Dense Associative Memory for Gaussian Distributions

Based on Tankala and Balasubramanian, 2026

- Classical associative memories retrieve a stored vector from a corrupted query.
- Modern representation learning often stores uncertainty as a distribution.
- Can we build an associative memory whose states are probability distributions.

$$\text{vector recall:} \quad \xi \in \mathbb{R}^d \longmapsto x_i,$$

$$\text{distributional recall:} \quad \xi \in \mathcal{P}_2(\mathbb{R}^d) \longmapsto X_i,$$

$$X_i = \mathcal{N}(\mu_i, \Sigma_i).$$

Can dense associative memory operate directly on Gaussian distributions?

Motivation: Gaussian embeddings are natural memory units

- A Gaussian embedding stores both location and uncertainty.
- Means encode central semantic content; covariances encode spread, ambiguity, and density.
- Examples include word, document, sentence, graph, and image embeddings.

Object	Gaussian representation
word	$\mathcal{N}(\mu_{\text{word}}, \Sigma_{\text{word}})$
document	$\mathcal{N}(\mu_{\text{doc}}, \Sigma_{\text{doc}})$
image	$\mathcal{N}(\mu_{\text{latent}}, \Sigma_{\text{latent}})$
sentence	$\mathcal{N}(\mu_{\text{sent}}, \Sigma_{\text{sent}})$

$$\Sigma_i \succ 0, \qquad X_i \in \mathcal{P}_2(\mathbb{R}^d).$$

Motivation: vector DAMs do not see covariance geometry

- Modern Dense AM models use an energy on vectors.

$$E_{\mathrm{vec}}(\xi) = -\frac{1}{\beta} \log \sum_{i=1}^{N} \exp\left(-\beta \|x_i - \xi\|^2\right),$$

- The log-sum-exp nonlinearity concentrates mass on the nearest stored pattern.

$$\xi_{\mathrm{new}} = \sum_{i=1}^{N} w_i(\xi) x_i.$$

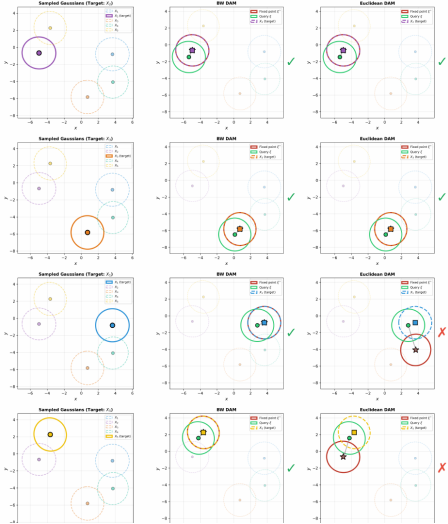
- For Gaussians, treating (μ, Σ) as a flat vector ignores the natural metric on covariance matrices.

A Gaussian memory needs a geometry-aware distance.

Motivation 4: geometry changes retrieval

- Eu-DAM stacks μ and Σ into one Euclidean vector.
- BW-DAM uses optimal transport maps between Gaussians.
- Two-dimensional example shows BW-DAM retrieving all targets, while Eu-DAM fails in two rows.

$$\|\Sigma_1 - \Sigma_2\|_F \not\equiv d_B(\Sigma_1, \Sigma_2).$$



Our contributions

- A log-sum-exp energy defined on Wasserstein space.
- A retrieval operator that aggregates optimal transport maps with Gibbs weights.
- Fixed points interpreted as self-consistent Wasserstein barycenters.
- Exponential storage capacity for randomly sampled Gaussian memories.
- Geometric convergence and exponential-in-dimension retrieval error bounds.
- Experiments on synthetic Gaussians, CelebA, CIFAR-10, text8, and NLI sentences.

Core idea: replace Euclidean averaging by Bures-Wasserstein transport aggregation.

Methodology 1: Gaussian Wasserstein geometry

For $X_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $X_2 = \mathcal{N}(\mu_2, \Sigma_2)$,

$$W_2^2(X_1, X_2) = \|\mu_1 - \mu_2\|^2 + \operatorname{tr}\Big(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Big).$$

The second term is the squared Bures distance between covariances.

If $\xi = \mathcal{N}(m, \Omega)$ and $X_i = \mathcal{N}(\mu_i, \Sigma_i)$, the optimal map is affine:

$$T_i(x) = \mu_i + A_i(x - m),$$
$$A_i = \Sigma_i^{1/2} M_i^{-1/2} \Sigma_i^{1/2},$$
$$M_i = \Sigma_i^{1/2} \Omega \Sigma_i^{1/2}.$$

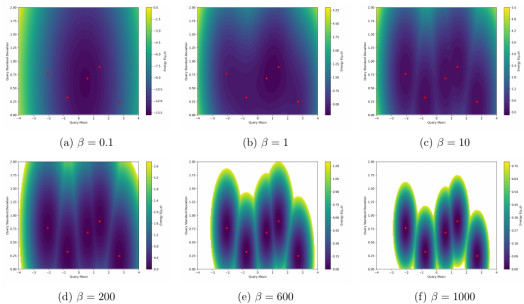
Closed-form distances and maps make BW-DAM computationally explicit.

