Mathematical Foundations for HRP: A Complete Dictionary from First Principles

Reference Guide for Portfolio Theory and Beyond

November 12, 2025

Abstract

This document provides a comprehensive, self-contained reference for all mathematical and statistical concepts needed to understand Hierarchical Risk Parity (HRP) and modern portfolio theory. Each entry includes formal definitions, intuitive explanations, examples, and connections to related concepts. This serves as both a dictionary for quick lookup and a tutorial for learning from scratch.

Contents

1 Probability and Statistics

1.1 Random Variable

Definition

A **random variable** X is a function that assigns a numerical value to each outcome of a random experiment. We write $X : \Omega \to \mathbb{R}$, where Ω is the sample space (set of all possible outcomes).

Intuition

A random variable is simply a quantity whose value depends on chance. Examples:

- X = the return of a stock tomorrow (could be any number)
- X = the number rolled on a die (1, 2, 3, 4, 5, or 6)
- X = the temperature at noon (continuous value)

Random variables come in two types:

- **Discrete**: Takes on countable values (e.g., die roll, coin flips)
- **Continuous**: Takes on any value in an interval (e.g., stock returns, temperature)

Notation

- Capital letters (X, Y, Z) denote random variables
- Lowercase letters (x, y, z) denote specific values
- $P(X = x)$ means "probability that X equals x"
- $P(X \leq x)$ means "probability that X is at most x"

1.2 Expected Value (Mean)

Definition

The **expected value** or **mean** of a random variable X is:
For discrete X:

$$\mathbb{E}[X] = \mu = \sum_i x_i P(X = x_i)$$

For continuous X:

$$\mathbb{E}[X] = \mu = \int_{-\infty}^{\infty} x f(x) dx$$

where $f(x)$ is the probability density function.

Intuition

The expected value is the long-run average if you repeat the experiment many times. It's the "center of mass" of the probability distribution.
Think of it as a weighted average where each outcome is weighted by its probability.

Example

Roll a fair six-sided die. What's the expected value?

$$\begin{aligned}\mathbb{E}[X] &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5\end{aligned}$$

Note: You can never roll 3.5, but this is the average outcome over many rolls.

Theorem

[Linearity of Expectation] For random variables X and Y and constants a, b:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

This holds even if X and Y are dependent!

1.3 Variance

Definition

The **variance** of a random variable X measures the spread or dispersion around the mean:

$$\text{Var}(X) = \sigma^2 = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Intuition

Variance measures "how spread out" the values are from the average.

- High variance: Values are widely dispersed (unpredictable)
- Low variance: Values cluster near the mean (predictable)
- Variance = 0: No randomness; X is constant

Why square the deviations? So positive and negative deviations don't cancel out, and to give more weight to large deviations.

Example

Compare two stocks:

- Stock A: Returns are always exactly 10% (variance = 0)
- Stock B: Returns are 0%, 10%, or 20% with equal probability (variance > 0)

Both have expected return 10%, but Stock B is riskier (higher variance).

1.4 Standard Deviation

Definition

The **standard deviation** is the square root of the variance:

$$\sigma = \sqrt{\text{Var}(X)} = \sqrt{\sigma^2}$$

Intuition

Standard deviation is more interpretable than variance because it's in the same units as the original data. If X is a stock return in percent, then σ is also in percent, while σ^2 is in "percent squared" (harder to interpret).
Rule of thumb for normal distributions:

- About 68% of values fall within $\mu \pm \sigma$
- About 95% of values fall within $\mu \pm 2\sigma$
- About 99.7% of values fall within $\mu \pm 3\sigma$

1.5 Covariance

Definition

The **covariance** between two random variables X and Y measures how they vary together:

$$\text{Cov}(X, Y) = \sigma_{XY} = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Intuition

Covariance tells us about the linear relationship between two variables:

- $\text{Cov}(X, Y) > 0$: When X is above its mean, Y tends to be above its mean (positive relationship)
- $\text{Cov}(X, Y) < 0$: When X is above its mean, Y tends to be below its mean (negative relationship)
- $\text{Cov}(X, Y) = 0$: No linear relationship (but could still have nonlinear relationship!)

Important: $\text{Cov}(X, X) = \text{Var}(X)$ (covariance of a variable with itself is its variance)

Example

Consider:

- X = temperature in Celsius
- Y = ice cream sales

We expect $\text{Cov}(X, Y) > 0$ because hot days (high X) typically have high ice cream sales (high Y).

Theorem

[Properties of Covariance]

1. Symmetry: $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
2. Linearity: $\text{Cov}(aX + b, Y) = a \cdot \text{Cov}(X, Y)$
3. Additivity: $\text{Cov}(X + Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z)$
4. Variance of sum: $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$

1.6 Correlation

Definition

The **correlation coefficient** between X and Y is:

$$\rho_{XY} = \text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Intuition

Correlation is covariance normalized to the range $[-1, 1]$. This makes it easier to interpret:

- $\rho = 1$: Perfect positive linear relationship ($Y = aX + b$ with $a > 0$)
- $\rho = -1$: Perfect negative linear relationship ($Y = aX + b$ with $a < 0$)
- $\rho = 0$: No linear relationship
- $|\rho|$ close to 1: Strong linear relationship
- $|\rho|$ close to 0: Weak linear relationship

Important: Correlation measures only *linear* relationships. Two variables can have correlation 0 but still be strongly related nonlinearly (e.g., $Y = X^2$).

