \text{fast decaying kernel function}$.

T. (Convolution Theorem) Any linear, translation-invariant transformation T can be written as a *convolution* with a suitable h .

T. (Convs are commutative and associative)

T. (Convs are shift-invariant), we define $f_{\Delta}(t) := f(t + \Delta)$. Then

$$(f_{\Delta} * h)(u) = (f * h)_{\Delta}(u)$$

D. (Fourier Transform) The Fourier transform of a function f is defined as

$$(Ff)(u) := \int_{-\infty}^{\infty} f(t) e^{-2\pi i u t} \, dt$$

and its inverse as

$$(F^{-1}f)(u) := \int_{-\infty}^{\infty} f(t) e^{2\pi i u t} \, dt$$

Com. Convolutional operators can be efficiently computed with point wise multiplication using the Fourier transform.

$$F(f * h) = Ff \cdot Fh$$

and then transformed back using the inverse Fourier transform.

$$F^{-1}(F(f * h)) = F^{-1}(Ff \cdot Fh) = f * h$$

— 12.2—Discrete Time Convolutions

D. (Discrete Convolution)

For $f, h: \mathbb{Z} \rightarrow \mathbb{R}$, we can define the discrete convolution via

$$(f * h)[u] := \sum_{t=-\infty}^{\infty} f[t] h[u - t] = \sum_{t=-\infty}^{\infty} f[u - t] h[t]$$

Com. Note that the use of rectangular brackets suggests that we're using "arrays" (discrete-time samples).

Com. Typically we use a h with finite support (window size).

D. (Multidimensional Discrete Convolution)

For $f, h: \mathbb{R}^d \rightarrow \mathbb{R}$ we have

$$(f * h)[u_1, \dots, u_d] =$$

$$\sum_{t_1=-\infty}^{\infty} \cdots \sum_{t_d=-\infty}^{\infty} f(t_1, \dots, t_d) h(u_1 - t_1, \dots, u_d - t_d) = \sum_{t_1=-\infty}^{\infty} \cdots \sum_{t_d=-\infty}^{\infty} f(u_1 - t_1, \dots, u_d - t_d) h(t_1, \dots, t_d)$$

D. (Discrete Cross-Correlation)

Let $f, h: \mathbb{Z} \rightarrow \mathbb{R}$, then

$$\begin{aligned} (h * f)[u] &:= \sum_{t=-\infty}^{\infty} h[t] f[u + t] = \sum_{t=-\infty}^{\infty} h[-t] f[u - t] \\ (\bar{h} * f)[u] &= (f * \bar{h})[u] \quad \text{where } \bar{h}(t) = h(-t). \end{aligned}$$

aka "sliding inner product", non-commutative, kernel "flipped over" ($u + t$ instead of $u - t$). If kernel symmetric: cross-correlation = convolution.

— 12.3—Convolution via Matrices

Represent the input signal, the kernel and the output as *vectors*. Copy the kernel as columns into the matrix offsetting it by one more very time (gives a band matrix (special case of Toeplitz matrix)). Then the convolution is just a matrix-vector product.

— 12.4—Border Handling

There are different options to do this

- D. (Padding of p)** Means we extend the image (or each dimension) by p on both sides (so $+2p$) and just fill in a constant there (e.g., zero).
- D. (Same Padding)** Padding with zeros = *same padding* ("same" constant, i.e., 0, and we'll get a tensor of the "same" dimensions)
- D. (Valid Padding)** Only retain values from windows that are fully-contained within the support of the signal f (see 2D example below) = *valid padding*

— 12.5—Backpropagation for Convolutions

D. (Receptive Field \mathcal{I}_l^i of x_l^i)

The *receptive field* \mathcal{I}_l^i of node x_l^i is defined as $\mathcal{I}_l^i := \{j \mid W_{ji}^l \neq 0\}$ where W^l is the Toeplitz matrix of the convolution at layer l .

Com. Hence, the receptive field of a node x_l^i are just nodes the which are connected to it and have a non-zero weight.

Com. One may extend the definition of the receptive field over several layers. The further we go back in layer, the bigger the receptive field becomes due to the nested convolutions. The receptive field may be even the entire image after a few layers. Hence, the convolutions have to be small.

