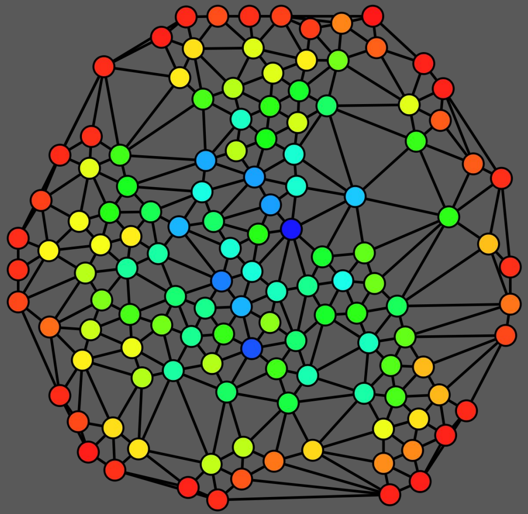
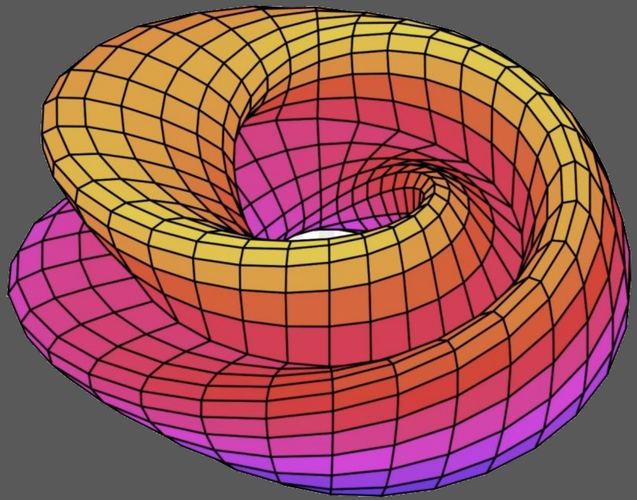
Stochastic Flows and Geometric Optimization on the Orthogonal Group

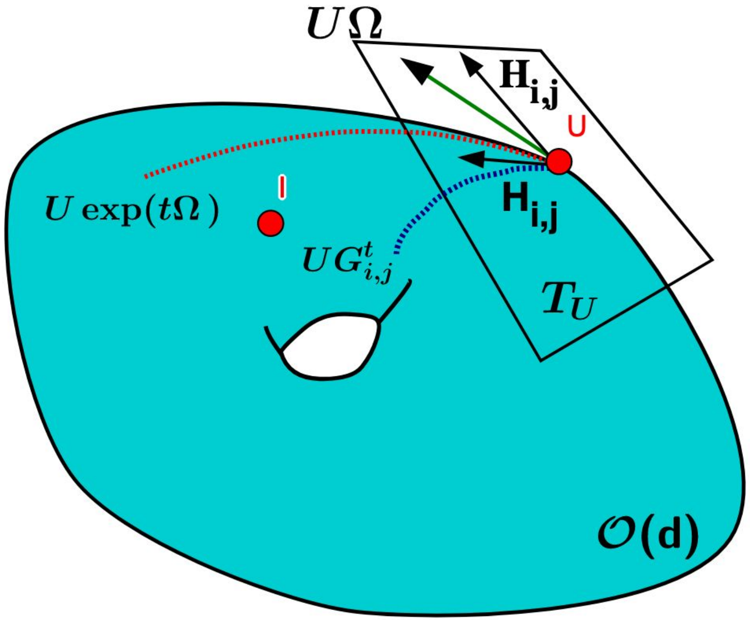




Krzysztof Choromanski, David Cheikhi, Jared Davis, Valerii Likhosherstov, Achille Nazaret, Achraf Bahamou, Xingyou Song, Mrugank Akarte, Jack Parker-Holder, Jacob Bergquist, Yuan Gao, Aldo Pacchiano, Tamas Sarlos, Adrian Weller, Vikas Sindhwani