Example

Real-world correlations:

- Height and weight: $\rho \approx 0.7$ (positive, strong)
- Price and demand: $\rho \approx -0.6$ (negative, moderate)
- Uncorrelated: Shoe size and IQ: $\rho \approx 0$

In finance:

- S&P 500 and NASDAQ: $\rho \approx 0.9$ (highly correlated)
- Stocks and bonds: $\rho \approx 0.2$ (weakly correlated)
- Gold and dollar: $\rho \approx -0.3$ (weakly negatively correlated)

1.7 Independence

Definition

Two random variables X and Y are **independent** if knowing the value of one gives no information about the other. Formally:

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y) \quad \text{for all } x, y$$

Or equivalently: $\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$

Intuition

Independence is a very strong condition. It means the variables are completely unrelated.
Key relationships:

- If X and Y are independent, then $\text{Cov}(X, Y) = 0$ and $\rho_{XY} = 0$
- But the converse is NOT true: $\text{Cov}(X, Y) = 0$ does not imply independence!
- Independence $\implies$ Uncorrelated
- Uncorrelated $\not\implies$ Independent

Example

Independent variables: Roll two dice separately. The outcomes are independent.
Dependent but uncorrelated: Let $X \sim \text{Uniform}(-1, 1)$ and $Y = X^2$. Then:

- X and Y are clearly dependent (knowing X tells you Y exactly)
- But $\text{Cov}(X, Y) = 0$ due to symmetry!
- This shows correlation only captures *linear* dependence

1.8 Normal (Gaussian) Distribution

Definition

A random variable X has a **normal distribution** with mean μ and variance σ^2, written $X \sim N(\mu, \sigma^2)$, if its probability density function is:

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

Intuition

The normal distribution is:

- Bell-shaped and symmetric around the mean
- Characterized by just two parameters: mean μ (location) and variance σ^2 (spread)
- The most important distribution in statistics due to the Central Limit Theorem
- Common in nature and finance (approximately)

The **standard normal** distribution has $\mu = 0$ and $\sigma^2 = 1$: $Z \sim N(0, 1)$.
Any normal variable can be standardized: If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$.

1.9 Sample vs. Population

Definition

- **Population**: The entire group we're interested in
- **Sample**: A subset of the population we actually observe
- **Population parameters**: True values (e.g., μ, σ^2) - usually unknown
- **Sample statistics**: Estimates computed from data (e.g., $\bar{x}$, s^2)

Notation

Quantity	Population	Sample
Mean	μ	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Variance	σ^2	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Std. Dev.	σ	$s = \sqrt{s^2}$
Covariance	σ_{XY}	$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
Correlation	ρ_{XY}	$r_{XY} = \frac{s_{XY}}{s_X s_Y}$

Intuition

Think of it this way:

- Population parameters are the "truth" we want to know
- Sample statistics are our best guesses based on limited data
- As sample size $n \to \infty$, sample statistics $\to$ population parameters

Why $n - 1$ instead of n in sample variance? This is called Bessel's correction. It makes the sample variance an unbiased estimator of the population variance. With n in the denominator, we'd systematically underestimate the true variance.

2 Linear Algebra

2.1 Vector

Definition

A **vector** is an ordered list of numbers. A vector with n components is written:

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$

This is a **column vector**. A **row vector** is written: $\mathbf{v}^T = (v_1, v_2, \ldots, v_n)$.

Intuition

Vectors represent:

- Points in space: (x, y, z) is a point in 3D space
- Directions: An arrow from origin to (x, y, z)
- Data: A portfolio weight vector $\mathbf{w} = (w_1, w_2, \ldots, w_N)$

In portfolio theory:

- Weight vector: $\mathbf{w}$ (how much in each asset)
- Return vector: $\mathbf{r}$ (return of each asset)
- Mean vector: $\boldsymbol{\mu}$ (expected return of each asset)

2.2 Matrix

Definition

A **matrix** is a rectangular array of numbers with m rows and n columns:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

This is an $m \times n$ matrix. Element a_{ij} is in row i, column j.

Intuition

Matrices represent:

- Systems of equations
- Linear transformations
- Relationships between multiple variables

In portfolio theory, the covariance matrix $\mathbf{\Sigma}$ is an $N \times N$ matrix where:

- Diagonal entries: $\sigma_{ii} = \text{Var}(r_i)$ (variance of asset i)
- Off-diagonal entries: $\sigma_{ij} = \text{Cov}(r_i, r_j)$ (covariance between assets i and j)

2.3 Matrix-Vector Multiplication

Definition

If $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{v}$ is an $n \times 1$ vector, their product $\mathbf{Av}$ is an $m \times 1$ vector:

$$(\mathbf{Av})_i = \sum_{j=1}^{n} a_{ij} v_j$$

In other words, the i-th component of $\mathbf{Av}$ is the dot product of the i-th row of $\mathbf{A}$ with $\mathbf{v}$.

