Mathematics of Information — Cheatsheet

Daniele Cambria

May 22, 2026

Course Overview — Roter Faden

Unifying question: How can we represent, recover, and learn high-dimensional objects from few measurements?

- **Ch 1 — Frames.** Generalize ONBs via redundancy. Key objects: frame operator $\mathbb{S} = \mathbb{T}^*\mathbb{T}$, dual frame. Algebraic backbone for everything.
- **Ch 2 — Uncertainty.** A signal cannot be sparse in two incompatible representations. Quantified by coherence between bases; low coherence $\Rightarrow$ sparse vectors are distinguishable.
- **Ch 3 — Compressed Sensing.** Practical recovery: $\min \|\mathbf{x}\|_1$ s.t. $\mathbf{Dx} = \mathbf{y}$. Works when coherence is low (or spark is high).
- **Ch 7 — RIP.** Stronger than coherence. Gives sharper, noise-robust recovery ($\delta_{2s} < \sqrt{2} - 1$).
- **Ch 8–9 — JL & RIP via JL.** Concentration of measure: random Gaussian/sub-Gaussian matrices satisfy RIP with $m \asymp s \log(n/s)$ measurements.
- **Ch 10 — Metric Entropy.** Information-theoretic limits: how compactly can a class $\mathcal{C}$ be encoded? $\log N(\varepsilon; \mathcal{C})$ measures complexity.
- **Ch 12 — Uniform Laws.** When does empirical risk approximate true risk? Rademacher / VC dimension play the role metric entropy plays for signals.

Key bridges: Coherence $\leftrightarrow$ spark $\leftrightarrow$ RIP (three levels of recovery certification). Concentration of measure underlies JL, RIP, McDiarmid, Rademacher. Covering numbers link Ch 10 $\leftrightarrow$ Ch 12.

1 A Short Course on Frame Theory

Why frames

ONBs are too rigid: many applications (denoising, robust transmission, redundancy against quantization or erasures, time–frequency analysis) need expansions that are non-orthogonal or have more vectors than the dimension. **Frames are the right generalization**: they enable signal expansions with stable analysis/synthesis even when the spanning set is redundant. Key payoff: the frame condition $A\|x\|^2 \le \sum_k |\langle x, g_k\rangle|^2 \le B\|x\|^2$ guarantees that coefficients carry essentially all signal energy, and the dual frame gives an explicit, stable reconstruction formula. Everything later in the course (CS, RIP) treats dictionaries as frames whose redundancy is exploited rather than removed.

Key terminology: analysis vs. synthesis vs. dual synthesis

Given frame $\{g_k\}$ with dual $\{\tilde{g}_k\}$ (so $\tilde{g}_k = \mathbb{S}^{-1} g_k$ for canonical dual):

- **Analysis matrix/operator $\mathbf{T}$ (rows $= g_k^{\mathrm{H}}$):** produces coefficients $\mathbf{c} = \mathbf{Tx}$, i.e. $c_k = \langle \mathbf{x}, g_k\rangle$.
- **Adjoint $\mathbf{T}^{\mathrm{H}}$ (columns $= g_k$):** acts as $\mathbf{T}^{\mathrm{H}}\mathbf{c} = \sum_k c_k\, g_k$. *Not a reconstruction in general*: $\mathbf{T}^{\mathrm{H}}\mathbf{T} = \mathbb{S}$ (frame op), and $\mathbb{S}\mathbf{x} \neq \mathbf{x}$ unless $\mathbb{S} = \mathbf{I}$.
- **(Dual) synthesis matrix/operator $\tilde{\mathbf{T}}^{\mathrm{H}}$ (columns $= \tilde{g}_k$):** performs reconstruction:

$$\mathbf{x} = \tilde{\mathbf{T}}^{\mathrm{H}}\mathbf{T}\mathbf{x} = \sum_k \langle \mathbf{x}, g_k\rangle\, \tilde{g}_k$$

  This is what the lecture notes call "*the* synthesis matrix" in the basis/frame case — **it is always built from the dual frame elements**.
- **Special case — ONB:** $\tilde{g}_k = g_k$, so $\tilde{\mathbf{T}} = \mathbf{T}$, and the dual synthesis coincides with the adjoint: $\tilde{\mathbf{T}}^{\mathrm{H}} = \mathbf{T}^{\mathrm{H}}$. *This is the only case where "synthesis" = "adjoint".*
- **Symmetric form (frame case):** both $\mathbf{x} = \sum \langle \mathbf{x}, g_k\rangle \tilde{g}_k = \sum \langle \mathbf{x}, \tilde{g}_k\rangle g_k$ are valid — you can swap which set carries the coefficients and which carries the synthesis vectors.

1.1 Examples of Signal Expansions

ONB in $\mathbb{R}^2$: $\mathbf{e}_1 = [1,0]^\top$, $\mathbf{e}_2 = [0,1]^\top$. Every $\mathbf{x} \in \mathbb{R}^2$ expands as

$$\mathbf{x} = \langle \mathbf{x}, \mathbf{e}_1\rangle \mathbf{e}_1 + \langle \mathbf{x}, \mathbf{e}_2\rangle \mathbf{e}_2.$$

Analysis and Synthesis Matrices

Define expansion coefficients in vector form: $\mathbf{c} = \mathbf{Tx}$ where

$$\mathbf{T} = \begin{bmatrix} \mathbf{e}_1^\top \\ \mathbf{e}_2^\top \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

is the *analysis matrix* (produces coefficients). Reconstruct signal from coefficients via

$$\mathbf{x} = \mathbf{T}^\top \mathbf{c} = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 \end{bmatrix} \mathbf{c}$$

where $\mathbf{T}^\top$ is the *synthesis matrix*. For ONB: perfect reconstruction holds $\mathbf{x} = \mathbf{T}^\top \mathbf{T}\mathbf{x}$.

Biorthonormal Bases in $\mathbb{R}^2$: Take non-orthogonal vectors $\mathbf{e}_1 = [1,0]^\top, \mathbf{e}_2 = \frac{1}{\sqrt{2}}[1,1]^\top$. Analysis matrix $\mathbf{T} = \begin{bmatrix} 1 & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$ is invertible. To reconstruct

$$\mathbf{x} = \langle \mathbf{x}, \mathbf{e}_1\rangle \tilde{\mathbf{e}}_1 + \langle \mathbf{x}, \mathbf{e}_2\rangle \tilde{\mathbf{e}}_2$$

we set $[\tilde{\mathbf{e}}_1\, \tilde{\mathbf{e}}_2] = \mathbf{T}^{-1}$, giving $\tilde{\mathbf{e}}_1 = [1,-1]^\top, \tilde{\mathbf{e}}_2 = [0,\sqrt{2}]^\top$.

Biorthonormality Property

The two bases satisfy $\langle \mathbf{e}_j, \tilde{\mathbf{e}}_k\rangle = \delta_{jk}$ (Kronecker delta). Matrix form: $\mathbf{T}[\tilde{\mathbf{e}}_1\, \tilde{\mathbf{e}}_2] = \mathbf{I}_2$, enabling perfect reconstruction $\mathbf{x} = \tilde{\mathbf{T}}^\top \mathbf{T}\mathbf{x}$ where $\tilde{\mathbf{T}}^\top = [\tilde{\mathbf{e}}_1\, \tilde{\mathbf{e}}_2]$.

Overcomplete Expansion (Frame) in $\mathbb{R}^2$: Three vectors in 2D are linearly dependent. Consider $\mathbf{g}_1 = [1,0]^\top, \mathbf{g}_2 = [0,1]^\top, \mathbf{g}_3 = [1,-1]^\top$ with $\mathbf{g}_3 = \mathbf{g}_1 - \mathbf{g}_2$. Analysis matrix $\mathbf{T} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & -1 \end{bmatrix}$ is 3×2 (tall, rank-deficient inverse).

We can represent $\mathbf{x}$ using these three vectors in multiple ways:

$$\mathbf{x} = \langle \mathbf{x}, \mathbf{g}_1\rangle \tilde{\mathbf{g}}_1 + \langle \mathbf{x}, \mathbf{g}_2\rangle \tilde{\mathbf{g}}_2 + \langle \mathbf{x}, \mathbf{g}_3\rangle \tilde{\mathbf{g}}_3$$

Design Freedom in Frames

Two valid dual frames exist:

1. $\tilde{\mathbf{g}}_1 = \mathbf{g}_1, \tilde{\mathbf{g}}_2 = \mathbf{g}_2, \tilde{\mathbf{g}}_3 = \mathbf{0}$ (ignores third vector)
2. $\tilde{\mathbf{g}}_1 = 2\mathbf{g}_1, \tilde{\mathbf{g}}_2 = \mathbf{g}_2 - \mathbf{g}_1, \tilde{\mathbf{g}}_3 = -\mathbf{g}_1$ (exploits redundancy)

Since $\mathbf{T}$ has infinitely many left-inverses $\tilde{\mathbf{T}}^\top$ satisfying $\tilde{\mathbf{T}}^\top \mathbf{T} = \mathbf{I}_2$, we have infinitely many dual frames. This *redundancy provides design freedom* in choosing how to synthesize the signal.

1.2 Signal Expansions in Finite-Dimensional Hilbert Spaces

1.2.1 Orthonormal Bases

Definition 1.4: Orthonormal Basis (ONB)

A set of vectors $\{\mathbf{e}_k\}_{k=1}^M$, $\mathbf{e}_k \in \mathbb{C}^M$, is an ONB for $\mathbb{C}^M$ if:

1. $\text{span}\{\mathbf{e}_k\}_{k=1}^M = \mathbb{C}^M$ (spanning)
2. $\langle \mathbf{e}_k, \mathbf{e}_j \rangle = \delta_{kj}$ (orthonormality)

Why ONBs matter: Every signal in $\mathbb{C}^M$ decomposes uniquely. For any $\mathbf{x} \in \mathbb{C}^M$:

$$\mathbf{x} = \sum_{k=1}^{M} \langle \mathbf{x}, \mathbf{e}_k \rangle \mathbf{e}_k$$

Coefficients $c_k = \langle \mathbf{x}, \mathbf{e}_k \rangle$ via orthogonal projection — no linear system needed.

Example: Standard basis $\mathbf{e}_1 = [1, 0]^\top, \mathbf{e}_2 = [0, 1]^\top$ in $\mathbb{R}^2$ is an ONB. For $\mathbf{x} = [3, 4]^\top$:

$$c_1 = \langle \mathbf{x}, \mathbf{e}_1 \rangle = 3, \quad c_2 = \langle \mathbf{x}, \mathbf{e}_2 \rangle = 4.$$

So $\mathbf{x} = 3\mathbf{e}_1 + 4\mathbf{e}_2$.

Analysis and Synthesis with ONBs

Stack basis vectors into analysis matrix:

$$\mathbf{T} = \begin{bmatrix} \mathbf{e}_1^{\mathrm{H}} \\ \vdots \\ \mathbf{e}_M^{\mathrm{H}} \end{bmatrix}$$

Coefficients: $\mathbf{c} = \mathbf{T}\mathbf{x}$. Since $\mathbf{e}_k$ are orthonormal, $\mathbf{T}$ is unitary: $\mathbf{T}^{\mathrm{H}}\mathbf{T} = \mathbf{I}_M$.
Synthesis: reconstruct from $\mathbf{c}$ via

$$\mathbf{x} = \mathbf{T}^{\mathrm{H}}\mathbf{c}$$

where $\mathbf{T}^{\mathrm{H}} = [\mathbf{e}_1 \cdots \mathbf{e}_M]$ is the synthesis matrix. *This works because for an ONB the dual coincides ($\tilde{\mathbf{T}} = \mathbf{T}$); in the general basis/frame case the synthesis matrix is $\tilde{\mathbf{T}}^{\mathrm{H}}$, made of the **dual** elements.*

Why this works: Unitarity means $\mathbf{T}^{\mathrm{H}}\mathbf{T} = \mathbf{I}_M$, so

$$\|\mathbf{c}\|^2 = \mathbf{c}^{\mathrm{H}}\mathbf{c} = \mathbf{x}^{\mathrm{H}}\mathbf{T}^{\mathrm{H}}\mathbf{T}\mathbf{x} = \mathbf{x}^{\mathrm{H}}\mathbf{x} = \|\mathbf{x}\|^2.$$

Key Property: Perfect Reconstruction

From $\mathbf{c} = \mathbf{T}\mathbf{x}$, we recover $\mathbf{x}$ exactly:

$$\mathbf{T}^{\mathrm{H}}\mathbf{c} = \mathbf{T}^{\mathrm{H}}\mathbf{T}\mathbf{x} = \mathbf{x}.$$

No information loss in analysis or synthesis.