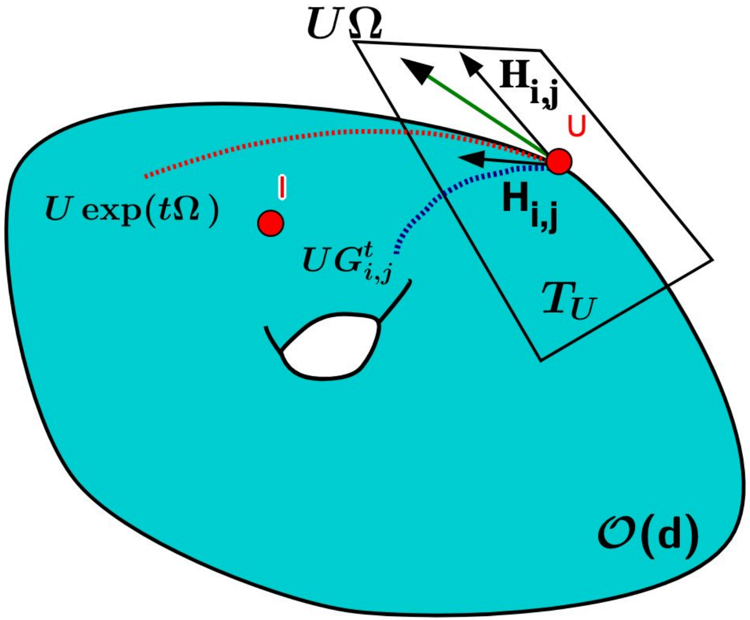
[https://arxiv.org/abs/2003.13563]

Problem Definition



Problem Definition

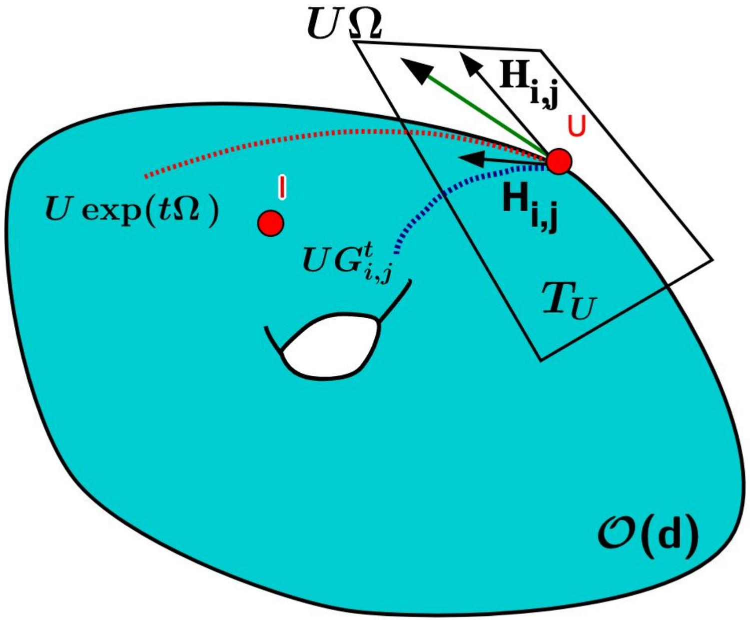
$$\max_{\mathbf{X}\in\mathbb{R}^{d\times k},\mathbf{X}^\top\mathbf{X}=\mathbf{I}_k} F(\mathbf{X})$$



Problem Definition

$$\max_{\mathbf{X}\in\mathbb{R}^{d\times k},\mathbf{X}^\top\mathbf{X}=\mathbf{I}_k} F(\mathbf{X})$$

