Core Mathematical Proofs for Machine Learning

Abdul Samad Khan

May 2025

Introduction

This document presents ten fundamental mathematical proofs relevant to machine learning (ML) and its applications, particularly in finance. As a grad student diving into ML, I've compiled these proofs to solidify my understanding and share insights with researchers. Each proof is concise, with occasional reflections on ML and finance applications.

1 Derivation of OLS Estimator

We aim to show that the ordinary least squares (OLS) estimator is $\hat{\beta} = (X^T X)^{-1} X^T y$.

Proof. The OLS objective is to minimize the sum of squared residuals: $\min_\beta \|y - X\beta\|^2$. This is equivalent to minimizing the quadratic form $(y - X\beta)^T(y - X\beta)$. Expanding:

$$(y - X\beta)^T(y - X\beta) = y^T y - 2\beta^T X^T y + \beta^T X^T X\beta.$$

Take the derivative with respect to β and set to zero:

$$\frac{\partial}{\partial\beta}(y^T y - 2\beta^T X^T y + \beta^T X^T X\beta) = -2X^T y + 2X^T X\beta = 0.$$

Solving: $X^T X\beta = X^T y$, and assuming $X^T X$ is invertible, we get:

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

□

ML Application: OLS is foundational in linear regression, used in predictive modeling. In finance, it's applied to estimate asset pricing models like CAPM.

2 Covariance Matrix for Two Random Variables

Compute the covariance matrix for random variables X_1, X_2.

Proof. The covariance matrix Σ for a random vector $\mathbf{X} = [X_1, X_2]^T$ is defined as:

$$\Sigma = E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T].$$

Let $E[X_1] = \mu_1$, $E[X_2] = \mu_2$. Then:

$$\Sigma = \begin{bmatrix} E[(X_1 - \mu_1)^2] & E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)^2] \end{bmatrix} = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) \end{bmatrix}.$$

Since $\text{Cov}(X_1, X_2) = \text{Cov}(X_2, X_1)$, the matrix is symmetric. □

ML Application: Covariance matrices are crucial in PCA for dimensionality reduction, often used in financial risk modeling to capture asset correlations.

3 SVD Decomposition for a 2x2 Matrix

Prove the singular value decomposition (SVD) for a 2x2 matrix A.

Proof. For a 2x2 matrix A, SVD states $A = U\Sigma V^T$, where U, V are orthogonal, and Σ is diagonal with non-negative singular values. Compute $A^T A$:

$$A^T A = V\Lambda V^T,$$

where $\Lambda = \text{diag}(\lambda_1, \lambda_2)$ contains eigenvalues, and V has eigenvectors. Singular values are $\sigma_i = \sqrt{\lambda_i}$. Define $\Sigma = \text{diag}(\sigma_1, \sigma_2)$. Set $U = AV\Sigma^{-1}$ (assuming $\sigma_i \neq 0$). Verify:

$$U^T U = \Sigma^{-1} V^T A^T A V \Sigma^{-1} = \Sigma^{-1} V^T (V\Lambda V^T) V \Sigma^{-1} = \Sigma^{-1}\Lambda\Sigma^{-1} = I.$$

Thus, U is orthogonal, and $A = U\Sigma V^T$. □

ML Application: SVD is used in matrix factorization for recommender systems and latent factor models in finance.

4 Matrix Gradient of $f(W) = W^T AW$

Derive the gradient $\nabla_W f(W) = W^T AW$.

Proof. Consider $f(W) = \text{tr}(W^T AW)$. For a perturbation $W + \delta W$, compute:

$$f(W + \delta W) = \text{tr}((W + \delta W)^T A(W + \delta W)) = \text{tr}(W^T AW) + \text{tr}(\delta W^T AW) + \text{tr}(W^T A\delta W) + o(\|\delta W\|).$$

The linear term is:

$$\text{tr}(\delta W^T AW) + \text{tr}(W^T A\delta W) = \text{tr}(\delta W^T (AW)) + \text{tr}((AW)^T \delta W).$$

Thus, the gradient is:

$$\nabla_W f(W) = AW + A^T W.$$

If A is symmetric, this simplifies to $2AW$. □

ML Application: This gradient appears in optimizing neural network weights, especially in financial forecasting models.

5 Bayes' Theorem for Conditional Probabilities

Show Bayes' theorem: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.