1.2.2 General (Biorthonormal) Bases

We relax orthonormality and allow non-orthogonal bases. Any basis $\{\mathbf{e}_k\}_{k=1}^M$ (spanning, linearly independent) has a **unique dual basis** $\{\tilde{\mathbf{e}}_k\}_{k=1}^M$ satisfying:

$$\langle \mathbf{e}_k, \tilde{\mathbf{e}}_j \rangle = \delta_{kj}$$

(biorthonormality). Perfect reconstruction holds:

$$\mathbf{x} = \sum_{k=1}^{M} \langle \mathbf{x}, \mathbf{e}_k \rangle \tilde{\mathbf{e}}_k$$

In matrix form: $\tilde{\mathbf{T}}^{\mathrm{H}}\mathbf{T} = \mathbf{I}_M$, where $\mathbf{T}$ is the analysis matrix of $\{\mathbf{e}_k\}$ and $\tilde{\mathbf{T}}^{\mathrm{H}}$ is the synthesis matrix of $\{\tilde{\mathbf{e}}_k\}$. Since $\mathbf{T}$ is square and full-rank, $\tilde{\mathbf{T}}^{\mathrm{H}} = \mathbf{T}^{-1}$ is unique.

Key distinction from ONBs: Biorthonormal bases are **not** norm-preserving. The norm of coefficients obeys:

$$\lambda_{\min}(\mathbf{T}^{\mathrm{H}}\mathbf{T})\|\mathbf{x}\|^2 \leq \|\mathbf{c}\|^2 \leq \lambda_{\max}(\mathbf{T}^{\mathrm{H}}\mathbf{T})\|\mathbf{x}\|^2$$

1.2.3 Redundant Signal Expansions (Frames in Finite Dimensions)

Consider $N > M$ vectors $\{\mathbf{g}_1, \ldots, \mathbf{g}_N\}$, $\mathbf{g}_k \in \mathbb{C}^M$, that **span** $\mathbb{C}^M$ but are linearly dependent. The analysis matrix $\mathbf{T} \in \mathbb{C}^{N \times M}$ (tall) has full column rank.

For any signal $\mathbf{x} \in \mathbb{C}^M$, compute coefficients:

$$\mathbf{c} = \mathbf{T}\mathbf{x} \in \mathbb{C}^N$$

A set $\{\tilde{\mathbf{g}}_1, \ldots, \tilde{\mathbf{g}}_N\}$ is a **dual frame** if:

$$\mathbf{x} = \sum_{k=1}^{N} \langle \mathbf{x}, \mathbf{g}_k \rangle \tilde{\mathbf{g}}_k \quad \text{for all } \mathbf{x} \in \mathbb{C}^M$$

This requires $\tilde{\mathbf{T}}^{\mathrm{H}}\mathbf{T} = \mathbf{I}_M$, where $\tilde{\mathbf{T}}^{\mathrm{H}} \in \mathbb{C}^{M \times N}$ is a synthesis matrix.

Theorem 1.6 (Moore-Penrose Inverse)

Let $\mathbf{T} \in \mathbb{C}^{N \times M}$ with $N \geq M$ and $\text{rank}(\mathbf{T}) = M$. The Moore-Penrose inverse is:

$$\mathbf{T}^\dagger = (\mathbf{T}^{\mathrm{H}}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{H}}$$

This is a left-inverse: $\mathbf{T}^\dagger\mathbf{T} = \mathbf{I}_M$.
All left-inverses (dual frame syntheses) are parametrized as:

$$\mathbf{L} = \mathbf{T}^\dagger + \mathbf{M}(\mathbf{I}_N - \mathbf{T}\mathbf{T}^\dagger)$$

for arbitrary $\mathbf{M} \in \mathbb{C}^{M \times N}$. This gives **infinitely many dual frames** when $N > M$ (design freedom).

1.3 Frames for General Hilbert Spaces

Motivation: From Finite to Infinite Dimensions

Finite-dimensional analysis matrices $\mathbf{T}$ generalize to **analysis operators** $\mathbb{T} : \mathcal{H} \to \ell^2$ in infinite-dimensional Hilbert spaces. A frame formalizes the concept of redundant signal expansions.

Bessel sequence: $\exists\, B < \infty$ such that $\sum_{k \in \mathcal{K}} |\langle x, g_k \rangle|^2 \leq B\|x\|^2$ for all $x \in \mathcal{H}$. The smallest such B is the *Bessel bound.*

Definition 1.8: Frame

A Bessel sequence $\{g_k\}_{k\in\mathcal{K}}, g_k \in \mathcal{H}$, is a **frame** for $\mathcal{H}$ if there exist constants $0 < A \leq B < \infty$ (frame bounds) such that:

$$A\|x\|^2 \leq \sum_{k\in\mathcal{K}} |\langle x, g_k\rangle|^2 \leq B\|x\|^2 \quad \text{for all } x \in \mathcal{H}$$

Equivalently: $A\|x\|^2 \leq \|\mathbb{T}x\|^2 \leq B\|x\|^2$.

Key properties of frames:

- **Completeness**: If $A > 0$ and $\langle x, g_k\rangle = 0$ for all k, then $x = 0$
- **Boundedness**: If $B < \infty$, then $\mathbb{T}$ is continuous
- **Lower bound → left-invertibility**: $A > 0$ ensures perfect reconstruction is possible

Example 1.9: Redundant ONB If $\{e_k\}$ is an ONB and $\{g_k\} = \{e_1, e_1, e_2, e_2, \ldots\}$ (each element repeated), then:

$$\sum_{k=1}^{\infty} |\langle x, g_k\rangle|^2 = 2\|x\|^2$$

So $\{g_k\}$ is a frame with bounds $A = B = 2$.

Example 1.10: Weighted frame Starting from ONB $\{e_k\}$, construct $\{g_k\} = \{e_1, \frac{1}{\sqrt{2}}e_2, \frac{1}{\sqrt{2}}e_2, \frac{1}{\sqrt{3}}e_3, \ldots\}$ where e_k appears k times with weight $\frac{1}{\sqrt{k}}$. This gives a tight frame with $A = B = 1$.

The adjoint operator: For a bounded linear operator $\mathbb{T} : \mathcal{H} \to \ell^2$, define the **adjoint** $\mathbb{T}^* : \ell^2 \to \mathcal{H}$ by:

$$\langle \mathbb{T}x, c\rangle = \langle x, \mathbb{T}^* c\rangle$$

For frame analysis, $\mathbb{T}^*$ acts as:

$$\mathbb{T}^*\{c_k\}_{k\in\mathcal{K}} = \sum_{k\in\mathcal{K}} c_k g_k$$

This generalizes the Hermitian transpose $\mathbf{T}^{\mathrm{H}}$ from finite to infinite dimensions.

Synthesis Operator $\mathbb{T}^*$: Adjoint vs. Dual Synthesis

The **synthesis operator** $\mathbb{T}^* : l^2 \to \mathcal{H}$ is the adjoint of the analysis operator: $\mathbb{T}^* c = \sum c_k g_k$.
Critical distinction: $\mathbb{T}^* \neq \tilde{\mathbb{T}}^*$ (dual synthesis):

- $\mathbb{T}^*$: Adjoint only. $\mathbb{T}^*\mathbb{T}x = \mathbb{S}x \neq x$ (NO perfect reconstruction).
- $\tilde{\mathbb{T}}^* = \mathbb{S}^{-1}\mathbb{T}^*$: Gives $\tilde{\mathbb{T}}^*\mathbb{T}x = x$ (perfect reconstruction).

Why? In redundant frames, $\mathbb{T}^*\mathbb{T} = \mathbb{S} \neq I$. Reconstruction requires:

$$x = \mathbb{T}^*\mathbb{S}^{-1}\mathbb{T}x \quad \text{or equivalently} \quad x = \tilde{\mathbb{T}}^*\mathbb{T}x$$

Example: Why $\mathbb{T}^*$ alone fails
Using $g_1 = g_2 = (1, 0)^T$, $g_3 = (0, 1)^T$: For $x = (a, b)^T$, coefficients are $c = \mathbb{T}x = (a, a, b)^T$.
Then $\mathbb{T}^*c = (a + a, b)^T = (2a, b)^T \neq x$. Apply $\mathbb{S}^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1 \end{pmatrix}$ to recover: $\mathbb{S}^{-1}(2a, b)^T = (a, b)^T$ ✓

1.3.1 The Frame Operator

Definition 1.14: Frame Operator

For a frame $\{g_k\}_{k\in\mathcal{K}}$ with analysis operator $\mathbb{T}$, the **frame operator** is:

$$\mathbb{S} = \mathbb{T}^*\mathbb{T} : \mathcal{H} \to \mathcal{H}$$

Acting on $x \in \mathcal{H}$:

$$\mathbb{S}x = \sum_{k\in\mathcal{K}} \langle x, g_k\rangle g_k$$

Key identity (eq. 1.39):

$$\sum_{k\in\mathcal{K}} |\langle x, g_k\rangle|^2 = \langle \mathbb{T}x, \mathbb{T}x\rangle = \langle \mathbb{T}^*\mathbb{T}x, x\rangle = \langle \mathbb{S}x, x\rangle$$

So the frame condition $A\|x\|^2 \leq \sum |\langle x, g_k\rangle|^2 \leq B\|x\|^2$ is equivalent to $A\|x\|^2 \leq \langle \mathbb{S}x, x\rangle \leq B\|x\|^2$.

Why is $\mathbb{S}$ central?

- In finite dimensions: $\mathbb{S} = \mathbf{T}^{\mathrm{H}}\mathbf{T}$ (Gram matrix of frame elements)

Theorem 1.15: Properties of the Frame Operator

1. **Linear and bounded**: $\mathbb{S}$ is continuous (composition of bounded operators)
2. **Self-adjoint**: $\mathbb{S}^* = (\mathbb{T}^*\mathbb{T})^* = \mathbb{T}^*\mathbb{T} = \mathbb{S}$
3. **Positive definite**: $\langle \mathbb{S}x, x\rangle = \|\mathbb{T}x\|^2 \geq A\|x\|^2 > 0$ for all $x \neq 0$
4. **Unique positive square root**: $\exists$ unique self-adjoint positive $\mathbb{S}^{1/2}$ with $\mathbb{S} = (\mathbb{S}^{1/2})^2$

Consequence: $\mathbb{S}$ is invertible, and the inverse $\mathbb{S}^{-1}$ is also self-adjoint, positive, bounded.

Proof outline:

- (1) Composition of bounded operators
- (2) $(AB)^* = B^*A^*$ gives self-adjointness
- (3) Lower bound: $\langle \mathbb{S}x, x\rangle \geq A\|x\|^2 > 0$
- (4) Spectral theorem: positive operators have unique positive square roots

Proof details:

- Property 1 follows because $\mathbb{S}$ is the composition $\mathbb{T}^*\mathbb{T}$ of bounded operators.
- Property 2 uses $(AB)^* = B^*A^*$, so $(\mathbb{T}^*\mathbb{T})^* = \mathbb{T}^*(\mathbb{T}^*)^* = \mathbb{T}^*\mathbb{T}$.
- Property 3: $\langle \mathbb{S}x, x\rangle = \sum |\langle x, g_k\rangle|^2 \geq A\|x\|^2 > 0$ for $x \neq 0$.
- Property 4 follows from general functional analysis: every self-adjoint positive bounded operator has unique self-adjoint positive square root (by spectral theorem).

Theorem 1.17: Frame Bounds from Spectrum
The **optimal** frame bounds are the extremal spectral values of $\mathbb{S}$:

$$A = \lambda_{\min}(\mathbb{S}) \quad \text{and} \quad B = \lambda_{\max}(\mathbb{S})$$

Numerical Example: Frame Operator in $\mathbb{R}^2$
Consider redundant frame in $\mathbb{R}^2$:

$$g_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad g_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad g_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

(First element repeated twice.) The analysis operator is $\mathbb{T}x = \{\langle x, g_k\rangle\}_{k=1}^3$, represented as:

$$\mathbb{T} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

The Gram matrix is:

$$\mathbb{S} = \mathbb{T}^*\mathbb{T} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

Eigenvalues: $\lambda_{\min} = 1$, $\lambda_{\max} = 2$, so frame bounds are $A = 1, B = 2$.