Methodology 2: Wasserstein log-sum-exp energy

The memory stores distributions $X_1, \ldots, X_N \in \mathcal{P}_2(\mathbb{R}^d)$ and scores a query ξ by

$$E(\xi) = -\frac{1}{\beta} \log \left(\sum_{i=1}^{N} \exp\{-\beta W_2^2(X_i, \xi)\} \right).$$

- small β: diffuse, overlapping basins;
- large β: sharp soft-min behavior;
- memories become attractors in Bures-Wasserstein geometry.



(a) $\beta = 0.1$ (b) $\beta = 1$ (c) $\beta = 10$

(d) $\beta = 200$ (e) $\beta = 600$ (f) $\beta = 1000$

Methodology 3: Gibbs weights generate competitive recall

The energy induces weights

$$w_i(\xi) = \frac{\exp\{-\beta W_2^2(X_i, \xi)\}}{\sum_{j=1}^{N} \exp\{-\beta W_2^2(X_j, \xi)\}}.$$

When X_i is much closer than all other memories,

$$w_i(\xi) \approx 1, \qquad w_j(\xi) \approx 0 \quad (j \neq i).$$

The Wasserstein gradient is a weighted transport field:

$$\nabla_W E(\xi)(x) = 2 \sum_{i=1}^{N} w_i(\xi) \Big(\mathrm{Id} - T_i \Big)(x).$$

Thus retrieval is not a weighted average of vectors; it is a weighted average of maps from the query to the memories.

Temperature β controls how strongly the nearest memory dominates.

Methodology 4: fixed points are self-consistent barycenters

At a stationary point ξ^*,

$$\sum_{i=1}^{N} w_i(\xi^*)\Big(\mathrm{Id} - T_i\Big) = 0$$

$$\iff \sum_{i=1}^{N} w_i(\xi^*)\, T_i = \mathrm{Id}.$$

Define the update operator

$$S_\xi = \sum_{i=1}^{N} w_i(\xi)\, T_i,$$

$$\Phi(\xi) = (S_\xi)_{\#}\, \xi.$$

Retrieval solves

$$\Phi(\xi^*) = \xi^*.$$

The fixed point is a Wasserstein barycentric object whose weights depend on the object itself.

$$\xi \xrightarrow{W_2^2(X_i,\xi)} \{w_i(\xi)\} \xrightarrow{\{T_i\}} \Phi(\xi).$$

This is the distributional analogue of classical DAM retrieval.

Methodology 5: one BW-DAM update for Gaussians

For a query $\xi = \mathcal{N}(m, \Omega)$ and stored memories $X_i = \mathcal{N}(\mu_i, \Sigma_i)$:

Step 1: distances and weights

$$D_i = \|\mu_i - m\|^2 + \mathrm{tr}\Big(\Sigma_i + \Omega - 2(\Sigma_i^{1/2}\Omega\Sigma_i^{1/2})^{1/2}\Big),$$

$$w_i = \frac{e^{-\beta D_i}}{\sum_{j=1}^{N} e^{-\beta D_j}}.$$

Step 2: transport and update

$$A_i = \Sigma_i^{1/2}(\Sigma_i^{1/2}\Omega\Sigma_i^{1/2})^{-1/2}\Sigma_i^{1/2},$$

$$\widetilde{A} = \sum_{i=1}^{N} w_i A_i,$$

$$m' = \sum_{i=1}^{N} w_i \mu_i, \qquad \Omega' = \widetilde{A}\Omega\widetilde{A}^{\top}.$$

Return $\Phi(\xi) = \mathcal{N}(m', \Omega')$ and iterate until convergence.

Theory 1: storage means a unique local attractor

For each stored Gaussian X_i, define a Wasserstein ball

$$B_i = \{\nu \in \mathcal{P}_2(\mathbb{R}^d) : W_2(X_i, \nu) < r\}.$$

We call X_i stored if:

- B_i contains a unique fixed point X_i^*;
- all queries in B_i converge to X_i^*;
- the basins B_i are pairwise disjoint.

With eigenvalues in $[\lambda_{\min}, \lambda_{\max}]$, the key separation assumption is

$$\min_{i \neq j} \Big(-\log \langle X_i, X_j \rangle_{L^2} \Big) \geq \frac{d}{2} \log(4\pi\lambda_{\max}) + d.$$

The proof uses the basin radius

$$r = \sqrt{\lambda_{\min}}.$$

This radius does not shrink with the number of stored distributions.

Theory 2: exponential storage capacity

Theorem 1, compressed. Let $N = \lfloor \sqrt{p}\, e^{\alpha d} \rfloor$ Gaussian measures be sampled on a Wasserstein sphere, where

$$\alpha = \frac{(1 + \log \kappa)^2}{4(3 + 2\log \kappa)^2}, \qquad \kappa = \frac{\lambda_{\max}}{\lambda_{\min}}.$$

For large enough d and $\beta \geq \max\{1, \beta_0\}$, all N memories are stored with probability at least $1 - p$.

$$N = \Theta(e^{\alpha d}).$$

- Capacity is exponential in dimension.
- The theorem extends vector DAM capacity to Gaussian distributions.
- The storage basin radius is $\sqrt{\lambda_{\min}}$, independent of N.

