In-context denoising with one-layer transformers

Connections between attention and associative memory retrieval

Matthew Smart

Center for Computational Biology
Flatiron Institute, New York, USA

July 17, 2025 | Vancouver, Canada | ICML

FLATIRON INSTITUTE

In-context denoising with one-layer transformers

Connections between attention and associative memory retrieval

Joint work with

Alberto Bietti
Flatiron Institute (CCM)

Anirvan Sengupta
Flatiron Institute (CCM + CCQ)
Rutgers (Physics)

Matthew Smart

Center for Computational Biology
Flatiron Institute, New York, USA

July 17, 2025 | Vancouver, Canada | ICML

FLATIRON INSTITUTE

Two (seemingly) unrelated architectures

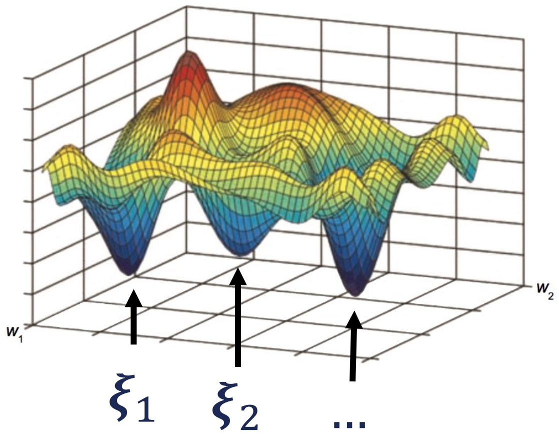
I - Associative memory networks

Classical formulation (1982, Hopfield)

Energy function $E(\mathbf{x}) = -\mathbf{x}^{\mathrm{T}}\mathbf{J}\,\mathbf{x}$

$\mathbf{x} \in [-1,+1]^n$ $\quad = -\sum_{\mu=1}^{p}\left(\boldsymbol{\xi}_\mu^{\mathrm{T}}\,\mathbf{x}\right)^2$

1. Select binary patterns $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_p$
2. Induce couplings $\mathbf{J} = \sum_{\mu=1}^{p} \boldsymbol{\xi}_\mu \boldsymbol{\xi}_\mu^{\mathrm{T}}$
3. Patterns are fixed points of $\mathbf{x}_{t+1} = \mathrm{sgn}(\mathbf{J}\mathbf{x}_t)$



Recent generalizations

Dense associative memory

$$E(\mathbf{x}) = -\sum_{\mu=1}^{p} F\left(\boldsymbol{\xi}_\mu^{\mathrm{T}}\mathbf{x}\right)$$

2016, Krotov & Hopfield $F = m^k$

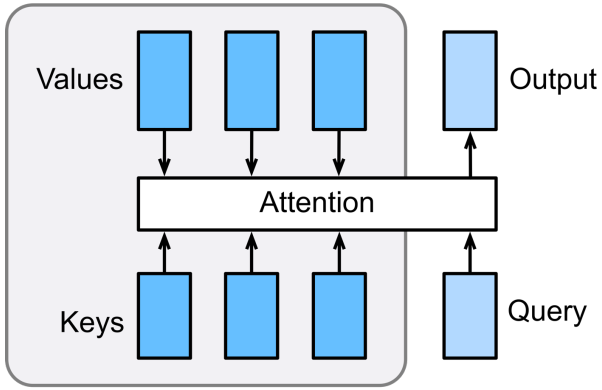
2017, Demircigil et al. $F = e^m$

Sharper nonlinearities:
→ enhanced storage capacity

II - Attention layer (transformers)

$$\mathbf{f}(\mathbf{X}) = \mathbf{V}\,\mathrm{softmax}\left(\frac{1}{\sqrt{d_k}}\mathbf{K}^{\mathrm{T}}\mathbf{Q}\right)$$

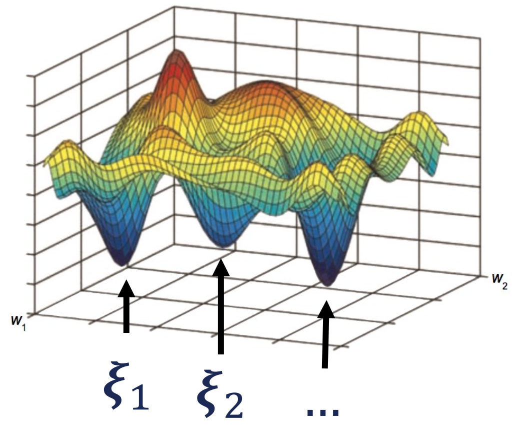
- Prompt (context tokens): $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_L] \in \mathbb{R}^{n\times L}$
- Let $\mathbf{W}_V$, $\mathbf{W}_K$, $\mathbf{W}_Q$ - learnable attention weights
- Define $\mathbf{V} = \mathbf{W}_V\mathbf{X}$, $\mathbf{K} = \mathbf{W}_K\mathbf{X}$, $\mathbf{Q} = \mathbf{W}_Q\mathbf{X}$,

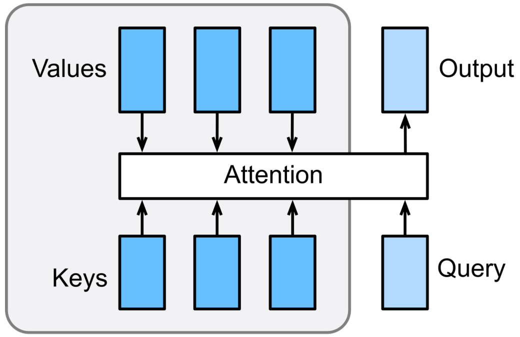


Attention is a key ingredient in transformers

Despite empirical success, underlying mechanisms remain unclear

Background: Bridging associative memory and attention





Transformer attention update can be viewed as one step of dense associative memory dynamics.

