Section #0 Notes, v 1.0
Due: never

1 Linear Algebra

1.1 Scalars and Vectors

A **scalar** is a single element of a field. For example, the real number $s \in \mathbb{R}$ is a scalar. We write scalars in lowercase.

A **vector** of n dimensions is an ordered collection of n coordinates, where each coordinate is a scalar of the underlying field. An n-dimensional vector $\mathbf{v}$ with real coordinates is an element of $\mathbb{R}^n$. Equivalently, the coordinates specify as single point in an n-dimensional space. By default, vectors will be columns and their transposes will be rows. We write vectors in bold lowercase, and the vector itself as a column of scalars:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Vectors may be scaled. $a\mathbf{x}$ scales each elements of $\mathbf{x}$ by scalar a. Vectors of the same dimension may be added coordinate-wise. Vectors have both a **direction** and a **magnitude**. The magnitude, typically the **L2 norm**, of a vector can be computed as the square root of the sum of the squares of the coordinates:

$$||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Refer to other vector norms such as the **L1**, **LP**, and **L∞** norms. Express the direction as a vector of magnitude one. Use

$$\frac{1}{||\mathbf{x}||_2}\mathbf{x}$$

.

An important product between vectors of the same dimension is the **inner product** (also called dot product or scalar product). For two vectors $\mathbf{u}$ and $\mathbf{v}$, this is equal to $\sum_{i=1}^{n} u_i v_i$. Note that a vector $\mathbf{u}$ dotted with itself equals the square of its L2 norm: $\langle \mathbf{u}, \mathbf{u} \rangle = ||\mathbf{u}||_2^2$. The **outer product** between two vectors is the matrix $\mathbf{W}$ whose entries are $w_{ij} = u_i v_j$.

1.2 Linear Independence

A set of non-zero vectors $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ is **linearly independent** if the equation $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 +, ..., +c_n\mathbf{v}_n = \mathbf{0}$ for scalars $c_1, ..., c_n$ can only be satisfied by setting $c_1, ..., c_n$ all to 0.

1.3 Spaces and Subspaces

A **vector space** $\mathcal{V}$ is a collection of vectors that follow several axioms regarding the properties of scaling and addition described above, and most importantly:

- closure under scaling: $\forall \mathbf{v} \in \mathcal{V}$ and scalars a, $a\mathbf{v} \in \mathcal{V}$
- closure under addition: $\forall \mathbf{u}, \mathbf{v} \in \mathcal{V}, (\mathbf{u} + \mathbf{v}) \in \mathcal{V}$

The set of vectors $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ form an **orthonormal basis** for $\mathcal{V}$ if they are all unit vectors (normal) and if $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0, \forall i \neq j$ (orthogonal) where $\langle , \rangle$ is the inner product. Let $\mathcal{S}$ be a vector space. If all of these hold:

- $\mathcal{S} \subseteq \mathcal{V}$
- closure under scaling: $\forall \mathbf{u} \in \mathcal{S}$ and scalars a, $a\mathbf{u} \in \mathcal{S}$
- closure under addition: $\forall \mathbf{u}, \mathbf{v} \in \mathcal{S}, (\mathbf{u} + \mathbf{v}) \in \mathcal{S}$

then $\mathcal{S}$ is a **subspace** of $\mathcal{V}$

1.4 Scalar, Vector, and Subspace Projection

For vectors $\mathbf{u}, \mathbf{v} \in \mathcal{V}$ and $\mathbf{v} \neq \mathbf{0}$, the **scalar projection** a of $\mathbf{u}$ onto $\mathbf{v}$ is computed as:

$$a = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{||\mathbf{v}||}$$

Using this, the **vector projection** $\mathbf{p}$ of $\mathbf{u}$ onto $\mathbf{v}$ can be computed as:

$$a\big(\frac{1}{||\mathbf{v}||}\mathbf{v}\big) = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle}$$