Example

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \begin{pmatrix} 7 \\ 8 \end{pmatrix} = \begin{pmatrix} 1(7) + 2(8) \\ 3(7) + 4(8) \\ 5(7) + 6(8) \end{pmatrix} = \begin{pmatrix} 23 \\ 53 \\ 83 \end{pmatrix}$$

2.4 Transpose

Definition

The **transpose** of a matrix $\mathbf{A}$, denoted $\mathbf{A}^T$, is obtained by swapping rows and columns:

$$(\mathbf{A}^T)_{ij} = a_{ji}$$

If $\mathbf{A}$ is $m \times n$, then $\mathbf{A}^T$ is $n \times m$.

Intuition

Transpose "flips" the matrix along its diagonal:

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}^T = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}$$

For a vector: $\mathbf{v}^T$ changes column to row, or row to column.
Properties:

- $(\mathbf{A}^T)^T = \mathbf{A}$
- $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$ (order reverses!)
- $(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$

2.5 Inner Product (Dot Product)

Definition

The **inner product** (or **dot product**) of two vectors $\mathbf{u}$ and $\mathbf{v}$ is:

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T\mathbf{v} = \sum_{i=1}^{n} u_i v_i$$

Result is a scalar (single number).

Intuition

The inner product measures how much two vectors "point in the same direction":

- $\mathbf{u}^T\mathbf{v} > 0$: Vectors point in similar directions
- $\mathbf{u}^T\mathbf{v} < 0$: Vectors point in opposite directions
- $\mathbf{u}^T\mathbf{v} = 0$: Vectors are orthogonal (perpendicular)

In portfolio theory: $\mathbf{w}^T\boldsymbol{\mu}$ gives the portfolio expected return (weighted sum of individual returns).

2.6 Quadratic Form

Definition

A **quadratic form** is an expression of the form:

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i a_{ij} x_j$$

where $\mathbf{A}$ is an $n \times n$ matrix and $\mathbf{x}$ is an $n \times 1$ vector.

Intuition

Quadratic forms generalize the concept of ax^2 (quadratic in one variable) to multiple variables.
In portfolio theory, $\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}$ is a quadratic form that gives portfolio variance:

$$\sigma_p^2 = \mathbf{w}^T\mathbf{\Sigma}\mathbf{w} = \sum_{i=1}^{N}\sum_{j=1}^{N} w_i \sigma_{ij} w_j$$

This captures both:

- Individual variances (diagonal terms: $w_i^2\sigma_i^2$)
- Covariances (off-diagonal terms: $2w_i w_j \sigma_{ij}$ when $i \neq j$)

2.7 Symmetric Matrix

Definition

A matrix $\mathbf{A}$ is **symmetric** if $\mathbf{A} = \mathbf{A}^T$, i.e., $a_{ij} = a_{ji}$ for all i, j.

Intuition

A symmetric matrix is mirror-symmetric across its main diagonal:

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{pmatrix}$$

Covariance and correlation matrices are always symmetric because $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.
Symmetric matrices have special properties:

- All eigenvalues are real
- Eigenvectors from different eigenvalues are orthogonal
- Can be diagonalized by an orthogonal matrix

2.8 Positive Definite Matrix

Definition

A symmetric matrix $\mathbf{A}$ is **positive definite** if:

$$\mathbf{x}^T\mathbf{A}\mathbf{x} > 0 \quad \text{for all } \mathbf{x} \neq \mathbf{0}$$

It is **positive semi-definite** if $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$.

Intuition

Positive definite matrices generalize the concept of "positive numbers" to matrices.
Equivalently, $\mathbf{A}$ is positive definite if all its eigenvalues are positive.
Covariance matrices are always positive semi-definite because:

$$\text{Var}(\mathbf{a}^T\mathbf{r}) = \mathbf{a}^T\mathbf{\Sigma}\mathbf{a} \geq 0$$

(variance is never negative)
A covariance matrix is positive definite if no asset is a perfect linear combination of others (no redundancy).

2.9 Matrix Inverse

Definition

The **inverse** of a square matrix $\mathbf{A}$, denoted $\mathbf{A}^{-1}$, satisfies:

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

where $\mathbf{I}$ is the identity matrix (1's on diagonal, 0's elsewhere).
Not all matrices have inverses. $\mathbf{A}^{-1}$ exists if and only if $\mathbf{A}$ is **non-singular** (determinant $\neq 0$).

Intuition

Matrix inverse generalizes division: $\mathbf{A}^{-1}$ is like "$1/\mathbf{A}$".
To solve $\mathbf{A}\mathbf{x} = \mathbf{b}$:

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

In portfolio theory, Markowitz optimization requires computing $\mathbf{\Sigma}^{-1}$, which is where problems arise when the covariance matrix is ill-conditioned.
Computing matrix inverses is:

- Expensive: $O(N^3)$ operations for $N \times N$ matrix
- Numerically unstable if the matrix is nearly singular

2.10 Eigenvalues and Eigenvectors

Definition

For a square matrix $\mathbf{A}$, a scalar λ is an **eigenvalue** and $\mathbf{v}$ is the corresponding **eigenvector** if:

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}, \quad \mathbf{v} \neq \mathbf{0}$$

Intuition

An eigenvector is a direction that the matrix just stretches (doesn't rotate). The eigenvalue tells you the stretching factor.
Geometrically: When $\mathbf{A}$ acts on $\mathbf{v}$, it simply scales $\mathbf{v}$ by λ.
For a covariance matrix:

- Eigenvalues represent variances along principal directions
- Eigenvectors represent these principal directions
- Largest eigenvalue: direction of maximum variance
- This is the foundation of Principal Component Analysis (PCA)

Every $N \times N$ matrix has N eigenvalues (counting multiplicities, including complex ones). For symmetric matrices, all eigenvalues are real.