"Hopfield networks is All You Need"
Ramsauer, ... , Hochreiter, ICLR 2021

Idea: Start from "Modern Hopfield network" with continuous state $\mathbf{x} \in \mathbb{R}^n$

$$E(\mathbf{x}) = -\beta^{-1} \log \left(\sum_{\mu=1}^{p} e^{\beta \boldsymbol{\xi}_\mu^{\mathrm{T}} \mathbf{x}} \right) + \frac{1}{2} \mathbf{x}^{\mathrm{T}} \mathbf{x}$$

$$\nabla_{\mathbf{x}} \downarrow$$

$$\mathbf{x}_{t+1} = \boldsymbol{\xi}\, \mathrm{softmax}(\boldsymbol{\xi}^{\mathrm{T}} \mathbf{x}_t) \approx \mathbf{V}\, \mathrm{softmax}\left(\frac{1}{\sqrt{d_k}} \mathbf{K}^{\mathrm{T}} \mathbf{Q} \right)$$

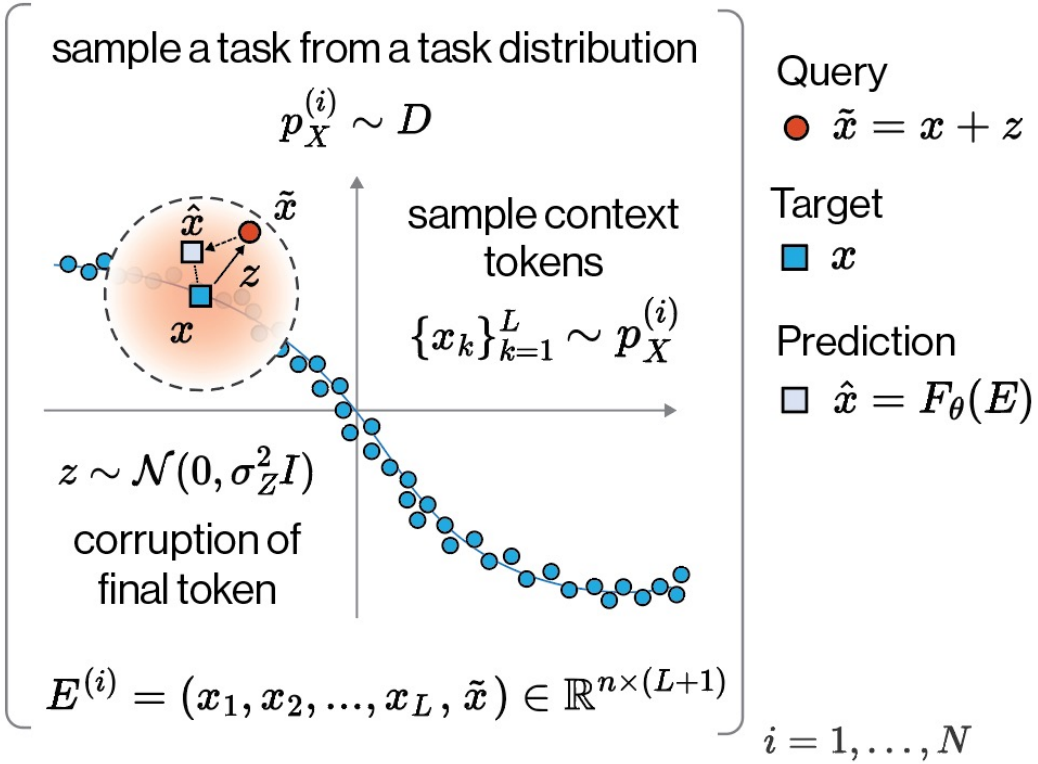
where $\boldsymbol{\xi} \equiv [\boldsymbol{\xi}_1 \ \cdots \ \boldsymbol{\xi}_p]$ $(n \times p)$

Attention update

<u>Motivation</u>:
How can this connection be further developed?

We propose a natural interface:
"In-context denoising"

In-context denoising – Overview



Steps

1. Select a data distribution p_X from a given class $\mathcal{D}$
2. Sample $L+1$ points ("pure" context tokens)
3. Corrupt the final token: $\tilde{x} \sim p_{\text{noise}}(\cdot \mid x_{L+1})$
4. Construct embedding $E = (\text{context}, \text{query})$.
   Use this to estimate the target $F_\theta(E) \mapsto \hat{x}$

Repeat to generate many ICL training pairs $(E^{(i)}, x_{L+1}^{(i)})$

In-context learning

Each prompt is a new, random task

see e.g.