This has the properties that $\langle \mathbf{u} - \mathbf{p}, \mathbf{p} \rangle = \mathbf{0}$ and $\mathbf{u} = \mathbf{p}$ if and only if $\mathbf{u}$ is a scaled multiple of $\mathbf{v}$.
Finally (this is important for ML), it is possible to project a vector $\mathbf{u}$ in a vector space $\mathcal{V}$ onto a subspace $\mathcal{S}$ of $\mathcal{V}$. If the set of vectors $\{\mathbf{s}_1, ..., \mathbf{s}_m\}$ form an orthonormal basis for $\mathcal{S}$, then the **subspace projection** $\mathbf{p}$ of $\mathbf{u}$ onto $\mathcal{S}$ can be expressed as the sum of the project of $\mathbf{u}$ onto each element of the basis of $\mathcal{S}$:

$$\mathbf{p} = \sum_{i=1}^{m} \frac{\langle \mathbf{u}, \mathbf{s}_i \rangle}{\langle \mathbf{s}_i, \mathbf{s}_i \rangle} \mathbf{s}_i$$

This has the properties that the vector $\mathbf{u} - \mathbf{p}$ is orthogonal to all vectors in $\mathcal{S}$, that $\mathbf{u} = \mathbf{p}$ if and only if $\mathbf{u} \in \mathcal{S}$, and that $\mathbf{p}$ is the closest vector in $\mathcal{S}$ to $\mathbf{u}$. $||\mathbf{u} - \mathbf{v}|| > ||\mathbf{u} - \mathbf{p}||, \forall \mathbf{v} \neq \mathbf{p}, \mathbf{v} \in \mathcal{S}$. A couple of connections: reconstruction loss of dimensionality reduction and projection as conditional probability.

1.5 Matrices

A **matrix** is a rectangular array of scalars. Primarily, an $n \times m$ matrix $\mathbf{A}$ is used to describe a **linear transformation** from m to n dimensions, where the matrix is an **operator**. If the underlying field is $\mathbb{R}$, then $\mathbf{A} \in \mathbb{R}^{n \times m}$. A_{ij} is the scalar found at the i^{th} row and j^{th} column. We write matrices in bold uppercase. A typical linear transformation looks like $\mathbf{y} = \mathbf{A}\mathbf{x}$ where $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{n \times m}$. What's linear? The property that $\mathbf{A}(\lambda_1 \mathbf{u} + \lambda_2 \mathbf{v}) = \lambda_1 \mathbf{A}\mathbf{u} + \lambda_2 \mathbf{A}\mathbf{v}$ for scalars λ_1 and λ_2. These notes do not go into the generalizations of many matrix properties from $\mathbb{R}$ to $\mathbb{C}$ (e.g. transpose to conjugate transpose, symmetric to Hermitian).

1.6 Matrix Properties

- $\mathbf{A}^\top$ is the **transpose** of $\mathbf{A}$ and has $A^\top_{ji} = A_{ij}$.
- $\mathbf{A}$ is **symmetric** if $A_{ij} = A_{ji}$. That is, $\mathbf{A} = \mathbf{A}^\top$. Only square matrices can be symmetric.
- $\mathbf{A}$ is said to be **orthogonal** if its rows and its columns are orthogonal unit vectors. Consequence: $\mathbf{A}^\top \mathbf{A} = \mathbf{A}\mathbf{A}^\top = \mathbf{I}$ where $\mathbf{I}$ is the **identity matrix** (ones on the main diagonal and zeros elsewhere). Orthogonal matrix $\mathbf{A}$ has $\mathbf{A}^\top = \mathbf{A}^{-1}$.

- **Diagonal** matrices have non-zero values on the main diagonal and zeros elsewhere. Diagonal matrices are easy to take powers of. Under certain conditions a matrix may be diagonalized, see eigendecomposition and SVD below.
- A matrix is **upper-triangular** if the only non-zero values are on the diagonal or above (top right of matrix). A matrix is **lower-triangular** if the only non-zero values are on the diagonal or below (bottom right of matrix).