D. (Recursive Receptive Field Formula) The receptive field size r_l of layer l is calculated as:

$$r_l = r_{l-1} + (k_l - 1) \times j_{l-1}$$

Where:

- r_l : Receptive field size of the current layer.
- r_{l-1} : Receptive field size of the previous layer.
- k_l : Kernel size of the current layer.
- j_{l-1} : Cumulative stride product of all previous layers ($j_{j-1} = \prod_{i=1}^{j-1} s_i$).

We have $\forall j \neq \mathcal{I}_l^i: \frac{\partial \mathcal{L}}{\partial x_j^l} = 0$,

Due to *weight-sharing*, the kernel weight h_o^j is re-used for every unit in the target layer at layer l , so when computing the derivative $\frac{\partial \mathcal{R}}{\partial h_o^l}$ we just build an additive combination of all the derivatives (note that some of them might be zero).

$$\frac{\partial \mathcal{R}}{\partial h_o^l} = \sum_{i=1}^m \frac{\partial \mathcal{R}}{\partial x_i^l} \frac{\partial x_i^l}{\partial h_o^l}$$

Backpropagations of Convolutions as Convolutions
 $\mathbf{y}^{(l)}$ output of l -th layer $\mathbf{y}^{(l-1)}$ output of $(l-1)$ -th layer / input to l -th layer \mathbf{w} convolution filter $\frac{\partial \mathcal{R}}{\partial \mathbf{y}^{(l)}}$ known $\mathbf{y}^{(l+1)} = \mathbf{y}^{(l)} * \mathbf{w}$

$$\begin{aligned} \frac{\partial \mathcal{R}}{\partial w_i} &= \sum_k \frac{\partial \mathcal{R}}{\partial y_k^{(l)}} \frac{\partial y_k^{(l)}}{\partial w_i} = \sum_k \frac{\partial \mathcal{R}}{\partial y_k^{(l)}} \frac{\partial}{\partial y_k^{(l)}} \left[\mathbf{y}^{(l)} * \mathbf{w} \right]_k \\ &= \sum_k \frac{\partial \mathcal{R}}{\partial y_k^{(l)}} \frac{\partial}{\partial w_i} \left[\sum_{o=-p}^p y_{k-o}^{(l-1)} w_o \right] = \sum_k \frac{\partial \mathcal{R}}{\partial y_k^{(l)}} y_{k-i}^{(l-1)} \\ &= \sum_k \frac{\partial \mathcal{R}}{\partial y_k^{(l)}} y_{-(k-i)}^{(l-1)} = \sum_k \frac{\partial \mathcal{R}}{\partial y_k^{(l)}} \text{rot180}(y^{(l-1)})_{k-i} \\ &= \left(\frac{\partial \mathcal{R}}{\partial \mathbf{y}^{(l)}} * \text{rot180}(y^{(l-1)}) \right)_i \end{aligned}$$

The derivative $\frac{\partial \mathcal{R}}{\partial \mathbf{y}^{(l)}}$ is analogous.

Note that we just used generalized indices i, k, o which may be multi-dimensional.

This example omits activation functions and biases, but that could be easily included with the chain-rule.

D. (Rotation180) $\forall i$: $\text{rot180}(\mathbf{x})_i = \mathbf{x}_{(-i)}$.

— 12.6—Pooling

There are min, max, avg, and softmax pooling. Max pooling is the most frequently used one.

D. (Max-Pooling)

- 1D: $x_i^{\text{max}} = \max \{x_{i+k} \mid 0 \leq k < r\}$
- 2D: $x_{ij}^{\text{max}} = \max \{x_{i+k, j+l} \mid 0 \leq k, l < r\}$

— 12.7—Sub-Sampling (aka "Strides")

Often, it is desirable to reduce the size of the feature maps. This can be achieved by skipping some of the input values in the convolution. The stride is the number of steps the kernel takes in each direction.

— 12.8—Hyperparameters & Output Dimensions

D. (Dilation) Dilation introduces gaps between the kernel elements, effectively widening the receptive field without increasing the number of parameters.

D. (Output Dimension Formula)

$$O = \left\lfloor \frac{I + 2P - d(K - 1) - 1}{S} + 1 \right\rfloor$$

where I is the input size, K the kernel size, P padding, S stride, and d dilation.