Garg et al. NeurIPS 2022
Zhang et al. JMLR 2024

Objective: find denoiser $F_\theta(E)$ that minimizes MSE

$$\min_\theta \mathbb{E}\left[\|X_{L+1} - F_\theta(E)\|^2\right]$$

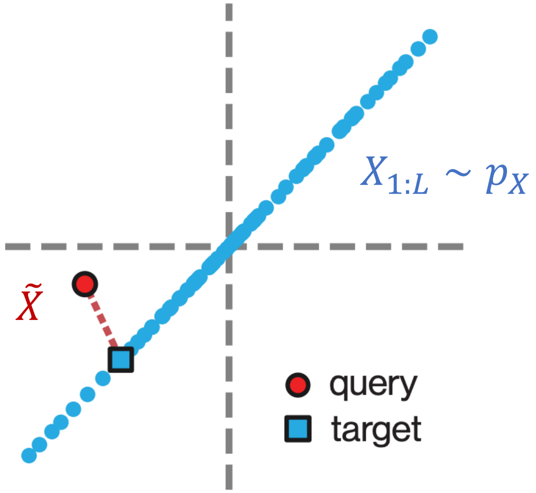
where expectation is over: $p_X \sim \mathcal{D},\ \ X_{1:L} \sim p_X,\ \ \tilde{X} \sim p_{\text{noise}}(\cdot \mid X_{L+1})$

FLATIRON INSTITUTE

In-context denoising – Cases

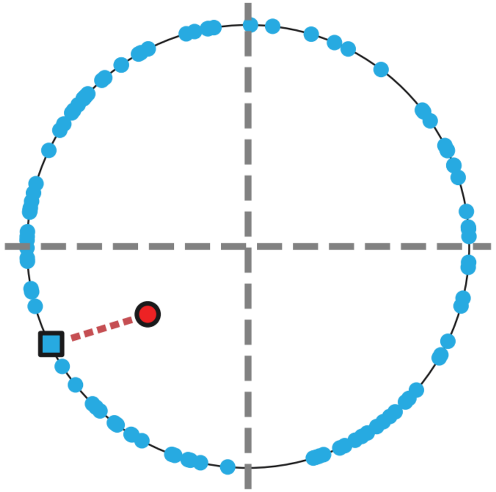
<u>Prompt:</u> Pure tokens from a data distribution and a single corrupted example

(prompts are randomly constructed from a pre-specified task distribution)



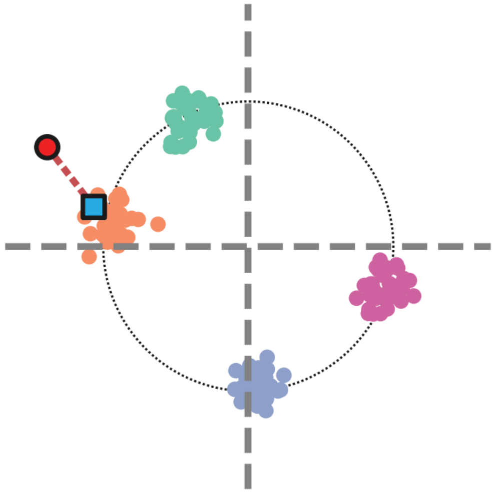
Case 1:
Linear manifolds

dimension d, sample variance σ_0^2



Case 2:
Nonlinear manifolds

d- spheres, radius R



Case 3:
Gaussian mixtures

p components: $\mathcal{N}(\mu_\alpha, \sigma_0^2 I)$

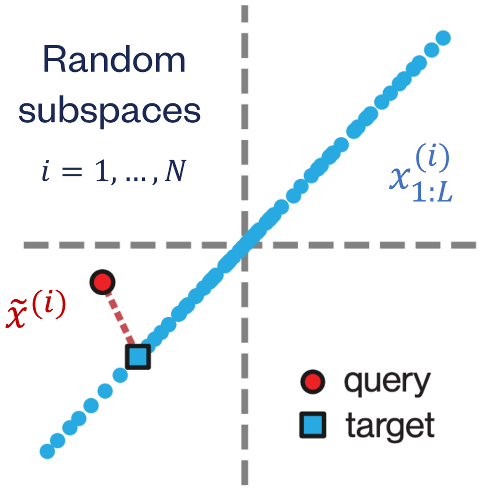
Objective: find denoiser $F_\theta(E)$ with $E = (\text{context, query})$ that minimizes MSE

$$\min_\theta \mathbb{E}\left[\|X_{L+1} - F_\theta(E)\|^2\right]$$

where expectation is over: $p_X \sim \mathcal{D}, \;\; X_{1:L} \sim p_X, \;\; \tilde{X} \sim p_{\text{noise}}(\cdot \mid X_{L+1})$