1.7 Matrix Multiplication Properties

AB is a valid **matrix product** if **A** is $p \times q$ and **B** is $q \times r$ (left matrix has same number of columns as right matrix has rows). There are many others important matrix products, such as the element-wise **Hadamard** product $\mathbf{A} \odot \mathbf{B}$ between matrices of the same shape. The standard matrix product corresponds to the composition of operators.

- Generally not commutative: $\mathbf{AB} \neq \mathbf{BA}$
- Left/Right Distributive over addition: $\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC}$. $(\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC}$.
- For some scalar λ: $\lambda(\mathbf{AB}) = (\lambda\mathbf{A})\mathbf{B}$ and $(\mathbf{AB})\lambda = \mathbf{A}(\mathbf{B}\lambda)$, and all four are equal if λ is real or complex.
- transpose of product: $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$

1.8 Rank, Determinant, Inverse

The rank of a matrix **A** is the **dimension** of the vector space spanned by its column vectors. A matrix is full rank if all its column vectors are linearly independent. The same holds for row vectors. If **A** is $n \times m$, then $rank(A) \leq min(n, m)$.

The **determinant** of a square matrix is a scalar quantity with various uses. Its computation differs for square matrices of different sizes. The existence of a matrix inverse depends on a non-zero determinant. $det(\mathbf{A})$ is also the product of the eigenvalues of **A**. You may see the determinant denoted with single bars, e.g. $|\mathbf{X}|$. However, the author (mark goldstein) prefers $det(\mathbf{X})$. Do not confuse $|\mathbf{X}|$ with double bars $||\mathbf{X}||$, which typically denote a norm.

The **inverse** $\mathbf{A}^{-1}$ of matrix operator **A** "undoes" **A** much like multiplying by $\frac{1}{x}$ undoes multiplying by x. $\mathbf{A}^{-1}$ only exists if $det(\mathbf{A}) \neq 0$. In general, matrix inversion is a complicated operation, but special cases that are easy to work with come up in the machine learning literature. Often analytical solutions to systems depend on the existence of the inverse of a matrix. $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.

The **Moore-Penrose pseudoinverse** $\mathbf{A}^+$ of **A** is a generalization of the inverse to non-square matrices, where $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$. $\mathbf{A}\mathbf{A}^+$ may not be the general identity matrix but maps all column vectors of **A** to themselves.

1.9 Eigen-Everything

Each linear operator (matrix) has some set of vectors in its domain that are simply mapped to a scaled version of the vector in the codomain. The operator preserves the direction of these vectors: $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ for some scalar value λ. In this case, λ is an **eigenvalue** of **A** and **x** is a corresponding **eigenvector**. These can be seen as the *invariant directions* of the operator.

The eigenvectors of the empirical covariance matrix of some data correspond to the directions of variance in the data. The eigenvectors with the largest associated eigenvalues correspond to the directions of highest

variance in the data (by construction). These directions may be linear combinations of the original coordinates (features) of the data. See **Principal Component Analysis** (PCA) and dimensionality reduction.

Low-rank approximations reduce the dimensionality of a matrix so that it is useful for fast computation or compression. In this problem one balances the rank-minimization goal with reconstruction error between the approximation and the true matrix. Of interest also are a low **matrix norm** an **sparsity**. One example is the entrywise **Frobenius norm**.

Eigen-decomposition: Let $\mathbf{A}$ be an $n \times n$ full-rank matrix with n linearly independent eigenvectors $\{\mathbf{q}_i\}_{i=1}^n$. $\mathbf{A}$ can be factored into $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$ where $\mathbf{Q}$ is $n \times n$ and has $\mathbf{q}_i$ for its i^{th} column. $\mathbf{\Lambda}$ is a diagonal matrix whose elements are the corresponding eigenvalues: $\Lambda_{ii} = \lambda_i$. If a matrix $\mathbf{A}$ can be eigen-decomposed and none of its eigenvalues are 0, then $\mathbf{A}$ is **nonsingular** and its inverse is given by $\mathbf{A}^{-1} = \mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^{-1}$ with $\mathbf{\Lambda}_{ii}^{-1} = \frac{1}{\lambda_i}$.

