Alternative Models

State Space Machines (SSMs)

ELL8299 · ELL881 · AIL861

Sourish Dasgupta
Associate Professor, DAU, Gandhinagar
https://daiict.ac.in/faculty/sourish-dasgupta

State Space Machines – *Language as a Diffusive Field*

Core idea. Instead of updating memory in discrete jumps (as in RWKV or LSTM), a State Space Model (SSM) treats hidden meaning as a *continuously evolving field*:

$$\frac{dx}{dt} = Ax(t) + Bu(t).$$

- $x(t)$ — the latent semantic field (what the model "feels" at any instant)
- $Ax(t)$ — how that field drifts or decays on its own (internal physics)
- $Bu(t)$ — how the current token *nudges* or perturbs the field

Example: "John loves Mary who lives in New York City."

- "John" introduces a subject wave — it starts the semantic field.
- "loves" injects a relation pulse, slightly reshaping the field.
- "Mary" adds a strong entity trace that propagates forward.
- Between "Mary" and "who", the field *morphs smoothly* — "Mary" shifts from object (of loves) to subject (of lives).

Takeaway. SSMs view language as a *fluid process*: meaning doesn't jump from token to token — it **flows continuously**, evolving and fading like ripples in a pond.

What exactly is "*smoothness of meaning*"

In an SSM, the state evolves continuously:

$$\frac{dx}{dt} = Ax + Bu.$$

If A is stable ($\mathrm{Re}\lambda_i(A) \leq 0$):

$$x(t+\Delta) = e^{A\Delta}x(t)$$

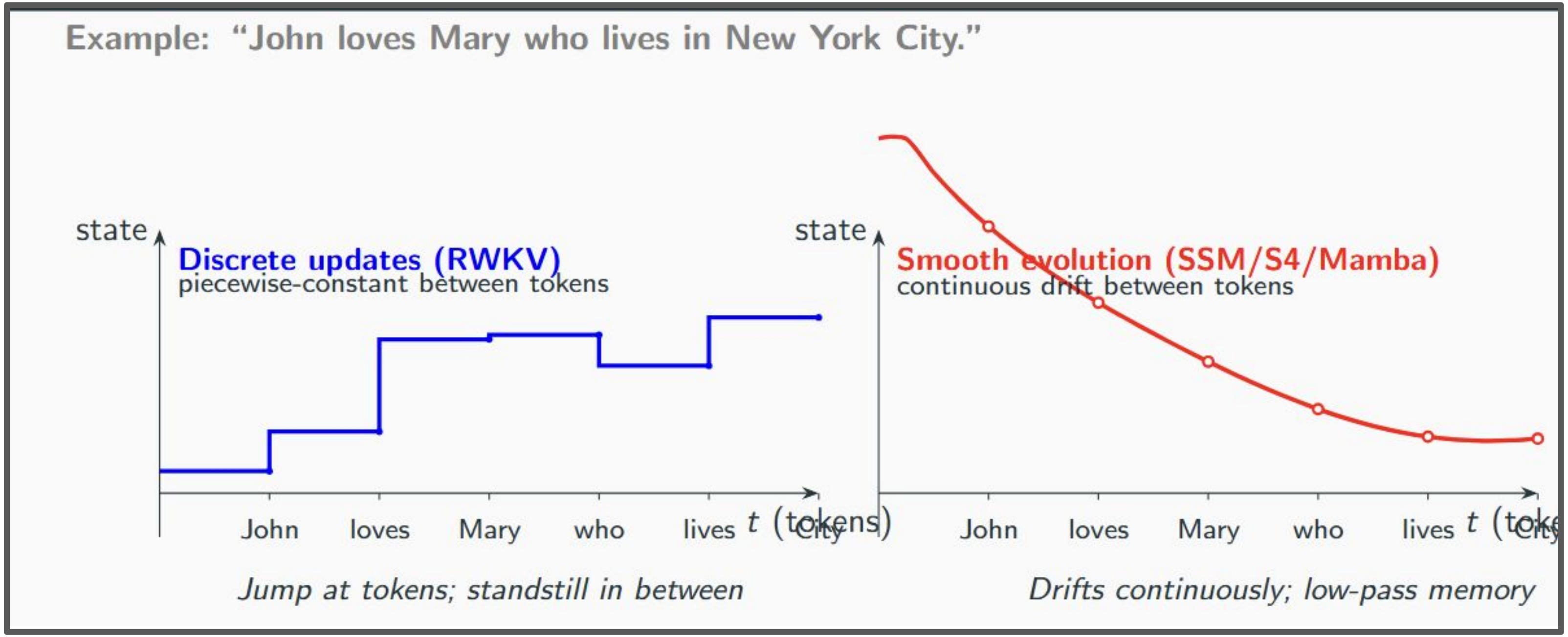
is an **analytic function** of t — i.e., infinitely differentiable and without jumps.