FLATIRON INSTITUTE

Linear case – Details and optimal predictor



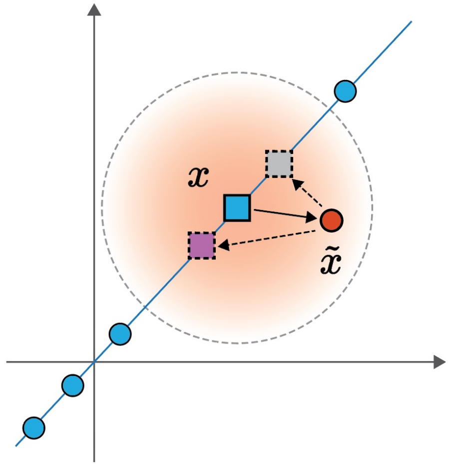
Goal: We seek a denoiser

$$f(\tilde{X}) \colon \mathbb{R}^n \to \mathbb{R}^n$$

that minimizes MSE:

$$C = \mathbb{E}_{X,\tilde{X}}\left[\left\|X - f(\tilde{X})\right\|^2\right]$$

Use knowledge of p_X, p_{noise} to derive the baseline



Linear baselines

Projection

$$\hat{x} = P^{(i)}\tilde{x}$$

Projection (shrunk)

$$\hat{x} = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_Z^2} P^{(i)}\tilde{x}$$

Dataset generation

Randomly select a d-dim subspace $S^{(i)}$ of $\mathbb{R}^n$

Let $P^{(i)}$ be the projection onto $S^{(i)}$

Sample $L+1$ tokens, corrupt final one

$X_t \sim P^{(i)} Y_t$ where $Y_t \sim N(0, \sigma_0^2 I)$

$\tilde{X} \sim X + Z$ where $Z \sim N(0, \sigma_Z^2 I)$

Parameters: $L, d, \sigma_0^2, \sigma_Z^2$

Heuristic approach

(linear case only!)

Ansatz – $f(\tilde{x}) = V\tilde{x}$

Plug in to objective, differentiate, solve

$$\begin{aligned} V_{\text{opt}} &= \operatorname{argmin}_V \mathbb{E}\left[\left\|X - V\tilde{X}\right\|^2\right] \\ &= \dots \\ &= \gamma \sigma_0^2 P \quad \text{where } \gamma \equiv \frac{1}{\sigma_0^2 + \sigma_Z^2} \end{aligned}$$

Resulting loss bound: $C(V_{\text{opt}}) = \gamma \sigma_0^2 \sigma_Z^2 d$

General approach

Can show **Bayes optimal denoiser** is:

$$f_{\text{opt}}(\tilde{x}) = \mathbb{E}[X \mid \tilde{X} = \tilde{x}]$$

For isotropic Gaussian noise,

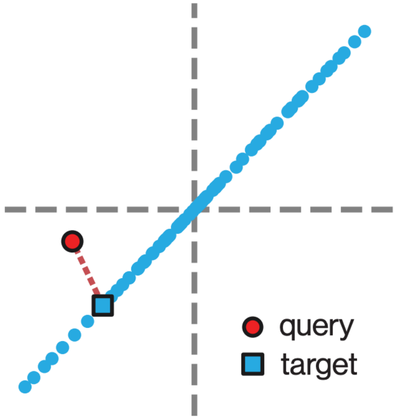
$$f_{\text{opt}}(\tilde{x}) = \frac{\int x\, e^{-\|\tilde{x}-x\|^2/2\sigma_Z^2}\, p_X(x)\, dx}{\int e^{-\|\tilde{x}-x\|^2/2\sigma_Z^2}\, p_X(x)\, dx}$$

Sub in p_X, solve $\rightarrow$ agrees!

Optimal denoisers $\leftrightarrow$ Trained attention layers

These solutions are **expressible by attention layers** $f_\theta(E)$

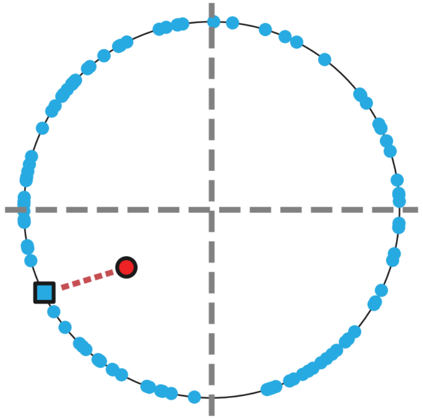
Attention input: $E = (X, \tilde{x}) = [x_1 \ldots x_L \ \tilde{x}] \in \mathbb{R}^{n\times(L+1)}$, weights $W_{KQ}, W_{PV} \in \mathbb{R}^{n\times n}$



Case 1:
Linear manifolds

dimension d, variance σ_0^2

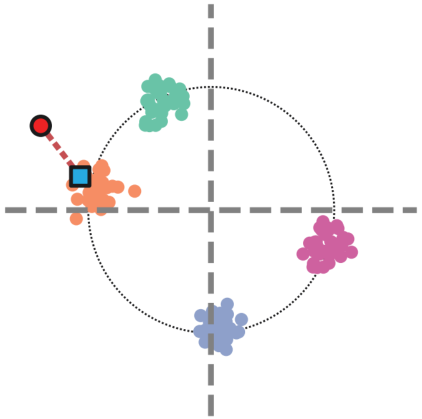
$$f_{\text{opt}}(\tilde{x}) = \gamma\, \sigma_0^2 P\, \tilde{x} = \gamma \mathbb{E}[XX^T]\, \tilde{x}$$