Ex. (Convolution with $K = 4$) For a kernel of size 4, stride 1, and dilation 1, the convolution (cross-correlation style) is:

$$o(i, j) = \sum_{k=0}^3 \sum_{l=0}^3 I(i + k, j + l) K(k, l)$$

— 12.9—Channels

Ex. Here we have

- an input signal that is 2D with 3 channels (7x7x3) (image x channels)
- and we want to learn two filters $W0$ and $W1$, which each process the 3 channels, and sum the results of the convolutions across each channel leading to a tensor of size 3x3x2 (convolution result x num convolutions)



Usually we convolve over all of the channels together, such that each convolution has the information of all channels at its disposition and the order of the channels hence doesn't matter.

— 12.10—CNNs in Computer Vision

So the typical use of convolution that we have in vision is: a sequence of convolutions

- that *reduce* the spatial dimensions (sub-sampling)
- that *increase* the number of channels

The deeper we go in the network, we transform the spatial information into a semantic representation. Usually, most of the parameters lie in the fully connected layers

— 12.10.1—Classic CNN Architectures

D. (LeNet-5 (1998)) The pioneering CNN for handwritten digit recognition (MNIST).

- Structure:** 2 Convolutional layers (with Average Pooling) followed by 3 Fully Connected layers.
- Key Features:** Introduced the concepts of local receptive fields, shared weights, and spatial subsampling. Used Sigmoid/Tanh activations (pre-ReLU).

D. (AlexNet (2012)) The breakthrough model that popularized Deep Learning on ImageNet.

- Structure:** Deeper than LeNet (5 Conv layers, 3 FC layers). Used large filters initially (11 x 11).
- Innovations:** First large-scale use of ReLU (to solve vanishing gradients), **Dropout** (for regularization), and **Data Augmentation**. Trained on GPUs.

D. (VGG Network (2014)) Focused on the effect of network depth using a uniform architecture.

- Philosophy:** Replace large filters (e.g., 5 x 5, 7 x 7) with stacks of small 3 x 3 **filters**.
- Reasoning:** Two stacked 3 x 3 layers have the same receptive field as a 5 x 5 layer but with fewer parameters and more non-linearities (ReLU between layers).

D. (Inception Network) Focused on computational efficiency and network "width". It was developed by Google in 2014.

- Inception Module:** Instead of choosing a filter size, it performs 1 x 1, 3 x 3, and 5 x 5 convolutions (and pooling) *in parallel* and concatenates the outputs.
- 1 x 1 **Convolutions:** Used as "Bottleneck layers" to reduce dimensionality (depth) before expensive operations, significantly reducing computational cost.

D. (U-Net (2015)) Designed for Biomedical Image Segmentation (pixel-wise classification).

- Structure:** Symmetrical Encoder-Decoder architecture (U-shape).
- Encoder:** Contracting path (Convs + Max Pooling) to capture context.
- Decoder:** Expansive path (Up-Convs) to enable precise localization.
- Skip Connections:** Concatenates high-resolution features from the encoder directly to the decoder to recover spatial details lost during downsampling.

— 12.10.2—Convolutions in Sequences, NLP & Audio

D. (1D Convolutions) Unlike 2D CNNs for images, sequence modeling (Temporal ConvNets) uses 1D filters that slide over the time axis.

- Input:** Tensor of shape *(Batch, Length, Channels)*. In NLP, "Channels" are the dimensions of the Word Embeddings.
- Function:** Captures local temporal dependencies (like n -grams in text) effectively.
- Advantage:** Highly parallelizable (unlike RNNs which are sequential) and computationally efficient.

D. (Embeddings & NLP) CNNs are often applied on top of pre-trained word embeddings (e.g., Word2Vec, GloVe).

- A sentence is represented as a matrix (Length x Embedding Dim).
- Filters of different widths (e.g., covering 2, 3, or 4 words) act as feature detectors for phrases or specific linguistic patterns (e.g., "very good", "not bad") regardless of their position in the sentence.

D. (Dilated Convolutions) To handle long sequences without losing resolution (pooling), *dilation* introduces gaps between kernel elements.

- Receptive Field:** Grows exponentially with the dilation factor d (1, 2, 4, 8, ...), allowing the network to capture long-range dependencies with few layers.