Example

Consider $\mathbf{A} = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$.
Eigenvalues: $\lambda_1 = 2, \lambda_2 = 3$
Eigenvectors: $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$
Check: $\mathbf{A}\mathbf{v}_1 = \begin{pmatrix} 2 \\ 0 \end{pmatrix} = 2\mathbf{v}_1$ ✓

2.11 Determinant

Definition

The **determinant** of a square matrix $\mathbf{A}$, denoted $\det(\mathbf{A})$ or $|\mathbf{A}|$, is a scalar value that encodes certain properties.
For 2×2: $\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$
For larger matrices, determinant is computed recursively (or using eigenvalues: $\det(\mathbf{A}) = \prod_i \lambda_i$).

Intuition

Determinant has several interpretations:

- Scaling factor: How much $\mathbf{A}$ scales volumes
- Invertibility: $\det(\mathbf{A}) \neq 0$ if and only if $\mathbf{A}$ is invertible
- Zero determinant: Matrix is singular (some information is lost/redundant)

For covariance matrices:

- $\det(\mathbf{\Sigma}) = 0$: Perfect multicollinearity (some assets are redundant)
- Small $\det(\mathbf{\Sigma})$: Near-multicollinearity (assets highly correlated)
- This relates to the condition number and instability

2.12 Condition Number

Definition

The **condition number** of a matrix $\mathbf{A}$ is:

$$\kappa(\mathbf{A}) = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{\text{largest eigenvalue}}{\text{smallest eigenvalue}}$$

For symmetric positive definite matrices, this measures the "shape" of the matrix.

Intuition

Condition number measures how "stretched" the matrix is in different directions:

- $\kappa(\mathbf{A}) = 1$: Matrix scales all directions equally (like a sphere)
- $\kappa(\mathbf{A}) \gg 1$: Matrix stretches some directions much more than others (like a cigar)
- $\kappa(\mathbf{A}) = \infty$: Matrix is singular (completely flattens some direction)

Why it matters for numerical stability:

- A small error in input gets amplified by up to $\kappa(\mathbf{A})$ in the output
- $\kappa(\mathbf{A}) = 10^6$ means a 0.01% input error can become a 10,000% output error!
- This is why Markowitz optimization fails when correlations are high

Rule of thumb:

- $\kappa < 100$: Well-conditioned
- $100 < \kappa < 10,000$: Moderately ill-conditioned
- $\kappa > 10,000$: Severely ill-conditioned

3 Optimization Theory

3.1 Objective Function

Definition

An **objective function** (or **cost function**) is the function we want to minimize or maximize. Written as:

$$f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$$

where $\mathbf{x} = (x_1, \ldots, x_n)$ are the decision variables.

Intuition

The objective function quantifies what we care about:

- In portfolio theory: We might minimize risk $f(\mathbf{w}) = \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w}$
- In machine learning: We might minimize prediction error
- In engineering: We might minimize cost or maximize efficiency

Maximizing $f(\mathbf{x})$ is equivalent to minimizing $-f(\mathbf{x})$, so we typically just talk about minimization.

3.2 Constraint

Definition

A **constraint** restricts the feasible values of decision variables:

- **Equality constraint**: $h(\mathbf{x}) = 0$
- **Inequality constraint**: $g(\mathbf{x}) \leq 0$

The set of all $\mathbf{x}$ satisfying all constraints is the **feasible region**.

Example

In portfolio optimization:

- Equality: $\sum_{i=1}^{N} w_i = 1$ (weights sum to 100%)
- Inequality: $w_i \geq 0$ for all i (no short selling)
- Inequality: $w_i \leq 0.1$ for all i (no more than 10% in any asset)

3.3 Unconstrained Optimization

Definition

Find $\mathbf{x}^*$ that minimizes $f(\mathbf{x})$ with no constraints:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$$

Theorem

[First-Order Condition] If f is differentiable and $\mathbf{x}^*$ is a local minimum, then:

$$\nabla f(\mathbf{x}^*) = \mathbf{0}$$

where ∇f is the gradient (vector of partial derivatives):

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

Intuition

At a minimum, the function isn't increasing in any direction, so all directional derivatives are zero (gradient is zero).
This is like the single-variable calculus condition $f'(x) = 0$ at extrema, but generalized to multiple dimensions.
Warning: $\nabla f = \mathbf{0}$ is necessary but not sufficient. The point could be a maximum or saddle point!

3.4 Lagrange Multipliers

Definition

To optimize $f(\mathbf{x})$ subject to equality constraints $h_i(\mathbf{x}) = 0$, form the **Lagrangian**:

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \sum_{i=1}^{m} \lambda_i h_i(\mathbf{x})$$

where λ_i are **Lagrange multipliers**.