Case 2:
Nonlinear manifolds

d- spheres, radius R

$$f_{\text{opt}}(\tilde{x}) = \frac{\int e^{\langle x, \tilde{x}_\parallel\rangle/\sigma_Z^2}\, x\, dS_x}{\int e^{\langle x, \tilde{x}_\parallel\rangle/\sigma_Z^2}\, dS_x}$$



Case 3:
Gaussian mixtures

p components: $\mathcal{N}(\mu_\alpha, \sigma_0^2 I)$

$$f_{\text{opt}}(\tilde{x}) \approx \gamma \sigma_Z^2 \frac{\sum_\alpha e^{\gamma\langle \mu_\alpha, \tilde{x}\rangle} \mu_\alpha}{\sum_\alpha e^{\gamma\langle \mu_\alpha, \tilde{x}\rangle}}$$

* small cluster variance $\sigma_0^2 \to 0$

Recall $\gamma \equiv 1/(\sigma_0^2 + \sigma_Z^2)$
Bayes optimal: $f_{\text{opt}} = \mathbb{E}[X \mid \tilde{X}]$

We **derive optimal denoisers** for three elementary cases

Optimal denoisers ↔ Trained attention layers

These solutions are **expressible by attention layers $f_\theta(E)$**

Attention input: $E = (X, \tilde{x}) = [x_1 \dots x_L \; \tilde{x}] \in \mathbb{R}^{n\times(L+1)}$, weights $W_{KQ}, W_{PV} \in \mathbb{R}^{n\times n}$

<ins>Linear attention</ins>

$$f_{\mathrm{LSA}}(E) = W_{PV}\left(L^{-1}XX^{\mathrm{T}}\right)W_{KQ}\tilde{x}$$

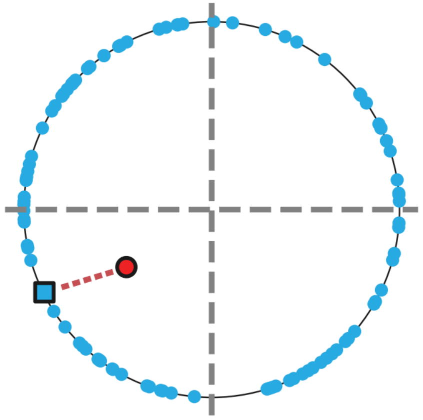
optimal for Case 1 (subspaces)

$$W^*_{PV} = \alpha I, \; W^*_{PV} = \beta I$$

with $\alpha\beta = 1/(\sigma_0^2 + \sigma_Z^2)$

↓ $L \to \infty$ *(large context limit)*

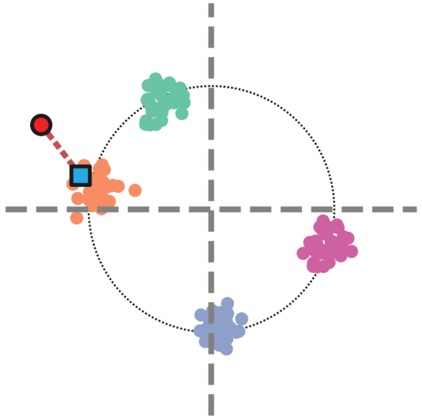
$$f_{\mathrm{opt}}(\tilde{x}) = \gamma\, \sigma_0^2 P\, \tilde{x}$$
$$= \gamma \mathbb{E}[XX^T]\, \tilde{x}$$



Case 2:
Nonlinear manifolds

d- spheres, radius R

$$f_{\mathrm{opt}}(\tilde{x}) = \frac{\int e^{\langle x, \tilde{x}_\parallel\rangle/\sigma_Z^2}\, x\; dS_x}{\int e^{\langle x, \tilde{x}_\parallel\rangle/\sigma_Z^2}\; dS_x}$$



Case 3:
Gaussian mixtures

p components: $\mathcal{N}(\mu_\alpha, \sigma_0^2 I)$

$$f_{\mathrm{opt}}(\tilde{x}) \approx \gamma\sigma_Z^2 \frac{\sum_\alpha e^{\gamma\langle\mu_\alpha, \tilde{x}\rangle}\mu_\alpha}{\sum_\alpha e^{\gamma\langle\mu_\alpha, \tilde{x}\rangle}}$$

* small cluster variance $\sigma_0^2 \to 0$

Recall $\gamma \equiv 1/(\sigma_0^2 + \sigma_Z^2)$
Bayes optimal: $f_{\mathrm{opt}} = \mathbb{E}[\,X \mid \tilde{X}\,]$