Singular Value Decomposition is a useful generalization of eigen-decomposition to rectangular matrices. Let $\mathbf{A}$ be an $m \times n$ matrix. Then $\mathbf{A}$ can be factored into $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{-1}$ where

- $\mathbf{U}$ is $m \times m$ and orthogonal. The columns of $\mathbf{U}$ are the **left-singular vectors** of $\mathbf{A}$.
- $\mathbf{\Sigma}$ is an $m \times n$ diagonal matrix with non-negative real entries. The diagonal values σ_i of $\mathbf{\Sigma}$ are known as the **singular values** of $\mathbf{A}$. These are also the square roots of the eigenvalues of $\mathbf{A}^\top\mathbf{A}$.
- $\mathbf{V}$ is an $n \times n$ orthogonal matrix. The columns of $\mathbf{V}$ are the **right-singular** vectors of $\mathbf{A}$.

1.10 Positive Definiteness

The symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is said to be **positive definite** if it satisfies the property

$$\mathbf{x}^\top\mathbf{A}\mathbf{x} > 0$$

and **positive semi-definite** if it satisfies

$$\mathbf{x}^\top\mathbf{A}\mathbf{x} \geq 0$$

for every non-zero vector $\mathbf{x} \in \mathbb{R}^n$. Positive definite matrices have all eigenvalues > 0 and positive semi-definite matrices have all eigenvalues ≥ 0.

1.11 Cholesky Decomposition

A symmetric positive-definite matrix may be factorized into a lower triangular matrix and its transpose. This is very useful for numerical efficiency. Example: We want to solve $\mathbf{A}\mathbf{x} = \mathbf{b}$ for $\mathbf{x}$. $\mathbf{A}$ is real, symmetric, and positive definite.

1. factor $\mathbf{A}$ into $\mathbf{L}\mathbf{L}^\top$.
2. solve $\mathbf{L}\mathbf{y} = \mathbf{b}$ for $\mathbf{y}$.
3. solve $\mathbf{L}^\top\mathbf{x} = \mathbf{y}$ for $\mathbf{x}$.

The key is that step 2 is easy to compute because $\mathbf{L}$ is lower-triangular (**forward substitution**). Step 3 is easy to compute because $\mathbf{L}^\top$ is upper-triangular (**backward substitution**). This decomposition is also very useful for simulating sampling from correlated random variables (see Cholesky Decomposition of MVN).

2 Calculus

2.1 Differentiation

You should be familiar with single-variable differentiation, including properties like:

$$\text{Chain rule: } \frac{d}{dx}f(g(x)) = f'(g(x))g'(x)$$
$$\text{Product rule: } \frac{d}{dx}f(x)g(x) = f'(x)g(x) + f(x)g'(x)$$
$$\text{Linearity: } \frac{d}{dx}(af(x) + bg(x)) = af'(x) + bg'(x)$$

for scalars a and b. In multivariable calculus, a function may have some number of inputs (say n) and some number of outputs (say m). In general, there is a partial derivative for every input-output pair. This is called the **Jacobian**. The j^{th} column is made up of the partial derivatives of f_j (the j^{th} output value of $\mathbf{f}$) with respect to all input elements, rows $i = 1$ to n.

$$\frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1(\mathbf{x})}{\partial x_n} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

If f is scalar-valued, its derivative is a column vector we call the **gradient vector** (like a single column of the Jacobian):

$$\frac{df(\mathbf{x})}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \cdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

The gradient vector points in the direction of steepest ascent in $f(\mathbf{x})$. This is useful for optimization.