Theory 3: retrieval converges geometrically and errors decay

Theorem 2: convergence. If $\xi^{(0)} \in B_i$ and $\xi^{(k+1)} = \Phi(\xi^{(k)})$, then

$$\xi^{(k)} \in B_i,$$
$$W_2(\xi^{(k)}, X_i^*) \le L^k W_2(\xi^{(0)}, X_i^*) \le 2\sqrt{\lambda_{\min}}\, L^k,$$

where $L < 1$. To reach accuracy ε,

$$k \ge \frac{\log(2\sqrt{\lambda_{\min}}/\varepsilon)}{\log(1/L)}.$$

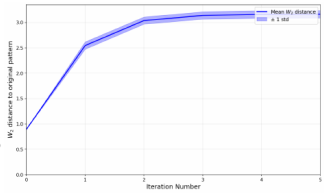
Theorem 3: one-step error. For random memories on the Wasserstein sphere and sufficiently large β,

$$W_2(X_i, \Phi(\xi^{(0)})) \le C\sqrt{d}\, e^{-\gamma d},$$
$$\gamma = \beta\lambda_{\min} - \frac{\alpha}{2}.$$

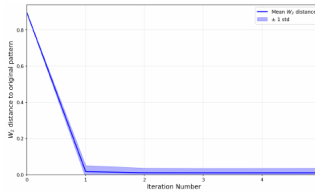
- Iteration contracts inside the correct basin.
- Finite-β fixed points are close to the stored Gaussian.
- Larger β increases weight concentration and improves retrieval.

Experiments 1: synthetic dynamics show the role of temperature

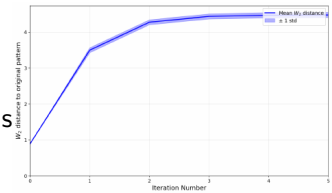
- Store $N = 1000$ Gaussian measures on a Wasserstein sphere.
- Perturb 75% of memories by distance $W_2 = \sqrt{\lambda_{\min}}$.
- For $\beta = 2$, retrieval reaches the original pattern in one iteration.
- For $\beta = 0.1$, retrieval fails and moves toward the wrong distribution.



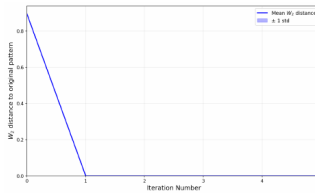
(a) $d = 10$, $\beta = 0.1$



(b) $d = 10$, $\beta = 2$

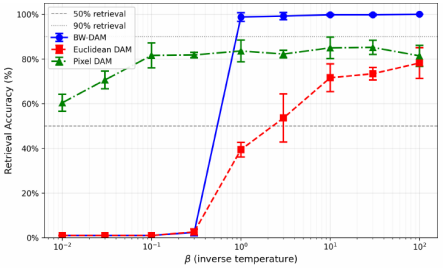


(c) $d = 20$, $\beta = 0.1$

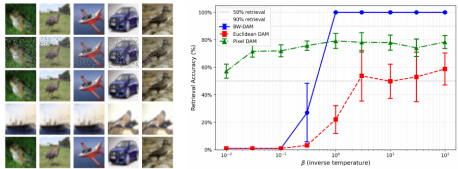


(d) $d = 20$, $\beta = 2$

Experiments 2: image retrieval on CelebA and CIFAR-10

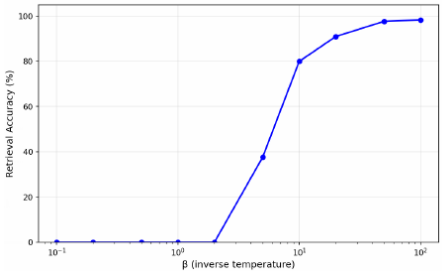


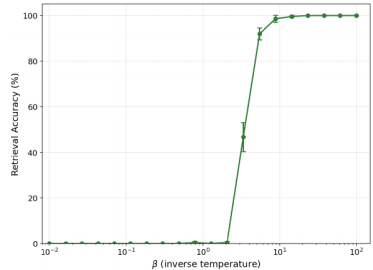
- CelebA queries mask 20% of pixels before Gaussian latent retrieval.
- BW-DAM reaches near-perfect retrieval at moderate β.



- CIFAR-10 shows the same phase transition in retrieval accuracy.
- Eu-DAM and Pixel-DAM plateau below BW-DAM because they ignore BW geometry.

Experiments 3: text retrieval and final takeaway





- Word2Gauss on text8: 11,815 word Gaussians in $\mathbb{R}^{50}$.
- Retrieval accuracy approaches 100% as β increases.
- GaussCSE on NLI: 1000 sentence Gaussians in $\mathbb{R}^{768}$.
- Sharp phase transition confirms the temperature-based basin mechanism.

Takeaway: respecting Wasserstein geometry gives robust distributional recall.