D. (WaveNet (2016)) A deep generative model for raw audio waveforms.

- Causal Convolutions:** Strict ordering ensures the prediction at time t only depends on samples $x_{<t}$ (cannot see the future).
- Dilated Causal Convolutions:** Stacks layers with increasing dilation factors. This allows the output neuron to have a receptive field of thousands of timesteps (milliseconds of audio) to generate realistic high-fidelity sound structure.
- Skip Connections:** Uses residual and parameterized skip connections to speed up convergence and allow training of very deep networks.

— 12.11—Comparison of #Parameters (CNNs, FC, LC)

Ex: input image $m \times n \times c$ (c = number of channels)

K convolution kernels: $p \times q$ (valid padding and stride 1) output dimensions: $(m - p + 1) \times (n - q + 1) \times K$

#parameters CNN: $K(pqc + 1)$

#parameters of fully-conn. NN with same number of outputs as CNN:

$mnc((m - p + 1)(n - q + 1) + 1)K$

#parameters of locally-conn. NN with same connections as CNN:

$pqc((m - p + 1)(n - q + 1) + 1)K$

13 Recurrent Neural Networks

— 13.1—Simple Recurrent Networks

D. (Concept) Unlike CNNs (fixed filter widths), RNNs model temporal/sequence data of variable length. They maintain a hidden state z_t acting as a "memory" of the history.

Formulation: Given input sequence x_1, \dots, x_T :

$$z_t = \phi(Uz_{t-1} + Vz_t)$$

$$\hat{y}_t = \psi(Wz_t)$$

where U, V, W are shared weight matrices and ϕ, ψ are non-linearities.

Unrolling: An RNN is equivalent to a deep feedforward network with T layers and *shared weights*.

Backpropagation Through Time (BPTT): Gradients are propagated backward through the unrolled graph. Since weights are shared, the total gradient is the sum of gradients at each time step:

$$\frac{\partial \mathcal{L}}{\partial w} = \sum_{t=1}^T \frac{\partial \mathcal{L}}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial z_t} \frac{\partial z_t}{\partial w}$$

Gradient Problems: The gradient involves repeated multiplication of the recurrent weight matrix U (specifically its Jacobian).

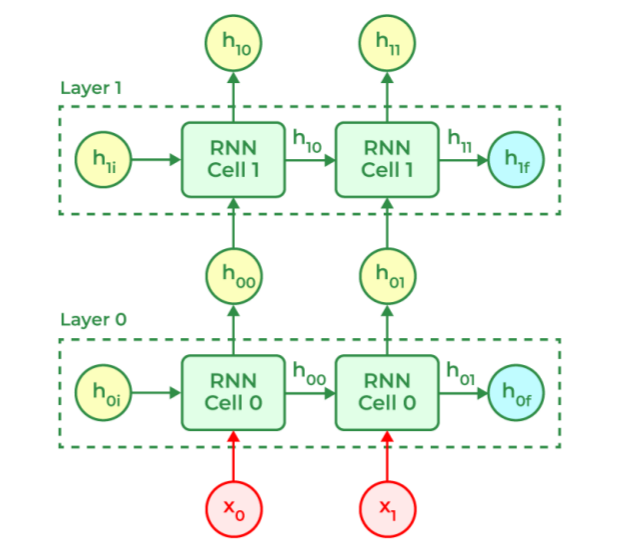
- Exploding Gradient:** If largest singular value $\sigma_{max}(U) > 1$.
- Vanishing Gradient:** If $\sigma_{max}(U) < 1$. This makes learning long-term dependencies difficult.

— 13.1.1—Structural Variants

D. (Bidirectional RNNs) Process sequence in both directions to capture past and future context.

$$\hat{y}_t = \psi(Wz_t^{\rightarrow} + \bar{W}z_t^{\leftarrow})$$

D. (Deep RNNs): Stacking multiple RNN layers to increase representational power. The output of layer l becomes the input to layer $l + 1$.



— 13.1.2—Gated Memory

Gates use multiplicative interactions (sigmoid σ) to control information flow, stabilizing gradients and allowing long-term memory.

D. (LSTMs) Long Short Term Memory models maintain a separate cell state C_t controlled by three gates.

$$C^$$