The **Hessian** matrix is like the Jacobian but with second-order derivatives. There are many interesting optimization topics related to the Hessian.

A few important derivatives:

$$\frac{d\mathbf{x}^\top\mathbf{a}}{d\mathbf{x}} = \frac{d\mathbf{a}^\top\mathbf{x}}{d\mathbf{x}} = \mathbf{a}$$
$$\frac{d\mathbf{a}^\top\mathbf{X}\mathbf{b}}{d\mathbf{X}} = \mathbf{a}\mathbf{b}^\top$$
$$\frac{d\mathbf{a}^\top\mathbf{X}^\top\mathbf{b}}{d\mathbf{X}} = \mathbf{b}\mathbf{a}^\top$$
$$\frac{d\mathbf{a}^\top\mathbf{X}\mathbf{a}}{d\mathbf{X}} = \frac{d\mathbf{a}^\top\mathbf{X}^\top\mathbf{a}}{d\mathbf{X}} = \mathbf{a}\mathbf{a}^\top$$
$$\frac{d\mathbf{X}}{dX_{ij}} = \mathbf{J}^{ij} \quad ***$$

*** $\mathbf{J}$ is NOT the Jacobian, but rather, a matrix with all zeros except for a 1 in the i, j entry.

Have you ever wondered how to differentiate the norm of a matrix? The eigenvalues? For more, see the **Matrix Cookbook** by Petersen and Pedersen (linked on course website).

2.2 Optimization

Local Extrema: Recall that the local extrema of a single-variable function can be found by setting its derivative to 0. The same is true here, using the condition $\frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \mathbf{0}$. However, this equation is often intractable. We can search for local minima numerically using gradient-based methods.

Gradient Descent: We start with an initial guess at at a useful value for a parameter $\mathbf{w}$: $\mathbf{w}_0$. Then at each step i we update our guess by going in the direction of greatest descent of a loss function (opposite the direction of the gradient vector):

$$\mathbf{w}_{i+1} = \mathbf{w}_i - \eta \frac{df(\mathbf{w})}{d\mathbf{w}}$$

where η is a learning rate. We stop when the value of the gradient is close to 0.

3 Probability Theory

3.1 Random Variables

A **random variable** can either be discrete or continuous. A discrete random variable X takes one of m values from sample space $\mathcal{X}$, each with a corresponding probability $p(x)$ for $x \in \mathcal{X}$. $p(x)$ is the **probability mass function** of X and can also be written as $p_X(x)$. We say that $x \sim X$ (x is sampled from X) when the value of x is picked in accordance with the distribution of X.

A continuous random variable can take on a continuous range of values. We use $p(x)$ or $p_X(x)$ for the **probability density function** of a continuous random variable. It's important to note that the probability of any one exact value is zero. It's important to think of the function as assigning densities that behave like *relative probabilities* rather than absolute masses. Among other things, $p(x)$ can be greater than 1.

3.2 Expectation

The **expected value** (or *expectation, mean*) of a random variable can be thought of as the "weighted average" of the possible outcomes of the random variable. For discrete random variables:

$$\mathbb{E}_{x \sim p(x)}[X] = \sum_{x \in \mathcal{X}} x \cdot p(x)$$
$$\mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} f(x)p(x)$$

For continuous random variables:

$$\mathbb{E}[X] = \int_{\mathcal{X}} x \cdot p(x)dx$$
$$\mathbb{E}[f(X)] = \int_{\mathcal{X}} f(x)p(x)dx$$

The most important property of expected values is the **linearity of expectation**. For **any** two random variables X and Y (regardless of independence)

- $\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$
- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ under independence

3.3 Variance

The variance of a random variable is its expected squared deviation from its mean

$$\begin{aligned} var(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \end{aligned}$$

Variance is a measure of the spread of a random variable. High variance variables are more spread out. Consider two normal distributions, one tall and skinny, and the other shorter and wider.