We **derive optimal denoisers** for three elementary cases

FLATIRON INSTITUTE

Optimal denoisers ↔ Trained attention layers

These solutions are **expressible by attention layers** $f_\theta(E)$

Attention input: $E = (X, \tilde{x}) = [x_1 \dots x_L \; \tilde{x}] \in \mathbb{R}^{n\times(L+1)}$, weights $W_{KQ}, W_{PV} \in \mathbb{R}^{n\times n}$

Linear attention

$$f_{\text{LSA}}(E) = W_{PV}\left(L^{-1}XX^{\mathrm{T}}\right)W_{KQ}\tilde{x}$$

optimal for Case 1 (subspaces)

$$W_{PV}^{*} = \alpha I, W_{PV}^{*} = \beta I$$

with $\alpha\beta = 1/(\sigma_0^2 + \sigma_Z^2)$

$L \to \infty$
(large context limit)

$$f_{\text{opt}}(\tilde{x}) = \gamma\, \sigma_0^2 P\, \tilde{x} = \gamma \mathbb{E}[XX^T]\, \tilde{x}$$

Softmax attention

$$f(E) = W_{PV}X \operatorname{softmax}\left(X^{\mathrm{T}}W_{KQ}\tilde{x}\right)$$

optimal for Case 2, 3 (spheres, GMM)

$$W_{PV}^{*} = \alpha I, W_{PV}^{*} = \beta I$$

with $\alpha = 1, \beta = 1/\sigma_Z^2$

$$f_{\text{opt}}(\tilde{x}) = \frac{\int e^{\langle x, \tilde{x}_{\parallel}\rangle/\sigma_Z^2}\, x\, dS_x}{\int e^{\langle x, \tilde{x}_{\parallel}\rangle/\sigma_Z^2}\, dS_x}$$

Provided $\|\mu_\alpha\| = R$

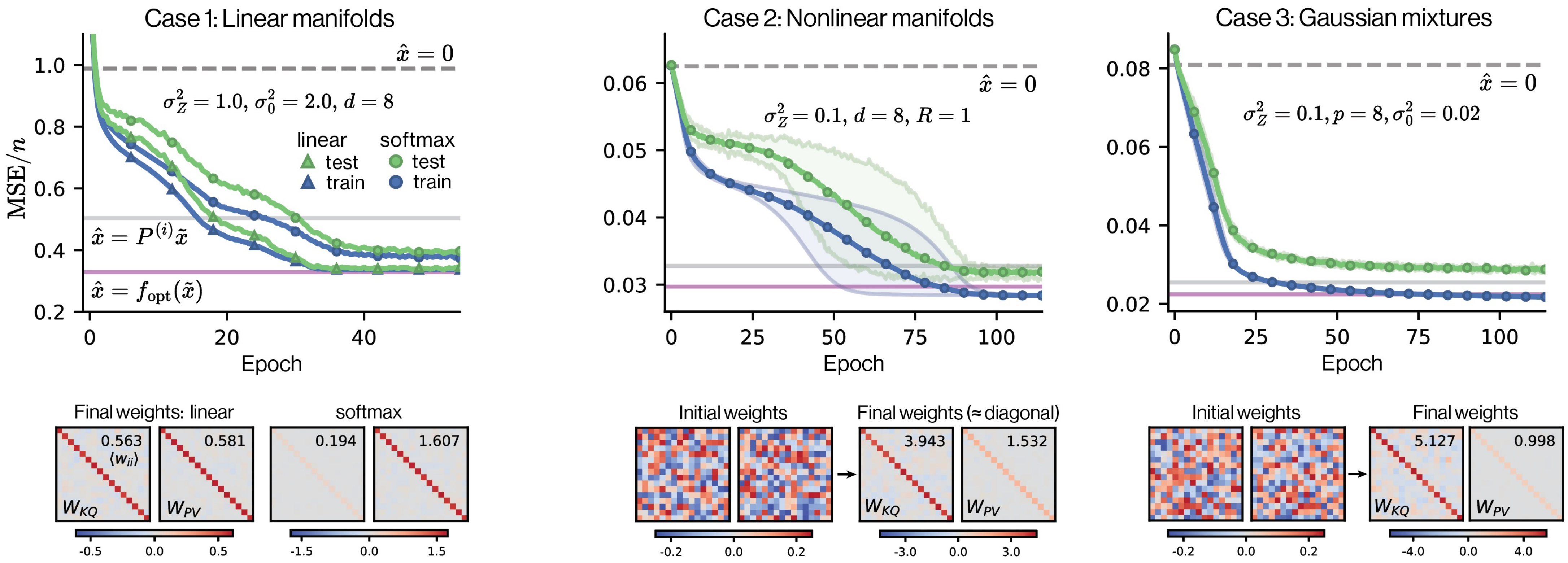
$$f_{\text{opt}}(\tilde{x}) \approx \gamma\sigma_Z^2 \frac{\sum_\alpha e^{\gamma\langle \mu_\alpha, \tilde{x}\rangle}\mu_\alpha}{\sum_\alpha e^{\gamma\langle \mu_\alpha, \tilde{x}\rangle}}$$

* small cluster variance $\sigma_0^2 \to 0$

Recall $\gamma \equiv 1/(\sigma_0^2 + \sigma_Z^2)$
Bayes optimal: $f_{\text{opt}} = \mathbb{E}[\, X \mid \tilde{X}\,]$