Implications.

- The state changes gradually; no abrupt token-boundary jumps.
- Rate of change bounded by $\|A\|$: $\|\dot{x}(t)\| \leq \|A\|\|x(t)\| + \|B\|\|u(t)\|$.
- Like clay reshaping slowly — not snapping to a new form.

Example. After "Mary", the representation smoothly morphs toward "who lives..." instead of resetting at each word.

RWKV (and RNN-styled models) vs. SSMs



- In RNNs/RWKV, memory changes only at discrete steps → “snapshot updates.”
- In SSMs, memory evolves continuously → “fluid updates.”

  This makes them naturally suited for:

  - Streaming signals (speech, video),
  - Long-term dependencies,
  - and modeling human-like temporal smoothness in thought and language.

Solving the continuous case …

In the continuous-time equation

$$\frac{dx}{dt} = Ax + Bu, \qquad x(t+\Delta) = e^{A\Delta}x(t) + \int_0^{\Delta} e^{A\tau} B\, u(t+\Delta-\tau)\, d\tau,$$

Intuition: information diffuses and decays smoothly; tokens *nudge* a field that keeps evolving in between.

The exponential integration factor: The Scalar Intuition

For the scalar case $\dot{x} = ax + bu$, multiply by e^{-at} (integrating factor):

$$e^{-at}\dot{x} = ae^{-at}x + be^{-at}u \quad \Rightarrow \quad \frac{d}{dt}\left(e^{-at}x(t)\right) = b\,e^{-at}u(t).$$

Integrate on $[t,\ t+\Delta]$:

$$e^{-a(t+\Delta)}x(t+\Delta) - e^{-at}x(t) = \int_t^{t+\Delta} b\,e^{-as}u(s)\,ds.$$

Solve for $x(t+\Delta)$:

$$x(t+\Delta) = e^{a\Delta}x(t) + \int_t^{t+\Delta} e^{a(t+\Delta-s)}\,b\,u(s)\,ds.$$

The exponential integration factor: The Matrix Form

For constant A and step Δ,

$$\boxed{\frac{d}{d\Delta} e^{A\Delta} = A\, e^{A\Delta} = e^{A\Delta} A.}$$

$$e^{A\Delta} = \sum_{k=0}^{\infty} \frac{(A\Delta)^k}{k!} = I + A\Delta + \frac{A^2\Delta^2}{2!} + \frac{A^3\Delta^3}{3!} + \cdots$$

$$\int_0^{\Delta} e^{A\tau}\, d\tau = \sum_{k=0}^{\infty} \frac{A^k\, \Delta^{k+1}}{(k+1)!} = \Delta I + \frac{A\Delta^2}{2!} + \frac{A^2\Delta^3}{3!} + \cdots$$

Solving the SSM ODE - I

Start with $\dot{x} = Ax + Bu$. Let $M(\tau) = e^{-A\tau}$. Then $\dot{M}(\tau) = -A\,M(\tau)$.