Proof. By definition, $P(A|B) = \frac{P(A\cap B)}{P(B)}$ and $P(B|A) = \frac{P(A\cap B)}{P(A)}$. Rearrange the latter:

$$P(A \cap B) = P(B|A)P(A).$$

Substitute into the former:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

□

ML Application: Bayes' theorem underpins Bayesian inference, used in probabilistic ML models and risk assessment in finance.

6 Expected Value of a Binomial Distribution

Compute $E[X]$ for a binomial random variable $X \sim \text{Bin}(n, p)$.

Proof. A binomial variable X is the sum of n independent Bernoulli trials, each with success probability p. Let $X = \sum_{i=1}^{n} X_i$, where $X_i \sim \text{Bern}(p)$. Then:

$$E[X] = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} p = np.$$

□

ML Application: Expected values are used in loss functions for classification, such as in credit default modeling.

7 Orthogonality of Eigenvectors for Symmetric Matrices

Prove that eigenvectors of a symmetric matrix corresponding to distinct eigenvalues are orthogonal.

Proof. Let A be symmetric, with eigenvalues $\lambda_1 \neq \lambda_2$ and eigenvectors v_1, v_2. Then $Av_1 = \lambda_1 v_1$ and $Av_2 = \lambda_2 v_2$. Compute:

$$v_2^T A v_1 = v_2^T(\lambda_1 v_1) = \lambda_1 v_2^T v_1.$$

Since A is symmetric, $v_2^T A = (Av_2)^T = (\lambda_2 v_2)^T = \lambda_2 v_2^T$. Thus:

$$v_2^T A v_1 = (v_2^T A)v_1 = \lambda_2 v_2^T v_1.$$

Equate: $\lambda_1 v_2^T v_1 = \lambda_2 v_2^T v_1$. Since $\lambda_1 \neq \lambda_2$, we have $v_2^T v_1 = 0$, so v_1, v_2 are orthogonal. □

ML Application: Orthogonal eigenvectors are key in spectral clustering and portfolio optimization.

8 Variance of a Linear Combination of Random Variables

Derive $\text{Var}(a_1X_1 + a_2X_2)$.

Proof. For random variables X_1, X_2 with coefficients a_1, a_2, compute:

$$\text{Var}(a_1X_1 + a_2X_2) = E\left[(a_1X_1 + a_2X_2 - E[a_1X_1 + a_2X_2])^2\right].$$

Since $E[a_1X_1 + a_2X_2] = a_1E[X_1] + a_2E[X_2]$, expand:

$$\text{Var} = E\left[(a_1(X_1 - E[X_1]) + a_2(X_2 - E[X_2]))^2\right] = a_1^2\text{Var}(X_1) + a_2^2\text{Var}(X_2) + 2a_1a_2\text{Cov}(X_1, X_2).$$

□

ML Application: This is used in variance reduction techniques, like in Monte Carlo simulations for option pricing.

9 Properties of Trace for Matrix Products

Show tr$(ABC) =$ tr$(BCA) =$ tr(CAB) for compatible matrices.

Proof. For matrices A, B, C where ABC is defined, the trace is:

$$\text{tr}(ABC) = \sum_i (ABC)_{ii} = \sum_i \sum_j \sum_k A_{ij} B_{jk} C_{ki}.$$

Compute tr(BCA):

$$\text{tr}(BCA) = \sum_i (BCA)_{ii} = \sum_i \sum_j \sum_k B_{ij} C_{jk} A_{ki}.$$

Relabel indices: let $i \to j$, $j \to k$, $k \to i$. The sum becomes:

$$\sum_j \sum_k \sum_i A_{jk} B_{ki} C_{ij} = \text{tr}(ABC).$$

Similarly for tr(CAB). Thus, the trace is cyclic. □

ML Application: Trace properties simplify computations in neural network loss functions.

10 Conditional Expectation for a Bivariate Normal

Compute $E[X_1|X_2]$ for a bivariate normal (X_1, X_2).

Proof. Let $(X_1, X_2) \sim N(\mu, \Sigma)$, with $\mu = [\mu_1, \mu_2]^T$, $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$. The conditional distribution $X_1|X_2 = x_2$ is normal with mean:

$$E[X_1|X_2 = x_2] = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2).$$

This follows from the bivariate normal density, where the conditional mean is derived via the covariance structure. □

ML Application: Conditional expectations are used in Gaussian processes, applied in algorithmic trading.

Conclusion

These proofs, while foundational, are the backbone of ML algorithms. As a graduate student, I find their applications in finance, such as risk modeling and portfolio optimization, particularly motivating. I've shared this as a Notion wiki and linked it in my research resume under "Foundational Math Portfolio."