Direct verification: For $x = \begin{pmatrix} a \\ b \end{pmatrix}$,

$$\sum_{k=1}^{3} |\langle x, g_k \rangle|^2 = a^2 + a^2 + b^2 = 2a^2 + b^2$$

With $\|x\|^2 = a^2 + b^2$, we confirm:

$$(a^2 + b^2) \leq 2a^2 + b^2 \leq 2(a^2 + b^2) \quad \checkmark$$

1.3.2 The Canonical Dual Frame

Theorem 1.19: Canonical Dual Frame

Let $\{g_k\}_{k\in\mathcal{K}}$ be a frame with frame operator $\mathbb{S} = \mathbb{T}^*\mathbb{T}$ and bounds A, B. The **canonical dual frame** is:

$$\tilde{g}_k = \mathbb{S}^{-1} g_k, \quad k \in \mathcal{K}$$

This is also a frame for $\mathcal{H}$ with bounds $\tilde{A} = 1/B$ and $\tilde{B} = 1/A$.

Why is the canonical dual frame important?

- **Perfect reconstruction**: $x = \sum_{k\in\mathcal{K}} \langle x, g_k \rangle \tilde{g}_k$ for all $x \in \mathcal{H}$
- **Uniqueness**: The dual frame is the unique frame for perfect reconstruction
- **Reciprocal duality**: If $\{\tilde{g}_k\}$ is the dual of $\{g_k\}$, then $\{g_k\}$ is the dual of $\{\tilde{g}_k\}$

Key relationship: $\tilde{\mathbb{T}} = \mathbb{T}\mathbb{S}^{-1} = \mathbb{T}(\mathbb{T}^*\mathbb{T})^{-1}, \quad \tilde{\mathbb{S}} = \mathbb{S}^{-1}$.

Proof of Theorem 1.19 (key steps):

1. $\mathbb{S}^{-1}$ **is self-adjoint.** $(\mathbb{S}\mathbb{S}^{-1})^* = (\mathbb{S}^{-1})^*\mathbb{S}^* = \mathbb{I}$, and $\mathbb{S}^* = \mathbb{S}$, so $(\mathbb{S}^{-1})^*\mathbb{S} = \mathbb{I}$, hence $(\mathbb{S}^{-1})^* = \mathbb{S}^{-1}$.

2. **Analysis operator.** Using $(\mathbb{S}^{-1})^* = \mathbb{S}^{-1}$:

$$\begin{aligned} \tilde{\mathbb{T}}x &= \{\langle x, \tilde{g}_k \rangle\} = \{\langle x, \mathbb{S}^{-1} g_k \rangle\} \\ &= \{\langle \mathbb{S}^{-1} x, g_k \rangle\} = \mathbb{T}(\mathbb{S}^{-1} x) \end{aligned}$$

3. **Frame bounds.** Since $(\mathbb{S}^{-1})^* = \mathbb{S}^{-1}$:

$$\begin{aligned} \sum_k |\langle x, \tilde{g}_k \rangle|^2 &= \sum_k |\langle \mathbb{S}^{-1} x, g_k \rangle|^2 \\ &= \langle \mathbb{S}(\mathbb{S}^{-1} x), \mathbb{S}^{-1} x \rangle = \langle \mathbb{S}^{-1} x, x \rangle \end{aligned}$$

Eigenvalues of $\mathbb{S}^{-1}$ lie in $[1/B, 1/A]$, giving: $\frac{1}{B}\|x\|^2 \leq \langle \mathbb{S}^{-1} x, x \rangle \leq \frac{1}{A}\|x\|^2$.

Why $(\mathbb{S}^{-1})^* = \mathbb{S}^{-1}$*:* Self-adjointness of an inverse is not automatic — it uses $\mathbb{S}^* = \mathbb{S}$. The argument: $(\mathbb{S}^{-1})^*\mathbb{S} = (\mathbb{S}^{-1})^*\mathbb{S}^* = (\mathbb{S}\mathbb{S}^{-1})^* = \mathbb{I}$, so multiply right by $\mathbb{S}^{-1}$.

Numerical Example: Dual Frame in $\mathbb{R}^2$

Continuing the previous example with $g_1 = g_2 = (1, 0)^T$ and $g_3 = (0, 1)^T$ where $\mathbb{S} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$:

$$\mathbb{S}^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1 \end{pmatrix}$$

The dual frame elements are:

$$\tilde{g}_1 = \mathbb{S}^{-1} g_1 = \begin{pmatrix} 1/2 \\ 0 \end{pmatrix}, \quad \tilde{g}_2 = \begin{pmatrix} 1/2 \\ 0 \end{pmatrix}, \quad \tilde{g}_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Verify reconstruction: For $x = (a, b)^T$:

$$\sum_{k=1}^{3} \langle x, g_k \rangle \tilde{g}_k = \tfrac{a}{2}(1, 0)^T + \tfrac{a}{2}(1, 0)^T + b(0, 1)^T = (a, b)^T = x \checkmark$$

The dual frame has bounds $\tilde{A} = 1/B = 1/2$ and $\tilde{B} = 1/A = 1$.

Theorem 1.21: Dual Frame Operator

The canonical dual frame operator satisfies:

$$\tilde{\mathbb{S}} = \mathbb{S}^{-1}$$

Proof of Theorem 1.21:

$$\begin{aligned} \tilde{\mathbb{S}}x &= \sum_k \langle x, \tilde{g}_k \rangle \tilde{g}_k = \sum_k \langle x, \mathbb{S}^{-1} g_k \rangle \, \mathbb{S}^{-1} g_k \\ &= \mathbb{S}^{-1} \sum_k \langle \mathbb{S}^{-1} x, g_k \rangle g_k = \mathbb{S}^{-1}\mathbb{S}\mathbb{S}^{-1} x = \mathbb{S}^{-1} x \end{aligned}$$

Step 3 uses $(\mathbb{S}^{-1})^* = \mathbb{S}^{-1}$ to move $\mathbb{S}^{-1}$ outside; step 4 uses the definition of $\mathbb{S}$.

1.3.3 Signal Expansions (Reconstruction)

Theorem 1.22: Frame Reconstruction

Let $\{g_k\}$ and $\{\tilde{g}_k\}$ be canonical dual frames. Every signal $x \in \mathcal{H}$ has two perfect reconstructions:

$$x = \sum_{k\in\mathcal{K}} \langle x, \tilde{g}_k \rangle g_k = \sum_{k\in\mathcal{K}} \langle x, g_k \rangle \tilde{g}_k$$

Equivalently: $\mathbb{T}^*\tilde{\mathbb{T}} = \mathbb{I}_{\mathcal{H}}$ and $\tilde{\mathbb{T}}^*\mathbb{T} = \mathbb{I}_{\mathcal{H}}$.

Proof of Theorem 1.22: Show $\mathbb{T}^*\tilde{\mathbb{T}} = \mathbb{I}$:

$$\begin{aligned} \mathbb{T}^*\tilde{\mathbb{T}}x &= \sum_k \langle x, \tilde{g}_k \rangle g_k = \sum_k \langle x, \mathbb{S}^{-1} g_k \rangle g_k \\ &\overset{(\mathbb{S}^{-1})^* = \mathbb{S}^{-1}}{=} \sum_k \langle \mathbb{S}^{-1} x, g_k \rangle g_k = \mathbb{S}\mathbb{S}^{-1} x = x \end{aligned}$$

The proof of $\tilde{\mathbb{T}}^*\mathbb{T} = \mathbb{I}$ is analogous (swap $g_k \leftrightarrow \tilde{g}_k$).

Note: Both forms avoid computing $\mathbb{S}^{-1}$ explicitly; the dual absorbs the inversion.

1.3.4 Tight Frames

A frame $\{g_k\}_{k\in\mathcal{K}}$ with $A = B$ is called **tight**.

Theorem 1.26: Tight Frame Characterization

A frame is tight with bound A if and only if:

$$\mathbb{S} = A\mathbb{I}_{\mathcal{H}} \quad \Longleftrightarrow \quad x = \frac{1}{A} \sum_{k\in\mathcal{K}} \langle x, g_k \rangle g_k$$

This is Parseval's identity for tight frames. The dual is simple: $\tilde{g}_k = \frac{1}{A} g_k$.

Tight frames allow perfect reconstruction without inverting $\mathbb{S}$ — just scale by $1/A$.

Theorem 1.29: A tight frame with $A = 1$ and $\|g_k\| = 1$ for all k is an ONB. *(Tight + unit norm + $A = 1 \Rightarrow$ no redundancy.)*

Theorem 1.28: Normalize Any Frame to Tight

Given any frame $\{g_k\}$ with frame operator $\mathbb{S}$, the scaled frame $\{\mathbb{S}^{-1/2} g_k\}$ is tight with bound $A = 1$:

$$x = \sum_{k\in\mathcal{K}} \langle x, \mathbb{S}^{-1/2} g_k \rangle \mathbb{S}^{-1/2} g_k$$

This provides a systematic way to design tight frames from redundant ones.

1.3.5 Exact Frames

Definition 1.32: Exact Frame

A frame $\{g_k\}_{k\in\mathcal{K}}$ is exact if removing any single element leaves an incomplete set. Otherwise it is inexact (redundant).

Key equivalences:

- Exact frame = Riesz basis = frame with no redundancy
- Analysis operator $\mathbb{T}$ is surjective (onto ℓ^2)
- Representation unique: $\sum c_k g_k = \sum a_k g_k \Rightarrow c_k = a_k$ (exact frames: coefficients $\langle x, \tilde{g}_k\rangle$ have minimal ℓ^2-norm)

Biorthonormality: Exact frames and their canonical duals satisfy:

$$\langle g_k, \tilde{g}_m\rangle = \delta_{k,m}$$

This is the infinite-dimensional generalization of biorthonormal bases.

Why exactness matters:

- Frames are generalizations of bases: exact = no redundancy, inexact = has redundancy
- Reconstruction coefficients in exact frames have unique minimum ℓ^2-norm (canonical dual gives this)
- Tight exact frames = orthonormal bases (special case: $A = 1$, no redundancy)

1.4.1 Sampling as Frame Expansion

Theorem 1.38: Shannon Sampling as Tight Frame

Let $x(t) \in \mathcal{L}^2(B)$ (bandlimited to B Hz), sampled at rate $1/T > 2B$. Define:

$$g_k(t) = 2B\,\mathrm{sinc}(2B(t - kT)), \quad k \in \mathbb{Z}$$

Then $\{g_k\}_{k\in\mathbb{Z}}$ is a **tight frame** for $\mathcal{L}^2(B)$ with bound $A = 1/T$, and:

$$x(t) = T\sum_{k=-\infty}^{\infty} \langle x, g_k\rangle\, g_k(t) = T\sum_{k=-\infty}^{\infty} x(kT)\, g_k(t)$$

Samples $x(kT) = \langle x, g_k\rangle$ are frame coefficients. $\mathbb{T} : x \mapsto \{x(kT)\}$ is the analysis operator (sampling); $\mathbb{T}^*$ is sinc interpolation. Frame is tight with $A = 1/T$ regardless of oversampling rate (as long as $1/T > 2B$).

2 Sparse Recovery & Compressed Sensing

Why compressed sensing

Classical sampling (Nyquist) measures every degree of freedom. But many real signals (images in a wavelet basis, sparse spectra, gene expression) have far fewer *active* components than ambient dimensions. **CS asks: if $\mathbf{x}$ is s-sparse, can we recover it from $m \ll n$ linear measurements $\mathbf{y} = \mathbf{D}\mathbf{x}$?** The ℓ_0-minimization that captures sparsity is NP-hard; the deep insight is that its *convex relaxation* (Basis Pursuit, ℓ_1) recovers the same solution provided the dictionary has a certifying property: low coherence, large spark, or RIP (Ch 7). Uncertainty relations (Ch 2) give the first such guarantee: two incompatible bases force any nonzero signal to be jointly non-sparse, so sparse signals are uniquely identifiable from partial measurements.

Setup: Dictionary $\mathbf{D} \in \mathbb{C}^{m\times n}$ with $m < n$ and unit-norm columns $\|d_\ell\|_2 = 1$. Observe $\mathbf{y} = \mathbf{D}\mathbf{x}$, recover $\mathbf{x}$ assuming *sparsity*.

Sparsity: $\|\mathbf{x}\|_0 \triangleq |\{i : x_i \neq 0\}|$ (quasi-norm). $\mathbf{x}$ is **s-sparse** if $\|\mathbf{x}\|_0 \le s$. Support: $\mathrm{supp}(\mathbf{x}) = \{i : x_i \neq 0\}$.