gradient flow approach

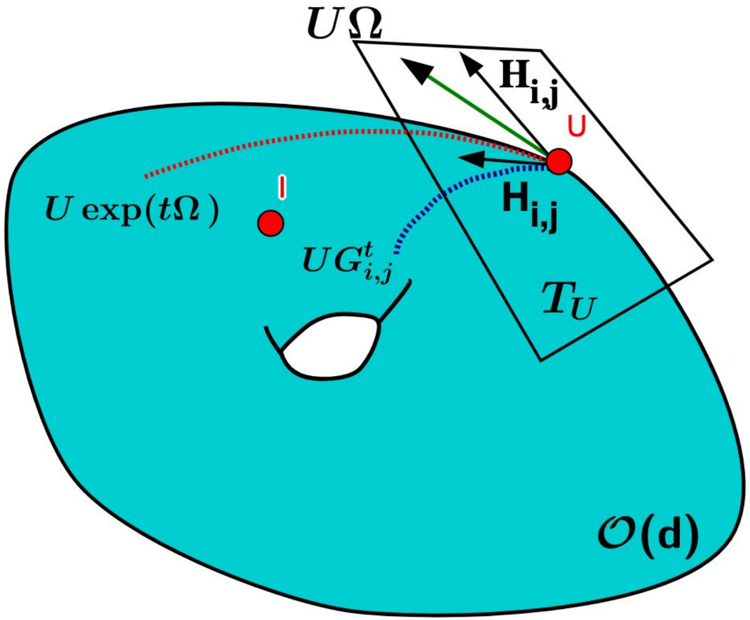


Problem Definition

$$\max_{\mathbf{X}\in\mathbb{R}^{d\times k},\mathbf{X}^\top\mathbf{X}=\mathbf{I}_k} F(\mathbf{X})$$

gradient flow approach

$$\begin{cases}\dot{\mathbf{X}}(t) = \Omega(t)\mathbf{X}(t) \\ \Omega = \Omega(\mathbf{X},\mathbf{G}) = \mathbf{G}\mathbf{X}^\top - \mathbf{X}\mathbf{G}^\top\end{cases}$$



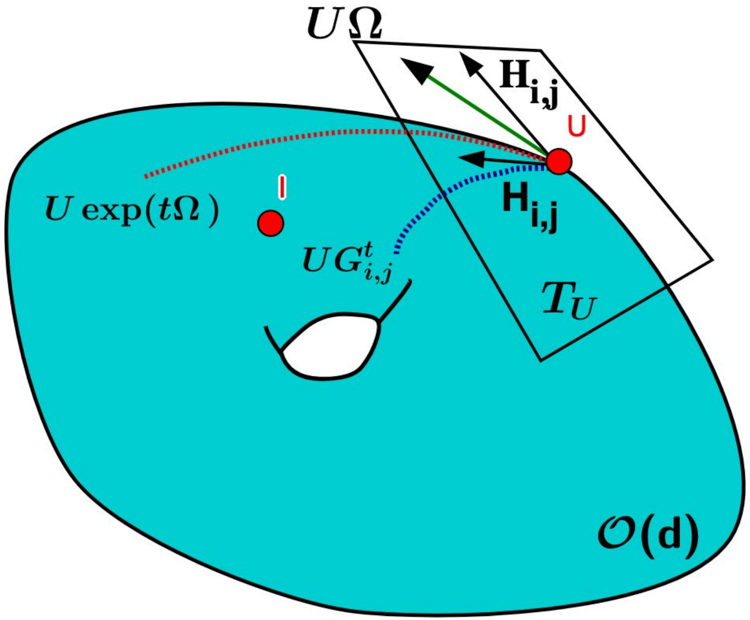
Problem Definition

$$\max_{\mathbf{X}\in\mathbb{R}^{d\times k},\mathbf{X}^\top\mathbf{X}=\mathbf{I}_k} F(\mathbf{X})$$

gradient flow approach

$$\begin{cases}\dot{\mathbf{X}}(t) = \Omega(t)\mathbf{X}(t) \\ \Omega = \Omega(\mathbf{X},\mathbf{G}) = \mathbf{G}\mathbf{X}^\top - \mathbf{X}\mathbf{G}^\top\end{cases}$$

discretization



Problem Definition

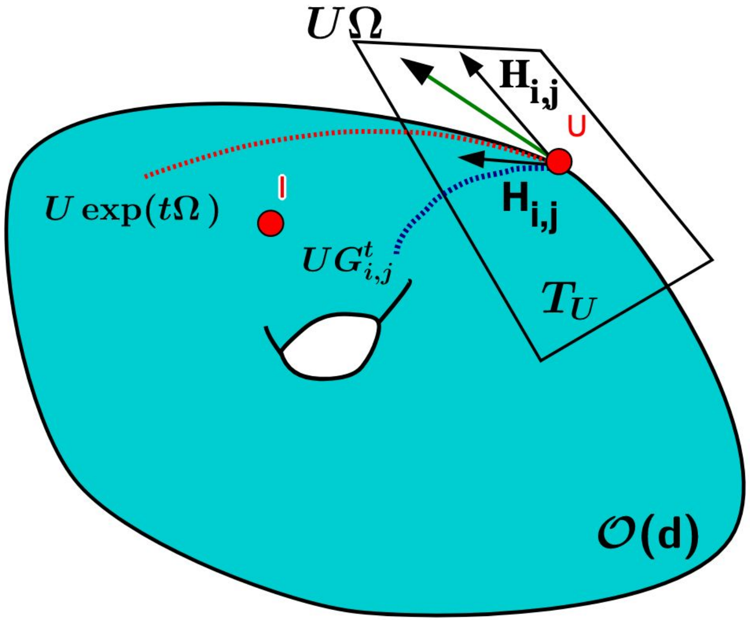
$$\max_{\mathbf{X}\in\mathbb{R}^{d\times k},\mathbf{X}^\top\mathbf{X}=\mathbf{I}_k} F(\mathbf{X})$$