Theorem

[Lagrange Multiplier Conditions] At an optimum $\mathbf{x}^*$, there exist multipliers $\boldsymbol{\lambda}^*$ such that:

$$\begin{aligned} \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) &= \mathbf{0} \quad \text{(gradient w.r.t. } \mathbf{x}\text{)} \\ h_i(\mathbf{x}^*) &= 0 \quad \text{for all } i \quad \text{(constraints)} \end{aligned}$$

Intuition

Lagrange multipliers convert a constrained problem into an unconstrained one. The multiplier λ_i represents the "shadow price" or marginal value of relaxing constraint i.
Geometrically: At the optimum, the gradient of f must be perpendicular to the constraint surface (otherwise you could move along the constraint to improve f).

Example

Minimize $f(x, y) = x^2 + y^2$ subject to $x + y = 1$.
Lagrangian: $\mathcal{L}(x, y, \lambda) = x^2 + y^2 - \lambda(x + y - 1)$
Conditions:

$$\frac{\partial \mathcal{L}}{\partial x} = 2x - \lambda = 0 \implies x = \lambda/2$$
$$\frac{\partial \mathcal{L}}{\partial y} = 2y - \lambda = 0 \implies y = \lambda/2$$
$$x + y = 1$$

Solving: $\lambda/2 + \lambda/2 = 1 \implies \lambda = 1 \implies x = y = 1/2$
Minimum value: $f(1/2, 1/2) = 1/4$

3.5 Convex Function

Definition

A function $f : \mathbb{R}^n \to \mathbb{R}$ is **convex** if for any $\mathbf{x}, \mathbf{y}$ and $0 \leq t \leq 1$:

$$f(t\mathbf{x} + (1 - t)\mathbf{y}) \leq tf(\mathbf{x}) + (1 - t)f(\mathbf{y})$$

Geometrically: The line segment connecting any two points on the graph lies above the graph.

Intuition

Convex functions are "bowl-shaped" - they curve upward.
Examples:

- Convex: x^2, e^x, $|x|$, $\mathbf{x}^T\mathbf{A}\mathbf{x}$ (if $\mathbf{A}$ positive definite)
- Not convex: $\sin(x)$, x^3, $-x^2$ (concave)

Why convexity matters:

- Convex functions have no local minima that aren't global minima
- If you find a point where $\nabla f = \mathbf{0}$, it's guaranteed to be the global minimum
- Optimization algorithms are guaranteed to converge

Portfolio variance $\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}$ is convex (since $\mathbf{\Sigma}$ is positive semi-definite), which is why Markowitz optimization is theoretically nice.

3.6 Quadratic Programming

Definition

A **quadratic program** (QP) has the form:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{c}^T\mathbf{x} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{G}\mathbf{x} \leq \mathbf{h} \end{aligned}$$

where $\mathbf{Q}$ is a symmetric matrix (objective is quadratic), and constraints are linear.

Intuition

Markowitz portfolio optimization is a quadratic program:

$$\begin{aligned} \text{minimize} \quad & \mathbf{w}^T\mathbf{\Sigma}\mathbf{w} \quad \text{(variance)} \\ \text{subject to} \quad & \mathbf{w}^T\mathbf{1} = 1 \quad \text{(weights sum to 1)} \\ & \mathbf{w}^T\boldsymbol{\mu} = \mu_{\text{target}} \quad \text{(target return)} \\ & \mathbf{w} \geq \mathbf{0} \quad \text{(no short selling)} \end{aligned}$$

QPs can be solved efficiently (though still requiring $O(N^3)$ operations), but the solution is sensitive to inputs when $\mathbf{Q}$ (here $\mathbf{\Sigma}$) is ill-conditioned.

4 Graph Theory

4.1 Graph

Definition

A **graph** $G = (V, E)$ consists of:

- V: A set of **vertices** (or **nodes**)
- E: A set of **edges** connecting pairs of vertices

An edge connecting vertices i and j is written (i, j) or $\{i, j\}$.

Intuition

Graphs represent relationships or connections:

- Social network: Vertices = people, edges = friendships
- Road network: Vertices = cities, edges = roads
- Portfolio: Vertices = assets, edges = correlations

Types of graphs:

- **Undirected**: Edges have no direction (friendship is mutual)
- **Directed**: Edges have direction (Twitter follow isn't mutual)
- **Weighted**: Edges have numerical weights (road distances, correlation strengths)

4.2 Complete Graph

Definition

A **complete graph** on N vertices has an edge between every pair of vertices. It has exactly $\binom{N}{2} = \frac{N(N-1)}{2}$ edges.

Intuition

In a complete graph, every node is directly connected to every other node.
In portfolio theory:

- A covariance matrix represents a complete graph
- Each asset (vertex) has a relationship (covariance) with every other asset
- For 50 assets: $\binom{50}{2} = 1225$ pairwise relationships!
- This complexity is part of why covariance estimation is hard

4.3 Tree

Definition

A **tree** is a connected graph with no cycles. Equivalently:

- A tree with N vertices has exactly $N - 1$ edges
- There is exactly one path between any two vertices
- Removing any edge disconnects the graph

Intuition

Trees are the simplest connected structures - like a family tree or organization chart.
Key properties:

- No redundant connections (removing any edge breaks connectivity)
- Hierarchical structure
- Very efficient: $N - 1$ edges instead of $\frac{N(N-1)}{2}$ for complete graph

For 50 assets:

- Complete graph: 1225 edges (relationships)
- Tree: 49 edges (relationships)
- This 25× reduction in complexity is why HRP is more stable!