2.1 Coherence

Coherence & Cumulative Coherence

$$\mu(\mathbf{D}) \triangleq \max_{r\neq\ell} |\langle d_\ell, d_r\rangle|$$

$$\mu_n(\mathbf{D}) \triangleq \max_{|S|\le n}\ \max_{\ell\notin S} \sum_{j\in S} |\langle d_\ell, d_j\rangle|$$

Welch bound: $\mu(\mathbf{D}) \ge \sqrt{\frac{n-m}{m(n-1)}}$; for $n \gg m$, $\mu \gtrsim 1/\sqrt{m}$.

Useful: $\mu_1(\mathbf{D}) = \mu(\mathbf{D})$ and $\mu_n(\mathbf{D}) \le n\,\mu(\mathbf{D})$. Low coherence $\Leftrightarrow$ columns near-orthogonal.

Between two ONBs $\{u_i\}, \{v_j\}$ of $\mathcal{H}^d$: $\mu(\mathcal{B}_1, \mathcal{B}_2) = \max_{i,j} |\langle u_i, v_j\rangle| \ge 1/\sqrt{d}$ (else $\|u_1\|^2 = \sum_j |\langle u_1, v_j\rangle|^2 < 1$, contradiction). Saturated by mutually unbiased bases.

Donoho–Stark uncertainty: For unitary $\mathbf{U}$ and index sets $\mathcal{P}, \mathcal{Q} \subseteq \{1, \dots, m\}$, let $\mathbf{D}_\mathcal{P}$ be the diagonal projection onto coords $\mathcal{P}$ and $\mathbf{P}_\mathcal{Q}(\mathbf{U}) = \mathbf{U}\mathbf{D}_\mathcal{Q}\mathbf{U}^{\mathrm{H}}$ the projection onto $\mathrm{span}\{u_j : j \in \mathcal{Q}\}$. Define

$$\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) \triangleq \|\mathbf{D}_\mathcal{P}\mathbf{P}_\mathcal{Q}(\mathbf{U})\|_2 = \max_{\mathbf{x}\in\mathrm{span}\{u_j\}_{j\in\mathcal{Q}},\, \mathbf{x}\neq 0} \frac{\|\mathbf{x}_\mathcal{P}\|_2}{\|\mathbf{x}\|_2}$$

Reading: How much of a $\mathcal{Q}$-band-limited vector can concentrate on $\mathcal{P}$. An uncertainty relation is a bound $\Delta_{\mathcal{P},\mathcal{Q}} \le c < 1$. Donoho–Stark's prototype: $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{F}) \le \sqrt{|\mathcal{P}||\mathcal{Q}|/m}$ for DFT $\mathbf{F}$ — bounds how time- and frequency-concentrated a signal can simultaneously be.

2.2 Spark and Uniqueness

Spark

$\mathrm{spark}(\mathbf{D})$ = cardinality of the smallest linearly dependent subset of columns.

Uniqueness ($\mathbf{P}_0$): If $\|\mathbf{x}\|_0 < \mathrm{spark}(\mathbf{D})/2$, then $\mathbf{x}$ is the unique $\|\cdot\|_0$-minimal solution of $\mathbf{D}\hat{\mathbf{x}} = \mathbf{y}$.

Coherence-spark bound: $\mathrm{spark}(\mathbf{D}) \ge 1 + \frac{1}{\mu(\mathbf{D})}$.

Hence $\|\mathbf{x}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right) \Rightarrow$ unique recovery.

Proof sketch (spark bound): Take $\mathbf{x}$ minimal nontrivial null-space vector, $\|\mathbf{x}\|_0 = \mathrm{spark}(\mathbf{D})$. From $\mathbf{D}\mathbf{x} = 0$, left-multiply by d_ℓ^{H}: $|x_\ell| \le \mu(\mathbf{D}) \sum_{r\neq\ell} |x_r|$. Add $\mu|x_\ell|$ to both sides, sum over $\mathrm{supp}(\mathbf{x})$: $(1+\mu)\|\mathbf{x}\|_1 \le \mu\,\|\mathbf{x}\|_1\, \mathrm{spark}(\mathbf{D})$.

2.3 Recovery Problems

(P_0) $\min \|\hat{\mathbf{x}}\|_0$ s.t. $\mathbf{D}\hat{\mathbf{x}} = \mathbf{y}$ (NP-hard)
(P_1) $\min \|\hat{\mathbf{x}}\|_1$ s.t. $\mathbf{D}\hat{\mathbf{x}} = \mathbf{y}$ (LP: Basis Pursuit)

Why ℓ_1? Convex relaxation of ℓ_0. Geometrically: ℓ_1-ball has vertices on sparse vectors, so the feasible affine subspace $\{\mathbf{x}\} + \mathcal{N}(\mathbf{D})$ first touches the ball at a sparse point.

2.4 Exact Recovery Condition (ERC)

ERC for support S

$$\frac{\max_{\ell\notin S} \sum_{j\in S} |\langle d_\ell, d_j\rangle|}{1 - \max_{\ell\in S} \sum_{j\in S\setminus\{\ell\}} |\langle d_\ell, d_j\rangle|} < 1$$

Sufficient (cumul. coherence): $\mu_{|S|-1}(\mathbf{D})+\mu_{|S|}(\mathbf{D}) < 1 \Rightarrow$ ERC holds, (P_1) recovers every $\mathbf{x}$ on S.

Key identity: $\langle \mathbf{x}, \operatorname{sgn}(\mathbf{x})\rangle = \|\mathbf{x}\|_1$ (since $x_i\,\overline{\operatorname{sgn}(x_i)} = |x_i|$).

Subspace duality: $\operatorname{im}(\mathbf{A}^{\mathrm{H}}) = \ker(\mathbf{A})^{\perp}$. So $\mathbf{v} \in \ker \mathbf{A}$, $\mathbf{u} \in \operatorname{im} \mathbf{A}^{\mathrm{H}} \Rightarrow \langle \mathbf{v}, \mathbf{u}\rangle = 0$ (lets you replace $\operatorname{sgn}(\mathbf{x})$ with $\operatorname{sgn}(\mathbf{x}) - \mathbf{u}$ inside null-space pairings).

2.5 Restricted Null Space Property (NSP)

$P_1(\mathcal{S}, \mathbf{D})$ — Restricted Null Space Property

$$P_1(\mathcal{S}, \mathbf{D}) \triangleq \max_{\mathbf{x}\in\mathcal{N}(\mathbf{D}),\, \mathbf{x}\neq 0} \frac{\sum_{k\in\mathcal{S}} |x_k|}{\sum_k |x_k|}$$

Quantifies how much mass any null-space vector can put on $\mathcal{S}$.

Theorem 3.4 (NSP $\Rightarrow$ BP recovery)

If $\mathbf{x}$ is supported on $\mathcal{S}$ and $P_1(\mathcal{S}, \mathbf{D}) < 1/2$, then $\mathbf{x}$ is the **unique** (P_1) minimizer of $\min \|\hat{\mathbf{x}}\|_1$ s.t. $\mathbf{D}\hat{\mathbf{x}} = \mathbf{D}\mathbf{x}$.

Reading: NSP is *exactly* the right condition for BP — coherence and RIP are sufficient certificates that imply NSP. Specifically: $|\mathcal{S}| < \frac{1}{2}(1 + 1/\mu) \Rightarrow P_1(\mathcal{S}, \mathbf{D}) < 1/2$, so coherence-based recovery factors through NSP.

2.6 Dual Certificate for BP

Theorem: If $\mathbf{x}$ has support S and there exists $\mathbf{z} \in \mathbb{C}^m$ with

$$\mathbf{D}_S^{\mathrm{H}}\mathbf{z} = \operatorname{sign}(\mathbf{x}_S), \quad \|\mathbf{D}_{S^c}^{\mathrm{H}}\mathbf{z}\|_\infty < 1,$$

and $\mathbf{D}_S$ has full column rank, then $\mathbf{x}$ is the *unique* (P_1) minimizer.

Reading: $\mathbf{z}$ is a witness; $\mathbf{D}^{\mathrm{H}}\mathbf{z}$ lies in $\partial\|\cdot\|_1$ at $\mathbf{x}$ ($= 1$ on S in modulus, < 1 off S). This is the KKT subgradient condition.

2.7 Compressibility

Best s-term approximation: $\sigma_s(\mathbf{x})_p = \inf_{\|\mathbf{z}\|_0 \le s} \|\mathbf{x} - \mathbf{z}\|_p$ = keep s largest-magnitude entries, zero the rest.

Weak-ℓ_p quasinorm: $\|\mathbf{x}\|_{p,\infty} = \sup_{t>0} t \cdot |\{i : |x_i| > t\}|^{1/p}$. Equivalently, if $|x_{(1)}| \ge |x_{(2)}| \ge \dots$ are sorted magnitudes:

$$\|\mathbf{x}\|_{p,\infty} = \sup_k k^{1/p}|x_{(k)}|$$

Decay $\Rightarrow$ approximability: $\|\mathbf{x}\|_{p,\infty} \le C \Rightarrow \sigma_s(\mathbf{x})_q \le C\, s^{1/q-1/p}$ for $q > p$.

3 Restricted Isometry Property (RIP)

Why RIP

Coherence is easy to compute but *pessimistic*: it certifies recovery only up to $s \sim 1/\mu \sim \sqrt{m}$, far from the optimal $s \sim m/\log n$. RIP is the right strengthening: $\mathbf{\Phi}$ acts as a near-isometry on *all* s-sparse vectors. From RIP follow (i) uniqueness of sparse representations, (ii) noise-robust ℓ_1 recovery (Thm 7.2/7.3), and (iii) stability to compressibility errors. **The trade-off:** RIP gives sharper guarantees but is NP-hard to verify deterministically — which is exactly why random matrices (Ch 8–9) become essential.

Setup: Measurement matrix $\mathbf{\Phi} \in \mathbb{C}^{m\times n}$, $m < n$. Observe $\mathbf{y} = \mathbf{\Phi}\mathbf{x}$ (or $\mathbf{y} = \mathbf{\Phi}\mathbf{x} + \mathbf{n}$, $\|\mathbf{n}\|_2 \le \varepsilon$).

3.1 RIP and Restricted Orthogonality

s-Restricted Isometry Constant δ_s

Smallest $\delta \ge 0$ such that, for all s-sparse $\mathbf{x}$:

$$(1-\delta_s)\|\mathbf{x}\|_2^2 \le \|\mathbf{\Phi}\mathbf{x}\|_2^2 \le (1+\delta_s)\|\mathbf{x}\|_2^2$$

Equivalently, $\big|\|\mathbf{\Phi}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\big| \le \delta_s\|\mathbf{x}\|_2^2$.

(s,t)-Restricted Orthogonality Constant $\theta_{s,t}$

Smallest $\theta \ge 0$ such that, for all *disjointly supported* s-sparse $\mathbf{u}$ and t-sparse $\mathbf{v}$:

$$|\langle \mathbf{\Phi}\mathbf{u}, \mathbf{\Phi}\mathbf{v}\rangle| \le \theta_{s,t}\,\|\mathbf{u}\|_2\,\|\mathbf{v}\|_2$$

Reading: δ_s controls energy preservation on sparse vectors; $\theta_{s,t}$ controls cross-correlation between disjoint sparse blocks.

3.2 Key Inequalities Between Constants

Monotonicity: $\delta_s \le \delta_{s+1}$, $\quad \theta_{s,t} \le \theta_{s+1,t}$.

θ–δ bound (Lemma 7.4):

$$\theta_{s,t} \le \delta_{s+t}$$

Reverse bound: $\delta_{s+t} \le \dfrac{s\,\delta_s + t\,\delta_t + 2\sqrt{st}\,\theta_{s,t}}{s+t}$

Telescoping (24-P4): $\delta_{ns} \le (n-1)\,\theta_{s,s} + \delta_s$ (split an ns-sparse vector into n disjoint s-sparse blocks and expand $\|\mathbf{\Phi}\mathbf{v}\|^2 - \|\mathbf{v}\|^2$).

