Notes: Cauchy–Schwarz, Quadratic Forms and Orthogonal Projection

1. The associated quadratic form

Let $(H, \langle \cdot, \cdot \rangle)$ be a Hilbert space. For $f, g \in H$ we define

$$Q(\lambda) \;=\; \|f - \lambda g\|^2, \qquad \lambda \in \mathbb{C}.$$

Expanding using the inner product:

$$Q(\lambda) \;=\; \|f\|^2 - 2\Re\big(\lambda \langle f, g\rangle\big) + |\lambda|^2 \|g\|^2.$$

That is, $Q(\lambda)$ is a *parabola* in the variable λ.

2. Optimization

Since $Q(\lambda) \geq 0$ for all λ, the minimum value is attained at

$$\lambda^* \;=\; \frac{\langle f, g\rangle}{\|g\|^2}.$$

Substituting:

$$Q(\lambda^*) \;=\; \|f\|^2 - \frac{|\langle f, g\rangle|^2}{\|g\|^2}.$$

3. Cauchy–Schwarz inequality

From non-negativity it follows that

$$Q(\lambda^*) \;\geq\; 0 \quad \Longrightarrow \quad |\langle f, g\rangle|^2 \;\leq\; \|f\|^2 \, \|g\|^2.$$

This is precisely the **Cauchy–Schwarz inequality**.

4. Geometric interpretation

- $Q(\lambda)$ represents the **squared distance** between f and the subspace spanned by g:

$$Q(\lambda) = \|f - \lambda g\|^2.$$

- The minimum $Q(\lambda^*)$ corresponds to the **orthogonal distance** from f to $\text{span}\{g\}$.
- The projected vector is

$$P_g(f) = \lambda^* g = \frac{\langle f, g\rangle}{\|g\|^2}\, g.$$

- The residual $f - P_g(f)$ is orthogonal to g:

$$\langle f - P_g(f),\, g\rangle = 0.$$

5. Continuity

The function $Q(\lambda)$ is quadratic and therefore continuous in λ. This guarantees that the passage to the minimum does not require ε–δ arguments, but rather that the geometry of the Hilbert space itself ensures continuity and convergence.

6. Applications in Machine Learning

The Cauchy–Schwarz inequality and the quadratic form interpretation appear in multiple areas of Machine Learning:

6.1 Cosine similarity

For embeddings $x, y \in \mathbb{R}^n$, define

$$\cos(\theta) = \frac{\langle x, y\rangle}{\|x\|\|y\|}.$$

Cauchy–Schwarz guarantees $-1 \leq \cos(\theta) \leq 1$, validating the notion of similarity used in NLP, vision, and recommender systems.

6.2 Regularization and stability

In optimization problems one minimizes

$$\min_w L(w) + \lambda\|w\|^2,$$

where the norm comes from an inner product. Cauchy–Schwarz guarantees continuity and stability of gradient-based algorithms.

6.3 Kernels and SVMs

If $K(x, y)$ is a kernel in a reproducing kernel Hilbert space (RKHS),

$$|K(x, y)| \leq \sqrt{K(x, x)}\,\sqrt{K(y, y)}.$$

This validates the kernel as a similarity measure and underpins methods such as SVMs and Gaussian Processes.

6.4 Error bounds

For random variables X, Y,

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]}\,\sqrt{\mathbb{E}[Y^2]}.$$

This bound is used in variance analysis, estimator consistency, and generalization theory (PAC learning).

6.5 Gradient and optimization

In gradient descent,

$$|\langle \nabla f(w), d\rangle| \leq \|\nabla f(w)\| \cdot \|d\|.$$

This ensures that the gradient acts as a continuous linear functional and that the step in direction d is controlled.

6.6 PCA and dimensionality reduction

In PCA one maximizes

$$\max_{\|v\|=1} v^T \Sigma v,$$

where Σ is the covariance matrix. The maximum is attained at a principal eigenvector thanks to Cauchy–Schwarz, which grounds dimensionality reduction.

Conclusion: The Cauchy–Schwarz inequality can be seen as the statement that the parabola

$$Q(\lambda) = \|f - \lambda g\|^2$$

never goes below the horizontal axis. Thus, Cauchy–Schwarz is not only an algebraic inequality, but also an expression of the *geometric continuity* of orthogonal projection.