$$\frac{d}{d\tau}(M(\tau)x(\tau)) = \dot{M}(\tau)x(\tau) + M(\tau)\dot{x}(\tau) = (-AM)x + M(Ax + Bu) = M(\tau)B\,u(\tau).$$

Integrate from $\tau = t$ to $\tau = t + \Delta$:

$$e^{-A(t+\Delta)}x(t+\Delta) - e^{-At}x(t) = \int_t^{t+\Delta} e^{-As}B\,u(s)\,ds.$$

Left-multiply by $e^{A(t+\Delta)}$:

$$x(t+\Delta) = e^{A(t+\Delta)}e^{-At}x(t) + \int_t^{t+\Delta} e^{A(t+\Delta)}e^{-As}B\,u(s)\,ds.$$

Using constancy of A: $e^{A(t+\Delta)}e^{-At} = e^{A\Delta}$ and $e^{A(t+\Delta)}e^{-As} = e^{A((t+\Delta)-s)}$.

Solving the SSM ODE - II

We have the equivalent "s-domain" expression:

$$x(t+\Delta) = e^{A\Delta}x(t) + \int_{s=t}^{t+\Delta} e^{A((t+\Delta)-s)}B\,u(s)\,ds.$$

Change variables: $\tau = (t+\Delta) - s \Rightarrow s = (t+\Delta) - \tau,\ ds = -d\tau$. When $s = t \Rightarrow \tau = \Delta$, and when $s = t+\Delta \Rightarrow \tau = 0$.

$$\int_{t}^{t+\Delta} e^{A((t+\Delta)-s)}B\,u(s)\,ds = \int_{\Delta}^{0} e^{A\tau}B\,u(t+\Delta-\tau)\,(-d\tau) = \int_{0}^{\Delta} e^{A\tau}B\,u(t+\Delta-\tau)\,d\tau.$$

Final closed form (Duhamel's formula):

$$\boxed{x(t+\Delta) = e^{A\Delta}x(t) + \int_{0}^{\Delta} e^{A\tau}\,B\,u(t+\Delta-\tau)\,d\tau}.$$

Some special cases

No input $u \equiv 0$: $x(t+\Delta) = e^{A\Delta}x(t)$.

No dynamics $A = 0$: $x(t+\Delta) = x(t) + \int_0^{\Delta} B\, u(t+\Delta-\tau)\, d\tau$.

Infinitesimal step $\Delta \to 0^+$: $e^{A\Delta} \approx I + A\Delta$ and

$$x(t+\Delta) - x(t) \approx \Delta\,(Ax(t) + Bu(t)),$$

recovering $\dot{x} = Ax + Bu$.

Constant input $u(\cdot) \equiv u_0$:

$$x(t+\Delta) = e^{A\Delta}x(t) + \left(\int_0^{\Delta} e^{A\tau}\, d\tau\right) B\, u_0, \quad \int_0^{\Delta} e^{A\tau}\, d\tau = \begin{cases} A^{-1}(e^{A\Delta} - I), & A \text{ invertible}, \\ \text{series/limit form}, & \text{general } A. \end{cases}$$

For a small Delta …

Retaining up to $O(\Delta^2)$:

$$x(t+\Delta) \approx \left(I + A\Delta + \tfrac{A^2\Delta^2}{2}\right)x(t) + \left(\Delta I + \tfrac{A\Delta^2}{2}\right)B\, u_t.$$

Subtract $x(t)$ and divide by Δ:

$$\frac{x(t+\Delta) - x(t)}{\Delta} \approx Ax(t) + Bu_t + O(\Delta),$$

which recovers $\dot{x} = Ax + Bu$ as $\Delta \to 0$.

How to interpret Delta?

Between "Mary" and "who",

- RWKV simply waits — discrete update at the next token.
- SSM treats that interval as a smooth semantic drift: the meaning of "Mary" gently morphs from object of "loves" to subject of "lives."