Decomposition identity: For $\mathbf{x} = \mathbf{u} + \mathbf{v}$ with disjoint $\operatorname{supp}(\mathbf{u})$, $\operatorname{supp}(\mathbf{v})$ ($|\operatorname{supp}\mathbf{u}| \le s$, $|\operatorname{supp}\mathbf{v}| \le t$):

$$\big|\|\mathbf{\Phi}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\big| \le \delta_s\|\mathbf{u}\|_2^2 + \delta_t\|\mathbf{v}\|_2^2 + 2\theta_{s,t}\|\mathbf{u}\|_2\|\mathbf{v}\|_2$$

3.3 Recovery Guarantees

Theorem 7.2 (Noiseless, ℓ_1 recovery)

If $\delta_{2s} < \sqrt{2} - 1 \approx 0.414$, then the BP solution $\mathbf{x}^* = \arg\min \|\tilde{\mathbf{x}}\|_1$ s.t. $\mathbf{\Phi}\tilde{\mathbf{x}} = \mathbf{y}$ satisfies

$$\|\mathbf{x}^* - \mathbf{x}\|_2 \le C_0\, s^{-1/2}\,\|\mathbf{x} - \mathbf{x}_s\|_1$$

where $\mathbf{x}_s$ = best s-term approximation of $\mathbf{x}$. **In particular, $\mathbf{x}$ exactly recovered if s-sparse.**

Theorem 7.3 (Noisy, robust ℓ_1 recovery)

If $\delta_{2s} < \sqrt{2} - 1$ and $\|\mathbf{n}\|_2 \le \varepsilon$, then $\mathbf{x}^* = \arg\min \|\tilde{\mathbf{x}}\|_1$ s.t. $\|\mathbf{y} - \mathbf{\Phi}\tilde{\mathbf{x}}\|_2 \le \varepsilon$ satisfies

$$\|\mathbf{x}^* - \mathbf{x}\|_2 \le C_0\, s^{-1/2}\,\|\mathbf{x} - \mathbf{x}_s\|_1 + C_1\varepsilon$$

Reading: Two error sources — *compressibility* (how far $\mathbf{x}$ is from s-sparse, $\|\mathbf{x} - \mathbf{x}_s\|_1$) and *noise* ($\varepsilon$).

3.4 RIP for General (Non-Sparse) x

Split $\mathbf{x}$ by sorted magnitudes into blocks $T_0, T_1, \dots$ of size s (largest entries first). Then:

$$\|\mathbf{\Phi}\mathbf{x}\|_2 \le \sqrt{1+\delta_s}\,\|\mathbf{x}_{T_0}\|_2 + \sum_{j\ge 1}\sqrt{1+\delta_s}\,\|\mathbf{x}_{T_j}\|_2$$

Using $\|\mathbf{x}_{T_j}\|_2 \le s^{-1/2}\|\mathbf{x}_{T_{j-1}}\|_1$ (block magnitudes decreasing): $\|\mathbf{\Phi}\mathbf{x}\|_2 \le \sqrt{1+\delta_s}\big(\|\mathbf{x}\|_2 + s^{-1/2}\|\mathbf{x}\|_1\big)$.

3.5 RIP & Coherence

For $s = 1$: $\delta_1 = \max_i \big|\|\mathbf{\Phi}_i\|_2^2 - 1\big|$; for unit-norm columns, $\delta_1 = 0$.

For $s = 2$ with unit columns: $\delta_2 = \mu(\mathbf{\Phi})$.
More generally, Gershgorin: $\delta_s \le (s-1)\,\mu(\mathbf{\Phi})$.

3.6 RIP ⇒ NSP (Theorem 7.7)

Theorem 7.7

If $\delta_{2s}(\mathbf{\Phi}) < 1/3$, then $P_1(\mathcal{S}, \mathbf{\Phi}) < 1/2$ for every $\mathcal{S}$ with $|\mathcal{S}| \le s$, i.e. the restricted null space property holds.

Reading: The bridge between Ch 7 (RIP) and Ch 3 (BP recovery). The full chain:

$$\delta_{2s} < \tfrac{1}{3} \xRightarrow{\text{RIP}} P_1 < \tfrac{1}{2} \xRightarrow{\text{NSP}} \text{BP recovers } s\text{-sparse } \mathbf{x}.$$

The sharper threshold $\delta_{2s} < \sqrt{2} - 1$ (Thm 7.2) avoids the NSP detour and gives stable/robust recovery directly. NSP is the cleaner conceptual condition; RIP is the verifiable certificate for it.

Proof sketch ($\theta_{s,t} \le \delta_{s+t}$): For unit-norm disjoint $\mathbf{v}, \mathbf{v}'$ (s- and t-sparse), $\mathbf{v} \pm \mathbf{v}'$ is $(s+t)$-sparse with $\|\mathbf{v} \pm \mathbf{v}'\|_2^2 = 2$. RIP gives $2(1-\delta_{s+t}) \le \|\mathbf{\Phi}(\mathbf{v} \pm \mathbf{v}')\|_2^2 \le 2(1+\delta_{s+t})$. Subtract the two cases and use the parallelogram identity: $|\langle \mathbf{\Phi}\mathbf{v}, \mathbf{\Phi}\mathbf{v}' \rangle| = \frac{1}{4}\big|\|\mathbf{\Phi}(\mathbf{v}+\mathbf{v}')\|_2^2 - \|\mathbf{\Phi}(\mathbf{v}-\mathbf{v}')\|_2^2\big| \le \delta_{s+t}$.

4 Johnson–Lindenstrauss Lemma

Why JL closes the CS loop

RIP is hard to verify, but *random* matrices concentrate sharply: $\|\mathbf{\Phi}\mathbf{u}\|^2$ is within $\varepsilon\|\mathbf{u}\|^2$ of its mean with exponential probability. JL says this extends to a *whole set* of points via a union bound, requiring only $k = \mathcal{O}(\log m/\varepsilon^2)$ dimensions. **The deep observation (Ch 9):** apply this to a fine ε-net of the unit sparse sphere Σ_s. The net has size $\sim (n/s)^s$, so $\log m \sim s\log(n/s)$ — exactly the optimal CS measurement count. Random Gaussian/Bernoulli matrices thus satisfy RIP with $m = \mathcal{O}(s\log(n/s))$. The same concentration of measure that gives JL also gives McDiarmid (Ch 12), tying together random projection, sparse recovery, and learning.

Goal: Embed m points $\mathcal{U} \subset \mathbb{R}^n$ into $\mathbb{R}^k$ with $k \ll n$ while approximately preserving all pairwise distances.

Lemma 8.1 (JL)

For $\varepsilon \in (0,1)$, if

$$k \ge \frac{8}{\varepsilon^2 - \varepsilon^3}\log(2m)$$

then there exists $f : \mathbb{R}^n \to \mathbb{R}^k$ such that for all $\mathbf{u}, \mathbf{u}' \in \mathcal{U}$:

$$(1-\varepsilon)\|\mathbf{u}-\mathbf{u}'\|^2 \le \|f(\mathbf{u}) - f(\mathbf{u}')\|^2 \le (1+\varepsilon)\|\mathbf{u}-\mathbf{u}'\|^2$$

Sample complexity: $k = \mathcal{O}\big(\varepsilon^{-2}\log m\big)$. Crucially, k is *independent of* n.

4.1 Concentration Lemma

Lemma 8.2 (Concentration)

Let $\mathbf{A} \in \mathbb{R}^{k\times n}$ have i.i.d. $\mathcal{N}(0, 1/k)$ entries. For fixed $\mathbf{u} \in \mathbb{R}^n$:

$$\mathbb{P}\big(\big|\|\mathbf{A}\mathbf{u}\|^2 - \|\mathbf{u}\|^2\big| \ge \varepsilon\|\mathbf{u}\|^2\big) < 2e^{-k(\varepsilon^2-\varepsilon^3)/4}$$

with $\mathbb{E}[\|\mathbf{A}\mathbf{u}\|^2] = \|\mathbf{u}\|^2$.

Generalization: Same bound holds for sub-Gaussian entries: $\mathbb{P}(|a_{ij}| > t) \le c_1 e^{-c_2 t^2}$.

JL via concentration: For each of the $\binom{m}{2}$ pairs $\mathbf{u}-\mathbf{u}'$, apply Lemma 8.2 with the union bound. Failure probability $< 2\binom{m}{2}e^{-k(\varepsilon^2-\varepsilon^3)/4} < 1$ provided k satisfies the JL bound. So a random Gaussian $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$ works with positive probability.

4.2 Converse (Lower Bound on k)

JL is essentially tight: any ε-distortion embedding of m points requires $k = \Omega(\varepsilon^{-2}\log m/\log(1/\varepsilon))$. Proof via packing argument in Euclidean ball (exam 21-22-P3).

4.3 JL ⇒ RIP (Ch 9)

For random matrix $\mathbf{\Phi}$ satisfying JL concentration: apply to a fine ε-net of unit-norm s-sparse vectors. With $m \ge c_1 k/\log(n/k)$, RIP holds with probability $\ge 1 - 2e^{-c_2 m}$. *Gaussian/sub-Gaussian random matrices satisfy RIP with $m = \mathcal{O}(s\log(n/s))$ measurements.*

5 Approximation Theory & Metric Entropy

Why metric entropy

Step back from algorithms: *how complex is a signal class $\mathcal{C}$, intrinsically?* The minimax code length $L(\varepsilon, \mathcal{C})$ — the minimum bits to represent any $f \in \mathcal{C}$ within error ε — equals $\lceil \log_2 N(\varepsilon; \mathcal{C}, \rho)\rceil$. So **metric entropy is description complexity**. It tells us the fundamental measurement / compression / sample budget for $\mathcal{C}$, independent of any specific algorithm. Examples calibrate intuition: a unit cube $[0,1]^d$ has entropy $d\log(1/\varepsilon)$ (volume-driven), a Cantor set has fractional Minkowski dimension (self-similar), and the sparse set Σ_s has entropy $s\log(n/s)$ — *matching* the CS measurement count, which is no accident. Ch 12 uses the same machinery for function classes, replacing $\mathcal{C}$ by $\mathcal{F}$.

5.0 Min-Max (Kolmogorov) Code Length

Definition 10.1 — Minimax code length $L(\varepsilon, \mathcal{C})$

$$L(\varepsilon, \mathcal{C}) := \min\big\{\ell : \exists\, E : \mathcal{C} \to \{0,1\}^\ell,\ D : \{0,1\}^\ell \to L^2,\ \sup_{f\in\mathcal{C}} \|D(E(f)) - f\|_{L^2} \le \varepsilon\big\}$$

Minimum bits to encode any $f \in \mathcal{C}$ within error ε.

Optimal exponent:

$$\gamma^*(\mathcal{C}) := \sup\big\{\gamma : L(\varepsilon, \mathcal{C}) = \mathcal{O}(\varepsilon^{-1/\gamma})\big\}$$

Larger γ^* = simpler class.

Bridge to metric entropy: The optimal encoder is precisely the nearest-neighbor encoder on an ε-covering. Hence

$$L(\varepsilon, \mathcal{C}) = \lceil \log_2 N(\varepsilon; \mathcal{C}, \rho)\rceil$$

So **metric entropy = description complexity**; covering numbers ARE the operational quantity, not just a technical tool.

5.1 Covering and Packing Numbers

Definitions

Let $(\mathcal{X}, \rho)$ metric space, $\mathcal{C} \subset \mathcal{X}$ compact, $\varepsilon > 0$.
ε-covering: $\{x_1, \dots, x_N\} \subset \mathcal{X}$ such that $\forall x \in \mathcal{C}$, $\exists i$ with $\rho(x, x_i) \le \varepsilon$. **Covering number** $N(\varepsilon; \mathcal{C}, \rho)$ = size of smallest ε-covering.
ε-packing: $\{x_1, \dots, x_M\} \subset \mathcal{C}$ with $\rho(x_i, x_j) > \varepsilon$ for $i \neq j$. **Packing number** $M(\varepsilon; \mathcal{C}, \rho)$ = size of largest ε-packing.

Metric entropy: $\log_2 N(\varepsilon; \mathcal{C}, \rho)$ — bits needed to encode $\mathcal{C}$ to precision ε.

5.2 Covering–Packing Duality

Lemma 10.4

$$M(2\varepsilon; \mathcal{C}, \rho) \le N(\varepsilon; \mathcal{C}, \rho) \le M(\varepsilon; \mathcal{C}, \rho)$$

Reading: Covering and packing have the *same scaling* in ε up to constants. Use packing for lower bounds, covering for upper bounds.