We **derive optimal denoisers** for three elementary cases

Experiment: Trained attention layers converge to optimal baseline



Linear attention

optimal for Case 1 (subspaces)

$$W_{PV}^* = \alpha I, W_{PV}^* = \beta I \text{ with } \alpha\beta = \frac{1}{\sigma_0^2+\sigma_Z^2}$$

Softmax attention

optimal for Case 2, 3 (spheres, GMM)

$$W_{PV}^* = \alpha I, W_{PV}^* = \beta I \text{ with } \alpha = 1, \beta = 1/\sigma_Z^2$$

Training: Adam, $N = 800$ samples, random initial weights
Plots show means + min/max envelope over 6 random seeds

FLATIRON INSTITUTE

In-context denoising bridges attention and associative memory

"Spherical" Hopfield model

$$\mathcal{E}_{\text{LSA}}(X,s) = -s^T J s + \frac{1}{2\alpha}\|s\|^2$$

$$J = \frac{\beta}{L} X X^{\text{T}}$$

Dense associative memory (MCHN)

$$\mathcal{E}(X,s) = -\frac{1}{\beta}\log\left(\sum_{t=1}^{L} e^{\beta x_t^T s}\right) + \frac{1}{2\alpha}\|s\|^2$$

$$s(t+1) = s(t) - \gamma\, \nabla_s \mathcal{E}\big(X, s(t)\big)$$
$$= \text{Attn}(X, s(t))$$

$$f^*_{\text{LSA}}(X,s) = \alpha\beta L^{-1}\, X X^{\text{T}} s$$

(Trained) **Linear attention**

$$f^*(X,s) = \alpha X\ \text{softmax}\big(\beta X^{\text{T}} s\big)$$

(Trained) **Softmax attention**

Remarks

1. For Cases 1-3, attention layers trained have scaled identity weights
2. Gradient descent on context-dependent AM landscape is mappable to attention (step size $\gamma = \alpha$ set by the Lagrange multiplier of $\|s\|^2$)

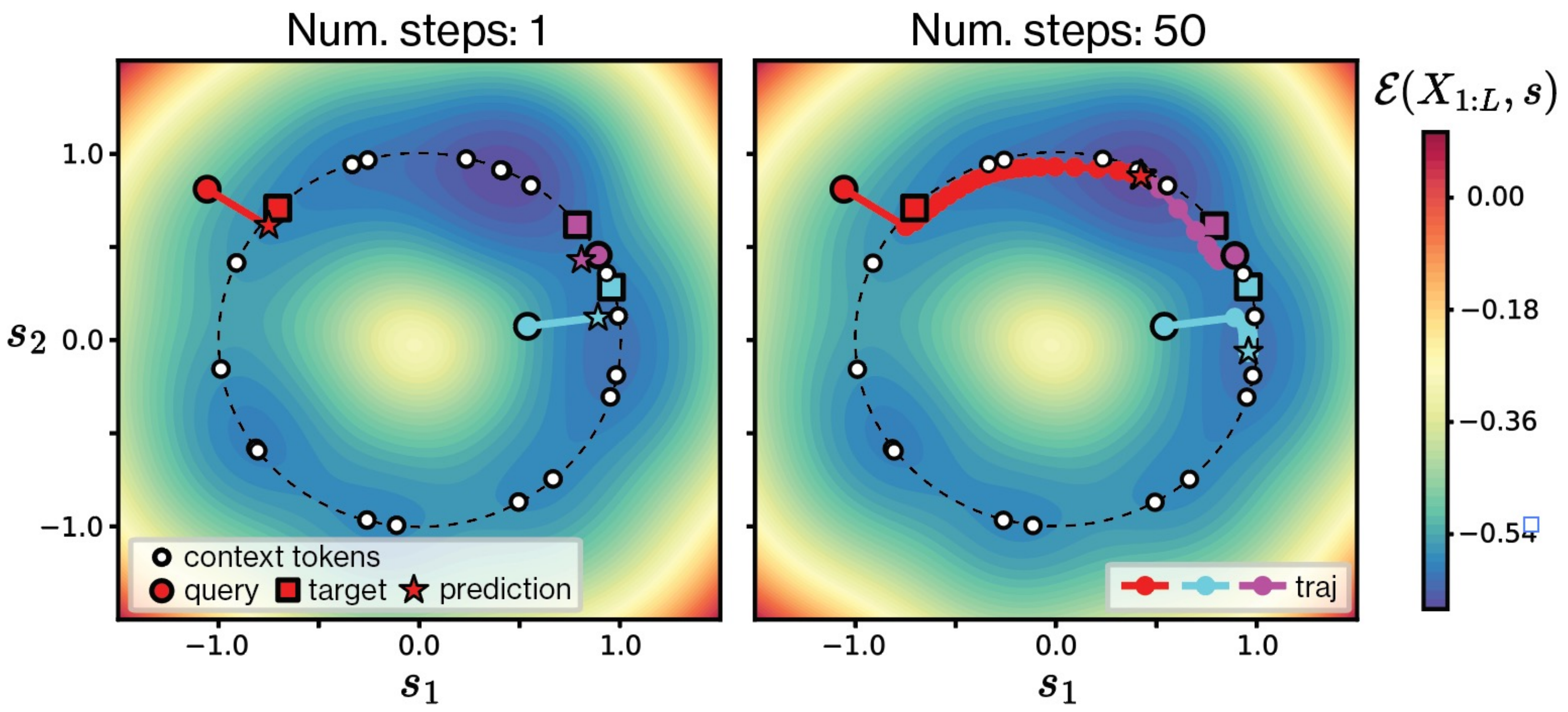
Context tokens → associative memory patterns (landscape)