$$var(aX + b) = a^2 var(X)$$

3.4 Joint Probability

The **joint probability** of $X = x$ and $Y = y$ is written as $p(x, y)$ or $p_{XY}(x, y)$. For independent random variables X and Y we have $p(x, y) = p(x)p(y)$. However, in the more general case we must **condition**: $p(x, y) = p(x)p(y|x) = p(y)p(x|y)$ (see next section). When you have a joint PMF or PDF of two or more random variables, its a common situation to want the **marginal distribution** of a single variable. For a pair of random variables X and Y, use the **sum rule**:

$$\begin{aligned} \text{Discrete: } p(x) &= \sum_{y \in \mathcal{Y}} p(x, y) \\ \text{Continuous: } p(x) &= \int_{y \in \mathcal{Y}} p(x, y) \end{aligned}$$

3.5 Conditional Probability

Receiving information about the value of a random variable Y can change the distribution of another variable X. We write the new conditional random variable as $X|Y$, and the new conditional distribution as $p(x|y)$. Manipulating the definition for the joint probability of random variables that may be dependent, we get:

$$p(x|y) = \frac{p(x, y)}{p(y)}$$

As mentioned above, when dealing with the joint probability of several dependent variables, factor into chains of conditional probabilities with the **product rule**:

$$\begin{aligned} p(x, y, z) &= p(x)p(y|x)p(z|x, y) \\ &= p(y)p(x|y)p(z|x, y) \\ &= p(z)p(x|z)p(y|x, z) \\ &= \text{etc...} \end{aligned}$$

See `http://colah.github.io/posts/2015-09-Visual-Information/` for some interesting visualizations of conditional probability and information theory.

3.6 Bayes' Theorem

This is a central theorem that we will use repeatedly in this course, and is an extension of the product rule.

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Since we are conditioning on y, y is held constant, and that means $p(y)$ is just a normalization constant. As a result, we often write the above property as

$$p(x|y) \propto p(y|x)p(x)$$

To see this concretely in terms of machine learning: say we observe data D, and we are interested in parameters $\mathbf{w}$. We can write the **posterior distribution** of the parameters given data by using Bayes' theorem.

$$\underbrace{p(\mathbf{w}|D)}_{\text{posterior}} = \frac{\overbrace{p(D|\mathbf{w})}^{\text{likelihood}}\overbrace{p(\mathbf{w})}^{\text{prior}}}{\underbrace{p(D)}_{\text{evidence}}}$$

Related to this, a maximum a posteriori (MAP) estimate for the parameter $\mathbf{w}$ is the value

$$\arg\max_{\mathbf{w}} p(\mathbf{w}|D)$$

A maximum likelihood estimate (MLE) is the value

$$\arg\max_{\mathbf{w}} p(D|\mathbf{w})$$

3.7 Covariance

The **covariance** between two jointly distributed random variables X and Y with finite variances is defined as the expected product of their deviations from their individual expected values. Intuitively, this asks: are X and Y likely to tend above $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ jointly (high covariance)? Or does X tend below $\mathbb{E}[X]$ while Y tends above $\mathbb{E}[Y]$ and vise versa (low covariance)?

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

When considering data in n dimensions, compute the $n \times n$ **covariance matrix** (often denoted $\mathbf{\Sigma}$), where $\mathbf{\Sigma}_{ij} = \text{cov}(X_i, X_j)$ is the empirical covariance between the i^{th} and j^{th} features.