Proof.
Left ($M(2\varepsilon) \le N(\varepsilon)$): Let $\{x_i\}$ be a maximal 2ε-packing and $\{y_j\}$ a minimal ε-covering. Each x_i lies in some ball $B(y_{j(i)}, \varepsilon)$. If $j(i) = j(i')$ for $i \neq i'$, the triangle inequality gives

$$\rho(x_i, x_{i'}) \le \rho(x_i, y_{j(i)}) + \rho(y_{j(i)}, x_{i'}) \le \varepsilon + \varepsilon = 2\varepsilon,$$

contradicting $\rho(x_i, x_{i'}) > 2\varepsilon$ from the packing. So $i \mapsto j(i)$ is injective, giving $M \le N$.
Right ($N(\varepsilon) \le M(\varepsilon)$): Let $\{x_i\}$ be a maximal ε-packing. Claim it is already an ε-covering. If not, $\exists\, x$ with $\rho(x, x_i) > \varepsilon$ for all i, making $\{x_i, x\}$ a larger packing — contradiction. Hence $N \le M$. $\square$

5.3 Volume Ratio Bound

Lemma 10.5 (Volume Bound)

For norms $\|\cdot\|, \|\cdot\|'$ on $\mathbb{R}^d$ with unit balls $\mathcal{B}, \mathcal{B}'$:

$$\left(\frac{1}{\varepsilon}\right)^d \frac{\mathrm{vol}(\mathcal{B})}{\mathrm{vol}(\mathcal{B}')} \le N(\varepsilon; \mathcal{B}, \|\cdot\|') \le \frac{\mathrm{vol}(\frac{2}{\varepsilon}\mathcal{B} + \mathcal{B}')}{\mathrm{vol}(\mathcal{B}')}$$

Same norm ($\mathcal{B} = \mathcal{B}'$): $\varepsilon^{-d} \le N(\varepsilon; \mathcal{B}, \|\cdot\|) \le (1 + 2/\varepsilon)^d$.

5.4 Minkowski Dimension

Minkowski (box-counting) Dimension

$$\dim_{\mathrm{M}}(\mathcal{C}) = \lim_{\varepsilon \to 0^+} \frac{\log N(\varepsilon; \mathcal{C}, \rho)}{\log(1/\varepsilon)}$$

when the limit exists; otherwise use $\limsup, \liminf$ for upper/lower dimensions.

Reading: If $N(\varepsilon) \asymp \varepsilon^{-d}$ then $\dim_{\mathrm{M}} = d$.

5.5 Key Examples

Interval $[-1, 1]$, $|\cdot|$: grid $x_i = -1 + 2(i-1)\varepsilon$ gives $M(\varepsilon) \ge \lfloor 1/\varepsilon \rfloor + 1$; covering: $N(\varepsilon) \le 1/\varepsilon + 1$. Hence $\log N \asymp \log(1/\varepsilon)$, $\dim_{\mathrm{M}} = 1$.

Unit cube $[0, 1]^d$, $\|\cdot\|_\infty$: $\log N(\varepsilon) \asymp d \log(1/\varepsilon)$, $\dim_{\mathrm{M}} = d$.

Euclidean ball $\mathcal{B}_2^d$: $\log N(\varepsilon; \mathcal{B}_2^d, \|\cdot\|_2) \asymp d \log(1/\varepsilon)$, $\dim_{\mathrm{M}} = d$.

Binary hypercube $\{0,1\}^d$, Hamming: $|\{0,1\}^d| = 2^d$, packing/covering depend on Hamming radius (use volume of Hamming ball $V(d, r) = \sum_{i \le r} \binom{d}{i}$).

Cantor set $C \subset [0, 1]$ (ratio 1/3): $N(3^{-n}) = 2^n$, so $\dim_{\mathrm{M}} = \log_2 / \log_3 \approx 0.631$.

5.6 Sparse Vectors

The set $\Sigma_s = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_0 \le s,\ \|\mathbf{x}\|_2 \le 1\}$ has

$$\log N(\varepsilon; \Sigma_s, \|\cdot\|_2) \asymp s \log(n/s) + s \log(1/\varepsilon)$$

(union of $\binom{n}{s}$ s-dim balls).

5.7 Toolbox for Metric Entropy Problems

Monotonicity:

- In ε: $\varepsilon \mapsto N(\varepsilon; \mathcal{C}, \rho)$ and $\varepsilon \mapsto M(\varepsilon; \mathcal{C}, \rho)$ are non-increasing.
- In set: $\mathcal{C} \subseteq \mathcal{C}' \Rightarrow N(\varepsilon; \mathcal{C}, \rho) \le N(\varepsilon; \mathcal{C}', \rho)$, same for M.

Hamming ball volume: In $\{0,1\}^d$ with Hamming distance, $|\mathcal{B}(x, m)| = \sum_{k=0}^m \binom{d}{k}$. Hence

$$\frac{2^d}{\sum_{k=0}^m \binom{d}{k}} \le N \le M \le \frac{2^d}{\sum_{k=0}^{\lfloor m/2 \rfloor} \binom{d}{k}}$$

where $N = N(m; \{0,1\}^d, d_H)$, $M = M(m; \{0,1\}^d, d_H)$.

Sauer–Shelah binomial bound: For $d \le n$:

$$\sum_{k=0}^{d} \binom{n}{k} \le \left(\frac{en}{d}\right)^d$$

Binary entropy & KL:

$$\phi(t) = -t \log t - (1-t) \log(1-t)$$
$$D(p\|q) = p \log \tfrac{p}{q} + (1-p) \log \tfrac{1-p}{1-q}$$

Note $D(\alpha \| \frac{1}{2}) = \log 2 - \phi(\alpha)$. Used in hypercube packing bounds:

$$\tfrac{1}{d} \log M(\varepsilon; \{0,1\}^d, d_H) \le D(\lfloor d\varepsilon/2 \rfloor / d \,\|\, \tfrac{1}{2}) + \tfrac{\log(d+1)}{d}$$

Self-similar / fractal scaling: If $\mathcal{C}$ splits into k disjoint copies each scaled by $\lambda < 1$, then $N(\lambda^n \varepsilon_0; \mathcal{C}) \asymp k^n$, so $\dim_{\mathrm{M}}(\mathcal{C}) = \log k / \log(1/\lambda)$.

Isometric embedding: If $V : \mathcal{C} \to \mathcal{C}'$ is a bijective isometry, then $N(\varepsilon; \mathcal{C}, \rho) = N(\varepsilon; \mathcal{C}', \rho')$ (e.g. vectorizing matrices: $\|\mathbf{A}\|_F = \|\mathrm{vec}(\mathbf{A})\|_2$).

Norm comparison: If $\|\cdot\| \le C \|\cdot\|'$ on K, then any ε-covering w.r.t. $\|\cdot\|'$ is also an ε-covering w.r.t. $\|\cdot\|$, so

$$N(\varepsilon; K, \|\cdot\|) \le N(\varepsilon/C; K, \|\cdot\|') \le N(\varepsilon; K, \|\cdot\|') \text{ if } C \ge 1$$

E.g. $\|\cdot\|_2 \le \|\cdot\|_1 \le \sqrt{n}\|\cdot\|_2$ gives $N(\varepsilon; K, \|\cdot\|_2) \le N(\varepsilon; K, \|\cdot\|_1) \le N(\frac{\varepsilon}{\sqrt{n}}; K, \|\cdot\|_2)$.

6 Uniform Laws & Learning Theory

Why uniform laws

The classical law of large numbers says $\frac{1}{n}\sum f(X_i) \to \mathbb{E}f(X)$ for any *fixed* f. Learning needs more: we choose f *after* seeing the data (empirical risk minimization), so we need the convergence to hold *uniformly* over the candidate class $\mathcal{F}$. **Three complexity measures certify this:** bracketing numbers $N_{[]}$ (classical, Ch 12.4), VC dimension (combinatorial, Ch 12.5), and Rademacher complexity $\mathcal{R}_n(\mathcal{F})$ (most flexible, Ch 12.3). Concentration of measure (Hoeffding, McDiarmid) turns a single-function bound into a uniform one. Sauer–Shelah bridges VC and covering, exactly parallel to how Ch 10 quantified signal classes. **Bottom line:** small complexity + concentration = generalization. The deep payoff is a single framework — entropy/Rademacher + concentration — covering empirical CDFs (Glivenko–Cantelli), linear classifiers, kernel methods (RKHS), and neural networks (Ledoux–Talagrand contraction for Lipschitz activations).

Setup: $X_1, \dots, X_n$ i.i.d. $\sim \mathbb{P}$. Empirical measure $\mathbb{P}_n = \frac{1}{n}\sum_i \delta_{X_i}$. Function class $\mathcal{F}$.

Uniform deviation: $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_i f(X_i) - \mathbb{E}[f(X)]\right|$.

6.1 Empirical CDF & Glivenko–Cantelli

$\hat{F}_n(t) = \frac{1}{n}|\{i : X_i \le t\}|$.

$\mathcal{F}$ is a **Glivenko–Cantelli class** for $\mathbb{P}$ if $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \xrightarrow{\text{a.s.}} 0$.

Theorem 12.6 (Glivenko–Cantelli)

For any distribution, $\|\hat{F}_n - F\|_\infty \xrightarrow{\text{a.s.}} 0$.

6.2 Rademacher Complexity

Empirical & Population Rademacher Complexity

Let $\varepsilon_i \in \{\pm 1\}$ i.i.d. uniform (Rademacher).

$$\hat{\mathcal{R}}_n(\mathcal{F}; x_1^n) = \mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\right|\right]$$

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_X[\hat{\mathcal{R}}_n(\mathcal{F}; X_1^n)]$$

Reading: $\mathcal{R}_n(\mathcal{F})$ measures how well $\mathcal{F}$ can correlate with random sign noise. Large $\mathcal{R}_n \Rightarrow$ class too rich.

Theorem 12.10 (Rademacher uniform bound)

If $\|f\|_\infty \le b$ for all $f \in \mathcal{F}$, then for all $\delta > 0$:

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \le 2\mathcal{R}_n(\mathcal{F}) + \delta$$

with probability $\ge 1 - e^{-n\delta^2/(2b^2)}$. Hence $\mathcal{R}_n(\mathcal{F}) = o(1) \Rightarrow$ Glivenko–Cantelli.

McDiarmid (bounded difference inequality)

If $\psi(X_1, \dots, X_n)$ satisfies $|\psi(\dots, x_i, \dots) - \psi(\dots, x_i', \dots)| \le L_i$ (changing one coordinate changes ψ by at most L_i), then for $\varepsilon > 0$:

$$\mathbb{P}\big(\psi - \mathbb{E}\psi > \varepsilon\big) \le \exp\Big(-\tfrac{2\varepsilon^2}{\sum_i L_i^2}\Big)$$

Rademacher concentration: $\hat{\mathcal{R}}_n$ has bounded differences $L_i = 2b/n$. By McDiarmid, w.p. $\ge 1 - \delta$:

$$\mathcal{R}_n(\mathcal{F}) \le \hat{\mathcal{R}}_n(\mathcal{F}; X_1^n) + \sqrt{\tfrac{2b^2\log(1/\delta)}{n}}$$

6.3 Rademacher Computations

Linear class: $\mathcal{F} = \{x \mapsto \langle w, x\rangle : \|w\|_2 \le W\}$ with $\|x\|_2 \le R$ a.s.:

$$\mathcal{R}_n(\mathcal{F}) \le \frac{WR}{\sqrt{n}}$$

Ledoux–Talagrand contraction: If $\phi : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz with $\phi(0) = 0$:

$$\mathcal{R}_n(\phi \circ \mathcal{F}) \le 2L\,\mathcal{R}_n(\mathcal{F})$$

(factor 2 for the absolute-supremum version used in this course; iteratively gives bounds for K-layer networks with Lipschitz activations: $\mathcal{R}_n(\mathcal{F}_K) \le \prod_{i=1}^K (2B_iL)\cdot\mathcal{R}_n(\mathcal{F}_0)$).

Convex hull: $\mathcal{R}_n(\operatorname{conv}\mathcal{F}) = \mathcal{R}_n(\mathcal{F})$.

Massart's finite-class: $|\mathcal{F}| < \infty$, $\sup_f \|f(X_1^n)\|_2 \le M$:

$$\hat{\mathcal{R}}_n(\mathcal{F}) \le \frac{M\sqrt{2\log|\mathcal{F}|}}{n}$$

Class arithmetic:

- Sum/diff: $\mathcal{R}_n(\mathcal{F}_1 \pm \mathcal{F}_2) \le \mathcal{R}_n(\mathcal{F}_1) + \mathcal{R}_n(\mathcal{F}_2)$
- Abs.: $\mathcal{R}_n(|\mathcal{F}_1 - \mathcal{F}_2|) \le 2(\mathcal{R}_n(\mathcal{F}_1) + \mathcal{R}_n(\mathcal{F}_2))$
- Max: $\mathcal{R}_n(\max\{\mathcal{F}_1, \mathcal{F}_2\}) \le \frac{3}{2}(\mathcal{R}_n(\mathcal{F}_1) + \mathcal{R}_n(\mathcal{F}_2))$ (from $\max\{a,b\} = \frac{|a-b|+a+b}{2}$)
- $y_i \in \{\pm 1\}$ Rademacher-symmetric: $\mathcal{R}_n(\{y\,f\}) = \mathcal{R}_n(\mathcal{F})$

RKHS class: For $\mathcal{F}_b = \{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}} \le b\}$ with kernel k:

$$\hat{\mathcal{R}}_n(\mathcal{F}_b) \le \tfrac{b}{n}\sqrt{\textstyle\sum_i k(x_i,x_i)} = \tfrac{b}{n}\sqrt{\operatorname{Tr}\mathbf{K}}$$

via reproducing property $f(x) = \langle f, \phi_x\rangle$ + Jensen.