4.4 Path

Definition

A **path** in a graph is a sequence of vertices $v_1, v_2, \ldots, v_k$ where consecutive vertices are connected by edges: $(v_1, v_2), (v_2, v_3), \ldots, (v_{k-1}, v_k) \in E$.
The **length** of a path is the number of edges.

4.5 Distance in Graphs

Definition

The **distance** between two vertices in a graph is the length of the shortest path between them.
In a weighted graph, distance is the sum of edge weights along the shortest path.

Intuition

In HRP:

- We convert correlation ρ_{ij} to distance $d_{ij} = \sqrt{\frac{1}{2}(1 - \rho_{ij})}$
- Highly correlated assets have small distance (close together)
- Weakly correlated assets have large distance (far apart)
- The hierarchical tree groups nearby (similar) assets

5 Metric Spaces and Distance Functions

5.1 Metric Space

Definition

A **metric space** is a set X with a **distance function** (or **metric**) $d : X \times X \to \mathbb{R}$ satisfying:

1. **Non-negativity**: $d(x, y) \geq 0$
2. **Identity of indiscernibles**: $d(x, y) = 0 \iff x = y$
3. **Symmetry**: $d(x, y) = d(y, x)$
4. **Triangle inequality**: $d(x, z) \leq d(x, y) + d(y, z)$

Intuition

A metric space is simply a set where you can measure distances in a consistent way. The axioms formalize our intuitive notions about distance:

1. Distances can't be negative
2. Only a point has zero distance to itself
3. Distance from A to B equals distance from B to A
4. Direct path is shortest (you can't shorten a trip by adding a detour)

Example

Common metric spaces:

- $\mathbb{R}^n$ with Euclidean distance: $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
- $\mathbb{R}^n$ with Manhattan distance: $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|$
- Assets with correlation distance: $d(i, j) = \sqrt{\frac{1}{2}(1 - \rho_{ij})}$

5.2 Euclidean Distance

Definition

In $\mathbb{R}^n$, the **Euclidean distance** between points $\mathbf{x}$ and $\mathbf{y}$ is:

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

This is the "straight-line" distance.

Intuition

Euclidean distance is what we measure with a ruler in physical space:

- In 2D: $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$ (Pythagorean theorem)
- In 3D: $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$

It generalizes naturally to any number of dimensions.

5.3 Correlation-Based Distance

Definition

For assets with correlation ρ_{ij}, the **correlation distance** is:

$$d_{ij} = \sqrt{\frac{1}{2}(1 - \rho_{ij})}$$

This satisfies the metric axioms.

Intuition

This formula converts correlation (which is NOT a distance) into a proper distance metric:

- $\rho_{ij} = 1$ (perfectly correlated) $\implies$ $d_{ij} = 0$ (zero distance, assets are "identical")
- $\rho_{ij} = 0$ (uncorrelated) $\implies$ $d_{ij} = \frac{1}{\sqrt{2}} \approx 0.707$ (moderate distance)
- $\rho_{ij} = -1$ (perfectly anti-correlated) $\implies$ $d_{ij} = 1$ (maximum distance, assets are "opposites")

Why not just use $d_{ij} = 1 - \rho_{ij}$? Because that doesn't satisfy the triangle inequality! The square root and factor of $\frac{1}{2}$ are carefully chosen to make it a proper metric.

6 Clustering

6.1 Clustering

Definition

Clustering is the task of grouping objects so that objects in the same group (cluster) are more similar to each other than to objects in other groups.
Input: Set of objects with pairwise distances/similarities
Output: Partition of objects into clusters

Intuition

Clustering is unsupervised learning: we find structure without being told what the groups are.
Applications:

- Customer segmentation: Group similar customers
- Image segmentation: Group pixels into objects
- Document clustering: Group similar articles
- Portfolio construction: Group similar assets

6.2 Hierarchical Clustering

Definition

Hierarchical clustering builds a tree (dendrogram) of clusters. Two approaches:

- **Agglomerative** (bottom-up): Start with each object as its own cluster; repeatedly merge closest clusters
- **Divisive** (top-down): Start with all objects in one cluster; repeatedly split clusters

HRP uses agglomerative clustering.

Intuition

Agglomerative algorithm:

1. Start: Each object is its own cluster
2. Repeat:
   (a) Find the two closest clusters
   (b) Merge them into a single cluster
3. Stop: When all objects are in one cluster

Output: A tree (dendrogram) showing the order and distance of merges. You can "cut" the tree at any level to get different numbers of clusters.