Query → corrupted initial state

FLATIRON INSTITUTE

Is one gradient step really better than many?

- Sample $L = 30$ context tokens for sphere case in dimension $n = 2$
- Energy landscape is a context-dependent DAM (trained softmax attention)
- Denoising trajectories shown for three initial queries



'One step' vs. recurrent denoising

- One-step optimally blends query information with contextual guidance
- Exact retrieval (iteration) is sub-optimal: loses detailed information about the query

FLATIRON INSTITUTE

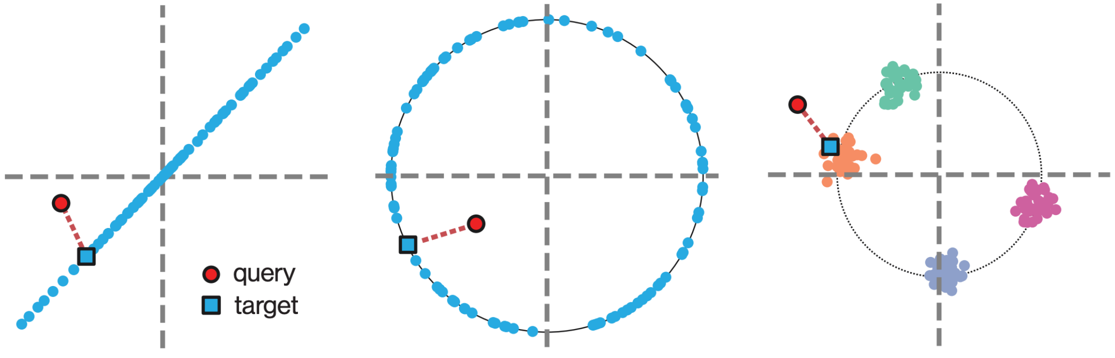
Summary & Outlook

Poster
East Building
#E-3207

Establishes a bridge between three communities:
In-context learning, Attention mechanisms, & Associative memory networks

Introduced "In-context denoising"

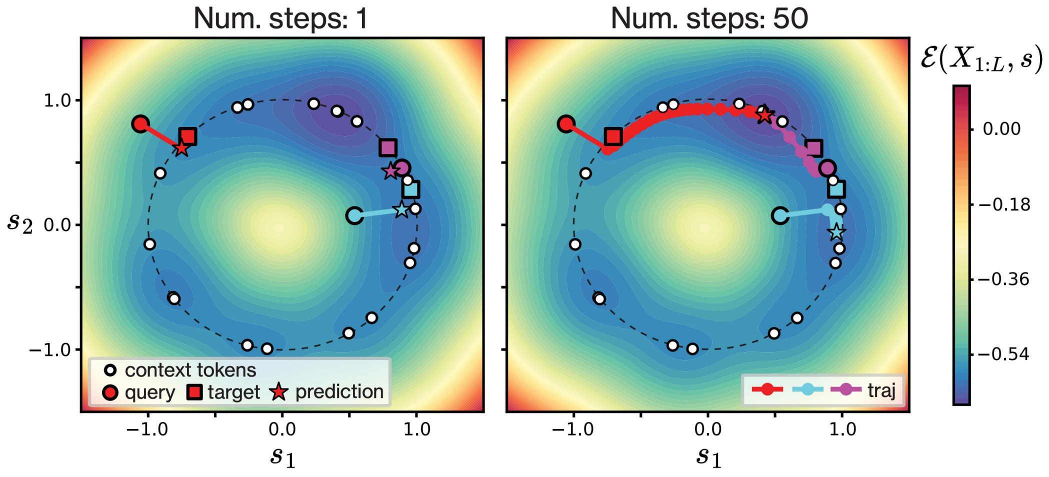
- Single attention step can express **Bayes optimal solutions** on certain restricted problems
- Trained attention layers converge to scaled identity weights



Case 1: Linear manifolds | Case 2: Nonlinear manifolds | Case 3: Gaussian mixtures

Refined "Attention ↔ Associative memory networks"

- Trained attention layers perform gradient descent on **context-dependent associative memory landscape**
- Recurrent iteration (**exact retrieval**) is **suboptimal**



Next steps and ongoing work

- Non-isotropic noise (non-trivial W_{KQ}, W_{PV})
- Positional embedding; dynamical inference
- Multi-layer / multi-head attention
- Connection to diffusion models

FLATIRON INSTITUTE