Properties of covariance: (supposing X, Y, Z have mean 0 and finite variances)

- Symmetric: $\text{cov}(X, Y) = \text{cov}(Y, X)$
- Positive Semi-definite: $\text{cov}(X, X) \geq 0$
- $\text{cov}(X, X) = 0$ implies X always takes the same value, its mean
- Bilinear: $\text{cov}(aX + bY, Z) = a\text{cov}(X, Z) + b\text{cov}(Y, Z)$
- Triangle Inequality: $|\text{cov}(X, Y)| \leq \sqrt{\text{var}(X)\text{var}(Y)}$

3.8 Conditional Expectation and Conditional Variance

It is common to determine the expectation and variance of a variable. If X and Y are random variables, then $\mathbb{E}[X|Y]$ is a random variable too, because it can take on several values depending on Y. $\mathbb{E}[X|Y = y]$ is the expected (or average) value of the random variable X given a particular observed value of Y. This is called the **conditional expectation** of X given $Y = y$.

Similarly, we can define **conditional variance** as

$$var(X|Y) = \mathbb{E}[(X - \mathbb{E}[X|Y]^2)|Y] = \mathbb{E}[X^2|Y] + \mathbb{E}[X|Y]^2$$

Adam's law (law of iterated expectations) gives

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$$

Eve's Law (or law of total variance) is the analogous case for variance

$$\text{var}[X] = \mathbb{E}[\text{var}[X|Y]] + \text{var}[\mathbb{E}[X|Y]]$$

3.9 Gaussians

3.9.1 Univariate PDF

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$$

- If X, Y are independent normals then $X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$
- $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$
- Any PDF proportional to $\exp(ax^2 + bx + c)$ must be a Gaussian PDF.

3.9.2 Multivariate PDF

Given dimension m, mean vector $\mu \in \mathbb{R}^m$, and covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{m \times m}$,

$$\mathcal{N}(\mathbf{x}; \mu, \mathbf{\Sigma}) = \frac{1}{det(2\pi\mathbf{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)\right)$$

Note: there are many ways to write the MVN PDF. You may notice the absence of m in the coefficient. This works because the 2π distributes nicely over $\mathbf{\Sigma}$ in the determinant.

3.10 Change of Random Variables

Suppose you have a random variable X that takes values from a set $\mathcal{S}$ and has a known PDF $f_X(x)$. Let $Y = r(X)$ for an arbitrary function $r : \mathcal{S} \to \mathcal{T}$. Let $\mathcal{B} \subseteq \mathcal{T}$ be the image of $r(X)$, then $r^{-1}(\mathcal{B}) = \{x \in \mathcal{S} : r(x) \in \mathcal{B}\}$ is the inverse image of $\mathcal{B}$ under r. What is the PDF of Y? This is in general a difficult problem. Let's cover a few simple cases.

X and Y are discrete. Reverse-engineering the problem, we can write down something like:

$$f_Y(y) = \begin{cases} \sum_{x \in r^{-1}(y)} f_X(x) & \text{if } y \in \mathcal{T} \\ 0 & \text{otherwise} \end{cases}$$

where $r^{-1}(y)$ is a slight abuse of notation to indicate the set of xs $\in \mathcal{S}$ that are mapped by r to the single value $y \in \mathcal{T}$. This could work also for a continuous-to-discrete transformation r, where you sweep across the areas of the original space $\mathcal{S}$ that map to y, and weigh them by their density. The main problem here is finding the set $r^{-1}(y)$.

What if the transformation is from continuous to continuous? When r is one-to-one and smooth, the **Change of Variables Formula** uses the CDF of X to define the PDF of Y (steps skipped here, result is shown). If r is strictly increasing or strictly decreasing on $\mathcal{S}$, then

$$f_y(y) = f_x[r^{-1}(y)]|\frac{d}{dy}r^{-1}(y)|$$

And the multivariable generalization is

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x})|det(\frac{d\mathbf{x}}{d\mathbf{y}})|$$

where the $\mathbf{x} = r^{-1}(\mathbf{y})$ as above. The term is the determinant of the Jacobian (first derivative) matrix of the inverse of the transformation. This seems complex, but luckily there are lots of special, easier, common cases, such as **linear transformations of random variables**. Examples:

- **Stretch + Shift**. Let X be a random variable with $x \in \mathcal{S} \subseteq \mathbb{R}$ (an interval). Let $a, b \in \mathbb{R} \setminus \{0\}$. Let Y be a random variable with $y \in \mathcal{T} = \{a + bx : x \in \mathcal{S}\}$.

$$f_Y(y) = \frac{1}{|b|} f_X(\frac{y-a}{b}), y \in \mathcal{T}$$

- **Multivariable Stretch + Shift**. Let X be an R.V. with $x \in \mathcal{S} \subseteq \mathbb{R}^n$. Let $Y = \mathbf{a} + \mathbf{B}X$ where $\mathbf{a} \in \mathbb{R}^n$ and $\mathbf{B} \in \mathbb{R}^{n \times n}$ and invertible. $\mathbf{y} \in \mathcal{T} = \{\mathbf{a} + \mathbf{B}x : x \in \mathcal{S}\}$.

$$f_Y(\mathbf{y}) = \frac{1}{|det(\mathbf{B})|} f_X[\mathbf{B}^{-1}(\mathbf{y} - \mathbf{a})], \mathbf{y} \in \mathcal{T}$$

- **Sum of Two Random Variables:** Let X and Y be random variables that take on values from the sets R and S respectively (subsets of $\mathbb{R}$). Let $Z = X + Y$. We have $z \in \mathcal{T} = \{z = x + y : x \in \mathcal{R}, y \in \mathcal{S}\}$. For $z \in \mathcal{T}$, let $\mathcal{D}_z = \{x \in \mathcal{R} : z - x \in \mathcal{S}\}$. This is the set of all xs for a fixed z to which a valid y could be added. If X,Y are discrete, then:

$$f_Z(z) = \sum_{x \in \mathcal{D}_z} f_{XY}(x, z - x), z \in \mathcal{T}$$

.

For continuous, replace the sum with an integral. If independent, the PDF can be factored:

$$f_Z(z) = \sum_{x \in \mathcal{D}_z} f_X(x) f_Y(z - x), z \in \mathcal{T}$$

.

This is exactly the **convolution** of $(f_X * f_Y)(z)$.

Be sure to read this great walkthrough (with proofs) by University of Alabama: `http://www.math.uah.edu/stat/dist/Transformations.html`

3.10.1 Cholesky decomposition of Multivariate Normal

Suppose we have an MVN R.V. $X = \begin{bmatrix} X_1 & \dots & X_m \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. How can we sample from the distribution of X? One way would be to generate a set of m i.i.d. standard normals $Z_1, \dots, Z_m$, and then represent the X_i as functions of the Z_i. How to actually do this?

- Since the standard normals Z_i are independent and have unit variance, we have $\mathbb{E}[Z_i Z_j] = \mathbf{I}_{ij}$. So $\mathbb{E}[ZZ^\top] = \mathbf{I}$.
- Since the desired covariance matrix $\mathbf{\Sigma}$ is symmetric and positive definite by definition of a covariance matrix, we can decompose it into $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\top$ by Cholesky Decomposition.
- Now investigate $X = \mathbf{L}Z$ where Z is the whole vector of uncorrelated normals Z_i.
- $\mathbb{E}\big[XX^\top\big] = \mathbb{E}\big[(\mathbf{L}Z)(\mathbf{L}Z)^\top\big] = \mathbb{E}\big[\mathbf{L}ZZ^\top\mathbf{L}^\top\big] = \mathbf{L}\mathbb{E}\big[ZZ^\top\big]\mathbf{L}^\top = \mathbf{L}\mathbf{I}\mathbf{L}^\top = \mathbf{L}\mathbf{L}^\top = \mathbf{\Sigma}$
- X has the desired covariance matrix $\mathbf{\Sigma}$.$\mathbb{E}\big[XX^\top\big] = \mathbf{\Sigma}$. $\mathbf{L}$ is our mixing function from i.i.d. standard normals to the MVN X.