6.3 Linkage Methods

Definition

A **linkage method** defines the distance between two clusters A and B:

- **Single linkage**: $d(A, B) = \min_{i \in A, j \in B} d(i, j)$ (closest members)
- **Complete linkage**: $d(A, B) = \max_{i \in A, j \in B} d(i, j)$ (farthest members)
- **Average linkage**: $d(A, B) = \frac{1}{|A||B|} \sum_{i \in A} \sum_{j \in B} d(i, j)$ (average distance)
- **Ward linkage**: Minimize increase in within-cluster variance

Intuition

Different linkage methods produce different tree structures:
Single linkage:

- Pro: Simple, fast
- Con: Tends to create "chains" (long, stringy clusters)
- Used in HRP paper

Complete linkage:

- Pro: Creates compact, spherical clusters
- Con: Sensitive to outliers

Average linkage:

- Pro: Balanced, robust
- Con: More computationally expensive

Ward linkage:

- Pro: Minimizes variance, often gives good results
- Con: Only works with Euclidean distance

6.4 Dendrogram

Definition

A **dendrogram** is a tree diagram that records the hierarchical clustering structure.

- Leaves: Individual objects
- Internal nodes: Clusters formed by merging
- Height of node: Distance at which merge occurred

Intuition

Reading a dendrogram:

- Bottom: Individual objects
- Moving up: Objects merge into clusters
- Height where branches merge: Indicates how similar the merged clusters are
- Low merge: Very similar clusters
- High merge: Dissimilar clusters

You can "cut" the dendrogram at any height to get a flat clustering with a chosen number of clusters.
In HRP, the dendrogram structure determines how we allocate capital hierarchically.

7 Statistical Estimation

7.1 Estimator

Definition

An **estimator** is a function that maps sample data to an estimate of a population parameter.
Example: The sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is an estimator of the population mean μ.

7.2 Bias

Definition

The **bias** of an estimator $\hat{\theta}$ for parameter θ is:

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

An estimator is **unbiased** if $\mathbb{E}[\hat{\theta}] = \theta$ (on average, it equals the true value).

Intuition

Bias is systematic error: consistently over- or under-estimating.

- Unbiased: Errors average to zero over many samples
- Biased: Consistently wrong in one direction

Example: Sample mean $\bar{x}$ is unbiased for μ, but sample variance with n in denominator is biased (underestimates). Using $n - 1$ makes it unbiased.

7.3 Variance of Estimator

Definition

The **variance** of an estimator $\hat{\theta}$ measures its variability across different samples:

$$\text{Var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$$

Low variance: Estimates are consistent across samples
High variance: Estimates vary widely across samples

7.4 Mean Squared Error

Definition

The **mean squared error** (MSE) combines bias and variance:

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$$

Intuition

MSE measures total error:

- Even if unbiased, high variance means poor estimates
- Even if low variance, bias means systematically wrong
- MSE balances both concerns

This is the **bias-variance tradeoff**: Sometimes accepting a little bias reduces variance enough to lower MSE.
In portfolio optimization:

- Markowitz: Unbiased (uses all covariance information) but high variance (unstable)
- HRP: Slightly biased (uses only tree structure) but low variance (stable)
- HRP has lower MSE out-of-sample!

7.5 Maximum Likelihood Estimation

Definition

Maximum likelihood estimation (MLE) chooses the parameter value that makes the observed data most probable.
Given data $x_1, \ldots, x_n$ and probability model with parameter θ, the **likelihood** is:

$$L(\theta) = P(\text{data}|\theta)$$

The MLE is:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} L(\theta)$$

Intuition

MLE asks: "What parameter value makes my data least surprising?"
Example: Estimate the bias of a coin. Flip 10 times, get 7 heads. What's the probability p of heads?
Likelihood: $L(p) = \binom{10}{7} p^7 (1-p)^3$
Maximize: $\hat{p}_{\text{MLE}} = 7/10 = 0.7$
Intuition: The data suggests the coin is 70% likely to land heads.
In finance, we often estimate means and covariances via MLE (which gives the sample mean and sample covariance matrix).

8 Key Concepts from Information Theory

8.1 Entropy

Definition

The **entropy** of a discrete random variable X measures its uncertainty or information content:

$$H(X) = -\sum_i P(X = x_i) \log_2 P(X = x_i)$$

Measured in bits (if using $\log_2$).

Intuition

Entropy quantifies "how much information" a random variable contains:

- High entropy: Very uncertain, many possible outcomes
- Low entropy: Predictable, few possible outcomes
- Entropy = 0: No uncertainty (deterministic)

Example:

- Fair coin: $H = -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) = 1$ bit
- Biased coin (90% heads): $H \approx 0.47$ bits (more predictable)
- Two-headed coin: $H = 0$ bits (no uncertainty)

8.2 Mutual Information

Definition

The **mutual information** between random variables X and Y measures how much knowing one reduces uncertainty about the other:

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

where $H(X,Y)$ is the joint entropy.

Intuition

Mutual information generalizes correlation to capture any dependence (not just linear):

- $I(X;Y) = 0$: X and Y are independent (knowing one tells you nothing about the other)
- $I(X;Y) > 0$: X and Y are dependent (knowing one reduces uncertainty about the other)
- $I(X;Y) = H(X)$: Y completely determines X

Mutual information is always non-negative: $I(X;Y) \geq 0$
Unlike correlation, mutual information can detect any type of relationship, including nonlinear ones.