6.4 VC Dimension

Shattering & VC dimension

For binary class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$: $\{x_1, \dots, x_n\}$ is **shattered** if $|\mathcal{F}(x_1^n)| = 2^n$ (all labelings realizable).
$\nu(\mathcal{F})$ = largest n such that some n-point set is shattered.

Theorem 12.17 (Sauer–Shelah)

For $\mathcal{F}$ with VC dimension $\nu < \infty$ and $n \ge \nu$:

$$|\mathcal{F}(x_1^n)| \le \sum_{i=0}^{\nu}\binom{n}{i} \le (n+1)^\nu \le \left(\frac{en}{\nu}\right)^\nu$$

VC–Rademacher link: For binary class with VC-dim ν:

$$\mathcal{R}_n(\mathcal{F}) \le C\sqrt{\frac{\nu\log(n/\nu)}{n}}$$

(Sauer–Shelah + Massart). Dudley/chaining removes the log.

6.5 VC Examples (exam-frequent)

Class	ν
Left half-intervals $(-\infty, a]$ in $\mathbb{R}$	1
Intervals $(a, b]$ in $\mathbb{R}$	2
Axis-aligned rectangles in $\mathbb{R}^d$	$2d$
Half-spaces in $\mathbb{R}^d$	$d+1$
Closed balls in $\mathbb{R}^d$	$d+1$
Linear classifiers $\text{sign}(\langle w, x\rangle - b)$ in $\mathbb{R}^d$	$d+1$
Any d-dim vector space of functions	d

Radon's theorem: Any $d+2$ points in $\mathbb{R}^d$ admit a partition $A \sqcup B$ with $\text{conv}(A) \cap \text{conv}(B) \neq \emptyset$. Used to upper-bound VC dim: convex-overlapping subsets cannot be shattered.

6.6 Generalization Bound (PAC)

With probability $\geq 1-\delta$, for all $f \in \mathcal{F}$ (binary, $\|f\|_\infty \leq 1$):

$$\mathbb{P}[f] \leq \mathbb{P}_n[f] + 2\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

For VC class: sample complexity $n = \mathcal{O}\left(\frac{\nu + \log(1/\delta)}{\varepsilon^2}\right)$ suffices for ε-uniform deviation.

6.7 Bracketing (study)

$[\ell, u]$ is an ε-bracket if $\ell \leq u$ and $\|u - \ell\|_{L^p(\mathbb{P})} \leq \varepsilon$. Bracketing number $N_{[]}(\varepsilon; \mathcal{F}, L^p)$ = min number of ε-brackets covering $\mathcal{F}$.

Finite bracketing $\Rightarrow$ GC: If $N_{[]}(\varepsilon; \mathcal{F}, L^1) < \infty$ for all $\varepsilon > 0$, then $\mathcal{F}$ is Glivenko–Cantelli.

Use: For the empirical CDF class, $N_{[]}(\varepsilon) \leq \lceil 1/\varepsilon \rceil$ (split $[0,1]$ into ε-quantile chunks), giving classical GC.

A Inequalities & Quick Reference

A.1 Notation

Symbol	Meaning
a^*	complex conjugate of scalar a
$\mathbf{a}^*$, $\mathbf{A}^*$	element-wise complex conjugate of vector/matrix
$\mathbf{A}^\top$	transpose (real)
$\mathbf{A}^{\mathrm{H}}$	Hermitian (conjugate) transpose: $(\mathbf{A}^{\mathrm{H}})_{ij} = \overline{A_{ji}}$
$\mathbb{A}^*$	adjoint of operator $\mathbb{A}$: $\langle \mathbb{A}x, y\rangle = \langle x, \mathbb{A}^* y\rangle$
$\mathbf{A}^\dagger$	Moore-Penrose pseudoinverse: $(\mathbf{A}^{\mathrm{H}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{H}}$

Bold $\mathbf{A}$ = matrix/vector; **Blackboard $\mathbb{A}$** = operator on $\mathcal{H}$. Adjoint of matrix = Hermitian transpose: $\mathbf{A}^* = \mathbf{A}^{\mathrm{H}}$ (context-dependent).

Complex conjugate rules: $a \cdot a^* = |a|^2$; $(ab)^* = a^* b^*$.

A.2 Quick Definitions

- **Complete space:** A metric space is complete if every Cauchy sequence converges within it. A complete normed space is a **Banach space**; a complete inner-product space is a **Hilbert space**.
- **Bounded operator:** $\mathbb{T} : \mathcal{H} \to \mathcal{H}'$ is bounded if $\exists\, C < \infty$ s.t. $\|\mathbb{T}x\| \leq C\|x\|$ for all x. Equivalently: continuous. The operator norm is $|||\mathbb{T}||| \triangleq \sup_{\|x\|=1} \|\mathbb{T}x\|$.
- **Adjoint:** $\mathbb{T}^*$ unique s.t. $\langle \mathbb{T}x, y\rangle = \langle x, \mathbb{T}^* y\rangle$ for all x, y; $|||\mathbb{T}^*||| = |||\mathbb{T}|||$; $\mathbb{T}^{**} = \mathbb{T}$. Matrix: $\mathbf{A}^* = \mathbf{A}^{\mathrm{H}}$.
- **Self-adjoint:** $\mathbb{A}^* = \mathbb{A}$. Complex matrix: **Hermitian**, $\mathbf{A}^{\mathrm{H}} = \mathbf{A}$, i.e. $A_{ij} = \overline{A_{ji}}$; eigenvalues real, unitarily diagonalizable. Real: **symmetric**, $\mathbf{A}^\top = \mathbf{A}$.
- **Normal:** $\mathbb{A}^*\mathbb{A} = \mathbb{A}\mathbb{A}^*$. Complex matrix: $\mathbf{A}^{\mathrm{H}}\mathbf{A} = \mathbf{A}\mathbf{A}^{\mathrm{H}}$; unitarily diagonalizable. Self-adjoint and unitary are special cases.
- **Unitary (operator):** $\mathbb{U}^*\mathbb{U} = \mathbb{U}\mathbb{U}^* = \mathbb{I}$; isometry. Complex matrix: $\mathbf{U}^{\mathrm{H}}\mathbf{U} = \mathbf{I}$, i.e. $\mathbf{U}^{-1} = \mathbf{U}^{\mathrm{H}}$. Real: **orthogonal**, $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$.
- **Spectral theorem:** Hermitian $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{H}}$, real λ_i. PSD $\Leftrightarrow$ all $\lambda_i \geq 0$. PSD invertible: $\left|\left|\left|\mathbf{A}^{-1}\right|\right|\right|_2 = 1/\lambda_{\min}$.
- **Gram matrix:** $\mathbf{G} \triangleq \mathbf{T}^{\mathrm{H}}\mathbf{T}$; entry $(k,j) = \langle \mathbf{e}_k, \mathbf{e}_j\rangle$. Hermitian, positive-definite. $\mathbf{G} = \mathbf{I}$ iff ONB.
- **Rayleigh-Ritz:** For Hermitian $\mathbf{G}$ with eigenvalues $0 < \lambda_{\min} \leq \lambda_{\max}$:

$$\lambda_{\min}\|\mathbf{x}\|^2 \leq \mathbf{x}^{\mathrm{H}}\mathbf{G}\mathbf{x} \leq \lambda_{\max}\|\mathbf{x}\|^2$$

  Applied to frames: $\lambda_{\min}\|\mathbf{x}\|^2 \leq \|\mathbf{c}\|^2 \leq \lambda_{\max}\|\mathbf{x}\|^2$. Condition number $\kappa = \lambda_{\max}/\lambda_{\min}$; $\kappa = 1$ iff ONB (Parseval).
- **Common Hilbert spaces** (each is complete; inner product induces the listed norm):

Space	Inner product $\langle\cdot,\cdot\rangle$	Norm
$\mathbb{R}^n$	$x^\top y = \sum_i x_i y_i$	$\|x\|_2$
$\mathbb{C}^N$	$y^H x = \sum_i x_i \overline{y_i}$	$\|x\|_2$
$\ell^2(\mathcal{K})$	$\sum_{k\in\mathcal{K}} a_k \overline{b_k}$, $\sum \lvert a_k\rvert^2 < \infty$	$\|a\|_{\ell^2}$
$L^2(\Omega)$	$\int_\Omega f\overline{g}\, d\mu$, $\int \lvert f\rvert^2 < \infty$	$\|f\|_{L^2}$
$\mathcal{L}^2(B)$	same as $L^2(\mathbb{R})$, restricted to $\hat{x}(f) = 0$, $\lvert f\rvert > B$	$\|\cdot\|_{L^2}$

  $\mathcal{L}^2(B) \subset L^2(\mathbb{R})$ is a closed subspace (bandlimited signals). $\mathbb{R}^n, \mathbb{C}^N$ are finite-dimensional; ℓ^2, L^2 are separable and infinite-dimensional.
- **Norm:** $\|\cdot\| : \mathcal{X} \to \mathbb{R}_{\geq 0}$ on vector space $\mathcal{X}$ over $\mathbb{F}$:
  - Positivity: $\|x\| \geq 0$; $\|x\| = 0 \Leftrightarrow x = 0$
  - Homogeneity: $\|\alpha x\| = |\alpha|\,\|x\|$ $(\alpha \in \mathbb{F})$
  - Triangle ineq.: $\|x + y\| \leq \|x\| + \|y\|$

  Inner product induces norm $\|x\| = \sqrt{\langle x, x\rangle}$; not every norm comes from an inner product (test: parallelogram law $\downarrow$).
- **Vector norms** ($\mathbf{x} \in \mathbb{K}^n$):
  - $\|\mathbf{x}\|_0 \triangleq |\{i : x_i \neq 0\}|$ (quasi-norm, sparsity)
  - $\|\mathbf{x}\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$, $1 \leq p < \infty$
  - $\|\mathbf{x}\|_\infty = \max_i |x_i|$
- **Inner product (definition):** $\langle\cdot,\cdot\rangle : \mathcal{X} \times \mathcal{X} \to \mathbb{F}$. *Convention here*: linear in 1st arg, conjugate-linear in 2nd.
  - **Conj. symmetry:** $\langle y, x\rangle = \overline{\langle x, y\rangle}$ (over $\mathbb{R}$: $\langle y, x\rangle = \langle x, y\rangle$, plain symmetry)
  - **1st arg** (same over $\mathbb{R}$ and $\mathbb{C}$): $\langle \alpha x + \beta y, z\rangle = \alpha\langle x, z\rangle + \beta\langle y, z\rangle$
  - **2nd arg — additive** (same over $\mathbb{R}$ and $\mathbb{C}$): $\langle x, y+z\rangle = \langle x, y\rangle + \langle x, z\rangle$
  - **2nd arg — scalar**: $\langle x, \alpha y\rangle = \overline{\alpha}\langle x, y\rangle$ over $\mathbb{C}$; $= \alpha\langle x, y\rangle$ over $\mathbb{R}$ (since $\overline{\alpha} = \alpha$)
  - **Positivity & definiteness:** $\langle x, x\rangle \geq 0$; $\langle x, x\rangle = 0 \Rightarrow x = 0$

  Sesquilinear (over $\mathbb{C}$): linear in 1st, conjugate-linear in 2nd (*"one-and-a-half linear": $\overline{\alpha}$ appears in 2nd*). **Bilinear** (over $\mathbb{R}$): linear in both (conjugate trivial).
- **Inner products (examples):**
  - ℓ^2: $\langle \mathbf{x}, \mathbf{y}\rangle = \sum_i x_i \overline{y_i} = \mathbf{y}^{\mathrm{H}}\mathbf{x}$; norm $\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle}$
  - $L^2(\Omega)$: $\langle f, g\rangle = \int_\Omega f(t)\overline{g(t)}\, dt$; norm $\|f\|_2 = \sqrt{\langle f, f\rangle}$
  - Real case: drop conjugate ($\overline{y_i} \to y_i$, $\overline{g} \to g$)

- **Parallelogram law & polarization:** $2\|x\|^2 + 2\|y\|^2 = \|x+y\|^2 + \|x-y\|^2$.
  A norm satisfies this $\Leftrightarrow$ it is induced by an inner product. Polarization recovers the inner product from the norm: $\mathbb{R}$: $\langle x, y\rangle = \frac{1}{4}(\|x+y\|^2 - \|x-y\|^2)$; $\mathbb{C}$: add $+\frac{i}{4}(\|x+iy\|^2 - \|x-iy\|^2)$.