Δ measures this conceptual span, not real seconds.

What Delta implies …

Example: "John loves Mary who lives in New York City."

Mathematical view.

$$x(t+\Delta) = e^{A\Delta}x(t) + \int_0^{\Delta} e^{A\tau} B\, u(t+\Delta-\tau)\, d\tau$$

Δ controls how long the internal state $x(t)$ evolves before the next token arrives.

Linguistic view. Δ acts as a *semantic timestep* — how far meaning drifts between words.

What Delta implies …

Scenario	Δ meaning	Linguistic effect / Example
Normal flow	$\Delta = 1$ per token	Smooth reading; steady semantic pace. "*John loves Mary who lives in New York City.*"
Short pause	Small $\Delta > 1$	Slight hesitation — previous concept evolves. After "Mary," the model lets her role drift from
Rapid speech / dense phrase	$\Delta < 1$	Tokens arrive faster than the model relaxes; meanings overlap (e.g., "in New York").
Irregular phrasing / punctuation	Variable Δ_t	Commas, conjunctions, or full stops create variable semantic distances — different "tempos" of meaning flow.

Intuition. Information diffuses and decays smoothly; each token *nudges* a continuously evolving semantic field — faster, slower, or paused depending on Δ.

What happens if the *flow* stops?

Case A: Computational stop. No new token $\Rightarrow$ the model halts; $x_{T+1} = x_T$. The semantics freeze, like pausing a movie frame.

Case B: Conceptual pause. If we imagine continuous evolution with $u(t) = 0$,

$$\frac{dx}{dt} = Ax \quad \Rightarrow \quad x(t + \Delta) = e^{A\Delta} x(t).$$

In practice, language models freeze state (Case A), but the math allows Case B—useful for continuous signals like audio.

How to interpret A …

A governs how internal meaning *drifts* when no input arrives:

$$\frac{dx}{dt} = Ax, \quad \Rightarrow \quad x(t+\Delta) = e^{A\Delta}x(t).$$

A deeper dive into A

If A is diagonalizable:

$$A = Q\Lambda Q^{-1}, \qquad e^{A\Delta} = Qe^{\Lambda\Delta}Q^{-1}.$$

Each eigenvalue λ_i defines one **mode of evolution**:

$$x_i(t) = c_i e^{\lambda_i t}.$$

Interpretation. Each eigenvector q_i is a *semantic direction* — a dimension of meaning (entity, topic, rhythm, etc.). Its eigenvalue λ_i determines how that semantic aspect changes over time.

Example: "John loves Mary who lives in New York City."

Let $\lambda_i = a_i + ib_i$. Then:

$$e^{\lambda_i t} = e^{a_i t}(\cos b_i t + i \sin b_i t).$$

Decomposition.

- Real part $a_i = \mathrm{Re}(\lambda_i) \to$ exponential growth/decay of amplitude.
- Imaginary part $b_i = \mathrm{Im}(\lambda_i) \to$ oscillation (rotation) in phase.

Linguistic intuition.

- $a_i < 0 \to$ memory decays: "Mary" eventually fades.
- $a_i = 0 \to$ memory persists: "John" stays in focus across the sentence.
- $b_i \neq 0 \to$ rhythm/recurrence: subject–verb–object or syntactic cycles reappear.

Together: each eigenvalue encodes *how fast* and *how rhythmically* meaning evolves.

Example: "John loves Mary who lives in New York City."

- q_1: the *entity mode* — tracks proper nouns like "Mary" or "City", → small negative real part (slow decay).
- q_2: the *predicate mode* — tracks ongoing relations like "loves", "lives", → medium negative real part (decays faster).
- q_3: the *syntactic rhythm mode* — organizes clause transitions like "who", → complex eigenvalue (oscillatory behavior).
- q_4: the *function word mode* — glues structure ("in", "the"), → large negative real part (fast fade).