9 Portfolio Theory Specifics

9.1 Portfolio

Definition

A **portfolio** is a collection of financial assets with associated weights $w_1, \ldots, w_N$ where:

- w_i = fraction of total capital invested in asset i
- $\sum_{i=1}^{N} w_i = 1$ (all capital is allocated)
- $w_i \geq 0$ if short-selling is not allowed

The portfolio is represented by a weight vector $\mathbf{w} = (w_1, \ldots, w_N)^T$.

9.2 Portfolio Return

Definition

The **portfolio return** over a period is:

$$r_p = \sum_{i=1}^{N} w_i r_i = \mathbf{w}^T \mathbf{r}$$

where r_i is the return of asset i.
The **expected portfolio return** is:

$$\mu_p = \mathbb{E}[r_p] = \sum_{i=1}^{N} w_i \mu_i = \mathbf{w}^T \boldsymbol{\mu}$$

Intuition

Portfolio return is simply the weighted average of individual asset returns. If you put 40% in asset A (return 10%) and 60% in asset B (return 5%), your portfolio return is:

$$r_p = 0.4(10\%) + 0.6(5\%) = 7\%$$

This follows from linearity: if you double your investment, you double your return.

9.3 Portfolio Variance

Definition

The **portfolio variance** is:

$$\sigma_p^2 = \text{Var}(r_p) = \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} = \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \sigma_{ij}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix with entries $\sigma_{ij} = \text{Cov}(r_i, r_j)$.
The **portfolio volatility** (standard deviation) is:

$$\sigma_p = \sqrt{\sigma_p^2}$$

Intuition

Unlike return, portfolio variance is NOT a simple weighted average of individual variances. It includes covariance terms:

$$\sigma_p^2 = \underbrace{\sum_{i=1}^{N} w_i^2 \sigma_i^2}_{\text{individual variances}} + \underbrace{\sum_{i \neq j} w_i w_j \sigma_{ij}}_{\text{covariances}}$$

This is why diversification works: if assets aren't perfectly correlated, the covariance terms reduce total risk.

9.4 Sharpe Ratio

Definition

The **Sharpe ratio** measures risk-adjusted return:

$$\text{Sharpe} = \frac{\mu_p - r_f}{\sigma_p}$$

where:

- μ_p = expected portfolio return
- r_f = risk-free rate (e.g., Treasury bill rate)
- σ_p = portfolio volatility

Intuition

Sharpe ratio = "excess return per unit of risk"
Higher Sharpe ratio = better risk-adjusted performance
Example:

- Portfolio A: 12% return, 20% volatility, $r_f = 2\%$: Sharpe $= (12-2)/20 = 0.5$
- Portfolio B: 8% return, 10% volatility, $r_f = 2\%$: Sharpe $= (8-2)/10 = 0.6$

Portfolio B is better on a risk-adjusted basis (higher return per unit of risk), even though A has higher absolute return.
Typical values:

- Sharpe < 1: Not great
- Sharpe $= 1$ - 2: Good
- Sharpe > 2: Excellent (rare in practice)

9.5 Efficient Frontier

Definition

The **efficient frontier** is the set of portfolios that achieve:

- Maximum expected return for each level of risk, OR
- Minimum risk for each level of expected return

No portfolio below the frontier can dominate an efficient portfolio (provide more return for the same risk, or less risk for the same return).

Intuition

The efficient frontier is a curve in (risk, return) space. Portfolios on this curve are "Pareto optimal" - you can't improve one dimension without worsening the other. Any rational investor should choose a portfolio on the efficient frontier. The specific choice depends on risk tolerance:

- Risk-averse: Choose low-risk portfolio on the frontier
- Risk-tolerant: Choose high-return portfolio on the frontier

The tangent portfolio (highest Sharpe ratio) is often considered "optimal" for all investors who can borrow/lend at the risk-free rate.

9.6 Risk Parity

Definition

A **risk parity** portfolio allocates weights so that each asset contributes equally to total portfolio risk.
The risk contribution of asset i is:

$$\text{RC}_i = w_i \frac{\partial \sigma_p}{\partial w_i} = w_i \frac{(\boldsymbol{\Sigma}\mathbf{w})_i}{\sigma_p}$$

Risk parity requires: $\text{RC}_1 = \text{RC}_2 = \cdots = \text{RC}_N$

Intuition

Risk parity is different from equal weighting:

- Equal weighting: $w_i = 1/N$ (same amount in each asset)
- Risk parity: Weight inversely to risk (more in low-risk, less in high-risk)

Example: If asset A has volatility 10% and asset B has volatility 30%, risk parity puts 3× as much weight in A as in B, so they contribute equally to portfolio risk. HRP uses a hierarchical version of risk parity: at each split, allocate inversely to cluster risk.

10 Conclusion

This mathematical dictionary covers all foundational concepts needed to understand HRP and modern portfolio theory. Each concept builds on previous ones:

Foundation: Probability and statistics provide the language of uncertainty.
Structure: Linear algebra provides tools for multi-dimensional analysis.
Optimization: Optimization theory formalizes the decision problem.
Relationships: Graph theory and metric spaces provide alternative representations.
Discovery: Clustering algorithms find hidden structure in data.
Application: Portfolio theory applies all these tools to financial decision-making.

When you encounter an unfamiliar term in the HRP document, refer back to this dictionary for definitions, intuition, and examples. Mathematics is a language - and like any language, it becomes clearer with practice and reference.