- **Parseval / Bessel:** $(\mathcal{H}, \langle\cdot,\cdot\rangle)$, induced norm $\|\cdot\|$:
  - **ONB** $\{e_k\}$: $\sum_k |\langle x, e_k\rangle|^2 = \|x\|^2$
  - **Polarized**: $\langle x, y\rangle = \sum_k \langle x, e_k\rangle\overline{\langle y, e_k\rangle}$
  - **Frame** $\{g_k\}$, $0 < A \le B$: $A\|x\|^2 \le \sum_k |\langle x, g_k\rangle|^2 \le B\|x\|^2$
  - **Bessel**: $\sum_k |\langle x, g_k\rangle|^2 \le B\|x\|^2$

  Induced norm only; $A = B = 1 \Leftrightarrow$ Parseval (ONB/tight).

- **Orthogonality:** $x \perp y \Leftrightarrow \langle x, y\rangle = 0$; $\mathcal{S}^\perp = \{x : \langle x, y\rangle = 0\ \forall y \in \mathcal{S}\}$ (always closed).
  Pythagorean: $x \perp y \Rightarrow \|x+y\|^2 = \|x\|^2 + \|y\|^2$; mutually $\perp$: $\|\sum_k x_k\|^2 = \sum_k \|x_k\|^2$.

- **SVD:** $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}}$ with $\mathbf{U}, \mathbf{V}$ unitary, $\mathbf{\Sigma}$ diagonal with $\sigma_1 \ge \cdots \ge \sigma_r > 0$.

- **Matrix norms (op. & Frobenius):** For $\mathbf{A} \in \mathbb{C}^{m\times n}$,
  - $\|\mathbf{A}\|_2 = \sup_{\|x\|_2=1} \|\mathbf{A}x\|_2 = \sigma_{\max}(\mathbf{A})$
  - $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} |A_{ij}|^2} = \sqrt{\mathrm{Tr}(\mathbf{A}^{\mathrm{H}}\mathbf{A})} = \sqrt{\sum_i \sigma_i^2}$
  - $\|\mathbf{A}\|_2 \le \|\mathbf{A}\|_F \le \sqrt{r}\,\|\mathbf{A}\|_2$, $r = \mathrm{rank}\,\mathbf{A}$
  - Submult.: $\|\mathbf{AB}\|_2 \le \|\mathbf{A}\|_2\|\mathbf{B}\|_2$; $\|\mathbf{A}^{\mathrm{H}}\|_2 = \|\mathbf{A}\|_2$

- **Trace properties:**
  - Cyclic: $\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$
  - Basis-invariant: for ONB $\{e_i\}$, $\mathrm{Tr}(\mathbf{C}) = \sum_i \langle \mathbf{C}e_i, e_i\rangle$
  - $\mathrm{Tr}(\mathbf{A}^{\mathrm{H}}\mathbf{A}) = \|\mathbf{A}\|_F^2$
  - Frobenius bound: $\|\mathbf{C}\|_2^2 \le \mathrm{Tr}(\mathbf{C}\mathbf{C}^{\mathrm{H}}) = \sum_{i,j} |\langle a_i, \mathbf{C}a_j\rangle|^2$ for any ONB $\{a_i\}$.

- **Matrix Hilbert space:** $\mathbb{C}^{m\times m}$ with $\langle \mathbf{A}, \mathbf{B}\rangle = \mathrm{Tr}(\mathbf{B}^{\mathrm{H}}\mathbf{A}) = \sum_{i,j} A_{ij}\overline{B_{ij}}$ is Hilbert, $\dim m^2$, isomorphic to $\mathbb{C}^{m^2}$ via vec. Frame tools apply.

A.3 Key Inequalities

- **Cauchy–Schwarz:** (x, y in same inner product space; eq. iff $x = \lambda y$)
  - $|\langle x, y\rangle|^p \le \|x\|^p\|y\|^p$ ($p > 0$; induced norm, same space)
  - $\mathbb{R}^n/\ell^2$: $\left|\sum_k a_k b_k\right| \le \|a\|_{\ell^2}\|b\|_{\ell^2}$
  - L^2: $\left|\int fg\right| \le \|f\|_{L^2}\|g\|_{L^2}$

  Cannot: mix spaces ($x \in \ell^2$, $y \in L^2$) or use non-induced norms.

- **Hölder:** $\left|\sum_k a_k b_k\right| \le \|a\|_{\ell^p}\|b\|_{\ell^q}$, $\frac{1}{p} + \frac{1}{q} = 1$; both in same measure space.
  Key pairings: ℓ^1–ℓ^∞, ℓ^2–ℓ^2 (= C-S), $\ell^{4/3}$–ℓ^4.

- **ℓ^p norm inequalities ($x \in \mathbb{R}^n$):** For $1 \le p \le q$: $\|x\|_{\ell^q} \le \|x\|_{\ell^p} \le n^{1/p-1/q}\|x\|_{\ell^q}$.
  Special cases: $\|x\|_{\ell^\infty} \le \|x\|_{\ell^2} \le \sqrt{n}\,\|x\|_{\ell^\infty}$; $\|x\|_{\ell^2} \le \|x\|_{\ell^1} \le \sqrt{n}\,\|x\|_{\ell^2}$.

- **Triangle inequality** (any normed space; $|\cdot|$ on $\mathbb{R}$ or $\mathbb{C}$ is a valid norm):
  - Finite: $\left\|\sum_{i=1}^n x_i\right\| \le \sum_{i=1}^n \|x_i\|$
  - Infinite (Banach): $\left\|\sum_{i=1}^\infty x_i\right\| \le \sum_{i=1}^\infty \|x_i\|$
  - Integral (Minkowski): $\left\|\int f(t)\,dt\right\| \le \int \|f(t)\|\,dt$
  - Reverse (2 terms): $\big|\|x\| - \|y\|\big| \le \|x-y\|$; no clean $n \ge 3$ form

- **AM-GM / Young:** $\sqrt{ab} \le \frac{a+b}{2}$; general: $(\prod_i a_i)^{1/n} \le \frac{1}{n}\sum_i a_i$ ($a_i \ge 0$).
  Useful: $\|u\|\|v\| \le \frac{1}{2}(\|u\|^2 + \|v\|^2)$. **Young's:** $ab \le \frac{a^p}{p} + \frac{b^q}{q}$, $\frac{1}{p} + \frac{1}{q} = 1$.

- **Jensen:** For convex ϕ: $\phi(\mathbb{E}[X]) \le \mathbb{E}[\phi(X)]$. For concave ϕ (e.g. $\sqrt{\cdot}$): $\mathbb{E}[\phi(X)] \le \phi(\mathbb{E}[X])$, so $\mathbb{E}[\sqrt{X}] \le \sqrt{\mathbb{E}[X]}$.

- **Markov / Chebyshev:** $\mathbb{P}(|X| \ge t) \le \mathbb{E}|X|/t$; $\mathbb{P}(|X - \mathbb{E}X| \ge t) \le \mathrm{Var}(X)/t^2$.

- **Hoeffding** ($\to$ Sec. 8): If $X_1, \ldots, X_m$ i.i.d. bounded,
$$\mathbb{P}\Big(\Big|\tfrac{1}{m}\sum X_i - \mathbb{E}[X]\Big| > \varepsilon\Big) \le 2e^{-2m\varepsilon^2}$$

- **Union bound:** $\mathbb{P}(\bigcup_i A_i) \le \sum_i \mathbb{P}(A_i)$.

- **Geometric series:** $\sum_{k=0}^{N-1} r^k = \frac{1-r^N}{1-r}$ ($r \ne 1$); $\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$ ($|r| < 1$); $\sum_{k=m}^{\infty} r^k = \frac{r^m}{1-r}$.
  E.g. $\sum_{k=1}^\infty (1/2)^k = \frac{1/2}{1-1/2} = 1$.

- **Abel summation** (summation by parts; $A_n := \sum_{k=1}^n a_k$):
$$\sum_{n=1}^{N} a_n b_n = A_N b_N - \sum_{n=1}^{N-1} A_n(b_{n+1} - b_n)$$
  Telescoping form (exam_23):
  $\sum_{n=1}^N (a_n - a_{n-1})b_n = a_N b_N - a_0 b_1 - \sum_{n=1}^{N-1} a_n(b_{n+1} - b_n)$
  Trigger: differences on wrong factor — swap via Abel.

A.4 Fourier Reference

Convention (lecture notes): frequency in Hz, $e^{-i2\pi ft}$.

The four forms:

- **CT-FT:** $\hat{x}(f) = \int x(t)e^{-i2\pi ft}dt$; $x(t) = \int \hat{x}(f)e^{i2\pi ft}df$
- **DTFT:** $\hat{x}(\theta) = \sum_k x[k]e^{-ik\theta}$; $x[n] = \frac{1}{2\pi}\int_0^{2\pi} \hat{x}(\theta)e^{in\theta}d\theta$
- **DFT:** $\hat{x}[k] = \frac{1}{\sqrt{N}}\sum_n x[n]\omega_N^{kn}$; $x[n] = \frac{1}{\sqrt{N}}\sum_k \hat{x}[k]\omega_N^{-kn}$
- **FS:** $c_n = \frac{1}{T}\int_0^T x(t)e^{-i2\pi nt/T}dt$; $x(t) = \sum_n c_n e^{i2\pi nt/T}$

Universal properties (hold for all four forms):

- **Linearity:** $\mathcal{F}\{\alpha x + \beta y\} = \alpha\hat{x} + \beta\hat{y}$
- **Time shift:** $x(t - t_0) \longleftrightarrow e^{-i2\pi ft_0}\hat{x}(f)$
- **Modulation:** $e^{i2\pi f_0 t}x(t) \longleftrightarrow \hat{x}(f - f_0)$
- **Convolution:** $(x * y) \longleftrightarrow \hat{x}\cdot\hat{y}$; product $\longleftrightarrow$ convolution (dual)

Form-specific:

- **Plancherel (CT-FT, L^2):** $\langle f, g\rangle_{L^2} = \langle \hat{f}, \hat{g}\rangle_{L^2}$, $\|f\|_{L^2} = \|\hat{f}\|_{L^2}$; FT unitary on $L^2(\mathbb{R})$. *Use:* switch time $\leftrightarrow$ freq in norm/frame proofs.
- **Parseval (DTFT):** $\sum_k |a_k|^2 = \frac{1}{2\pi}\int_0^{2\pi} |\hat{a}(\theta)|^2 d\theta$.
- **Parseval (DFT):** $\sum_{n=0}^{N-1} |x[n]|^2 = \sum_{k=0}^{N-1} |\hat{x}[k]|^2$ (from unitarity of $\mathbf{F}_N$).
- **Poisson summation:** $\sum_{k\in\mathbb{Z}} x(k) = \sum_{k\in\mathbb{Z}} \hat{x}(k)$. *Sampling form:*
$$\hat{x}_d(f) = \frac{1}{T}\sum_{k\in\mathbb{Z}} \hat{x}\left(\frac{f+k}{T}\right)$$
  Sampling $\Rightarrow$ periodization; aliasing iff copies overlap.
- **DFT matrix $\mathbf{F}_N$:** $[\mathbf{F}_N]_{kn} = \frac{1}{\sqrt{N}}\omega_N^{kn}$, unitary ($\mathbf{F}_N\mathbf{F}_N^H = \mathbf{I}$). Columns form ONB for $\mathbb{C}^N$: $\mathbf{x} = \mathbf{F}_N^H\hat{\mathbf{x}} = \sum_\ell \langle \mathbf{x}, f_\ell\rangle f_\ell$.
- **Oversampled DFT ($M > N$):** columns of $F_o \in \mathbb{C}^{M\times N}$ (first N cols of $\mathbf{F}_M$) form tight frame for $\mathbb{C}^N$ with $A = B = N/M$.

Bandlimited: $x \in \mathcal{L}^2(B)$ iff $\hat{x}(f) = 0$ for $|f| > B$. Nyquist: $1/T \ge 2B$.

$$x(t) = 2BT\sum_{k\in\mathbb{Z}} x(kT)\,\mathrm{sinc}(2B(t - kT))$$