Then the total state $x(t)$ is just the *sum* of these modes' contributions:

$$x(t) = \sum_i \underline{c_i e^{\lambda_i t} q_i}.$$

Example: "John loves Mary who lives in New York City."

Think of e^{At} as an orchestra:

- Each eigenvector q_i is an instrument (a semantic component).
- $\text{Re}(\lambda_i)$ = how quickly that instrument's note fades.
- $\text{Im}(\lambda_i)$ = how often it repeats (its rhythm).

Running example.

- "Mary" $\rightarrow$ low-frequency mode (slow decay, persistent topic).
- "in", "the" $\rightarrow$ high-frequency, fast-decaying modes.
- "who lives in" $\rightarrow$ oscillatory mid-range (syntactic pattern).

What does it mean for A to be diagonalizable?

$$A = Q\Lambda Q^{-1}, \qquad e^{At} = Qe^{\Lambda t}Q^{-1}$$

$$\text{State evolution:} \quad x(t) = \sum_i c_i\, e^{\lambda_i t}\, q_i$$

Mathematical object	Linguistic interpretation (semantic mode)
Eigenvector q_i	An independent semantic channel (e.g., entity, predicate, syntax, function)
Eigenvalue $\lambda_i = a_i + ib_i$	Temporal behavior of that channel: a_i = decay/persistence, b_i = rhythm/oscillation
Coefficients c_i	How much each mode is present for the current sentence/context
$A = Q\Lambda Q^{-1}$	Memory splits into separable threads that evolve independently
Diagonalizable A	Clear, disentangled roles; interpretable time-scales per semantic aspect
Non-diagonalizable A	Entangled dynamics; harder to attribute roles to subspaces

Why diagonalizability matters?

- Easy to exponentiate: $e^{At} = Qe^{\Lambda t}Q^{-1}$.
- Each mode $e^{\lambda_i t}$ evolves independently.
- If not diagonalizable: modes couple and cause mixed dynamics.

How to ensure diagonalizability?

Not every matrix A is diagonalizable — but we can **design** it to be.

A matrix A is diagonalizable if

$$A = Q\Lambda Q^{-1}, \qquad \text{where } Q \text{ has } n \text{ linearly independent eigenvectors.}$$

That happens when eigenvalues are distinct or A is symmetric / normal.

Key guarantees:

Method	**Condition on A**	**Guarantee**
Distinct eigenvalues	All λ_i unique	Full independent eigenbasis.
Symmetric / Hermitian	$A = A^\top$ or $A = A^*$	Orthogonal diagonalization.
Normal matrices	$AA^* = A^*A$	Unitary diagonalization.
Explicit spectrum form	$A = Q\,\mathrm{diag}(\lambda)\,Q^{-1}$	Diagonalizable by construction.
Structured bases (HiPPO, Legendre)	Polynomial projection operators	Proven full-rank eigenspaces in $\mathbb{C}$.

How to ensure diagonalizability?

Modern SSMs and RWKV-like models **guarantee diagonalizability by construction**.

Model	**How A is constrained**	**Effect**
RWKV	$A = -wI$ (scalar per channel)	Trivially diagonal; independent exponential drifts.
S4 / S4D	$A = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ (complex)	Multi-mode decays / oscillations, diagonal by design.
Mamba	$A(u_t) = A_0 + \mathrm{diag}(a_t)$	Token-dependent diagonal; adaptive, still diagonalizable.

Interpretation. Diagonalizability ensures A acts like a set of independent *semantic resonators*: each eigenvalue controls its own memory rhythm, and all combine linearly to form meaning.

Coming up …

SSM connection.

- **S4:** learns multiple stable modes (diverse λ_i) — smooth long-term dynamics.
- **Mamba:** makes λ_i depend on the current token — selective temporal rhythm.

$\rightarrow$ *The real parts control* **memory**, *the imaginary parts control* **structure**.

Questions?