gradient flow approach

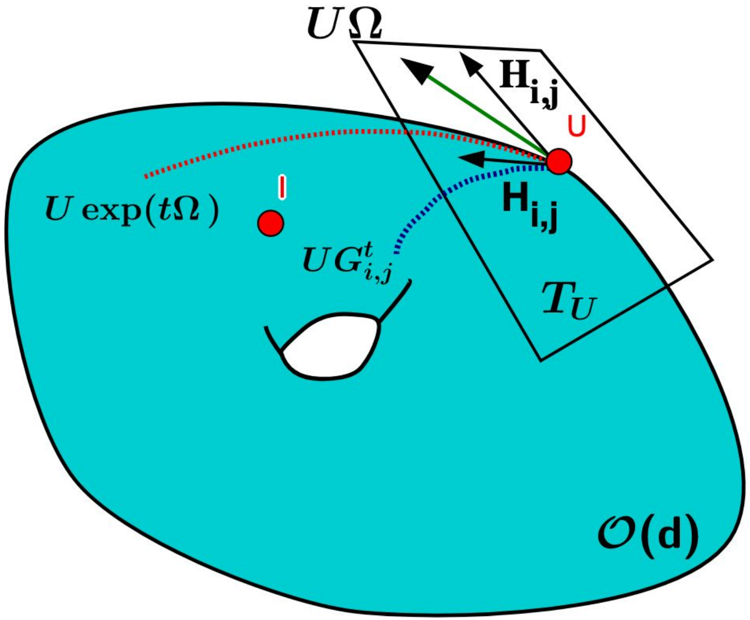
$$\begin{cases}\dot{\mathbf{X}}(t) = \Omega(t)\mathbf{X}(t) \\ \Omega = \Omega(\mathbf{X},\mathbf{G}) = \mathbf{G}\mathbf{X}^\top - \mathbf{X}\mathbf{G}^\top\end{cases}$$

discretization

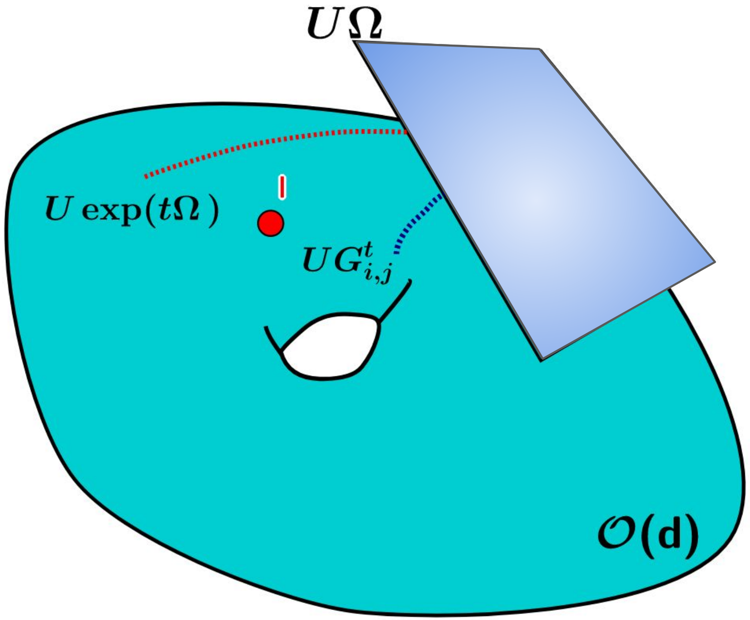
$$\mathbf{X}_{i+1} = \Gamma(\eta\Omega(\mathbf{X}_i,\mathbf{G}_i))\mathbf{X}_i$$



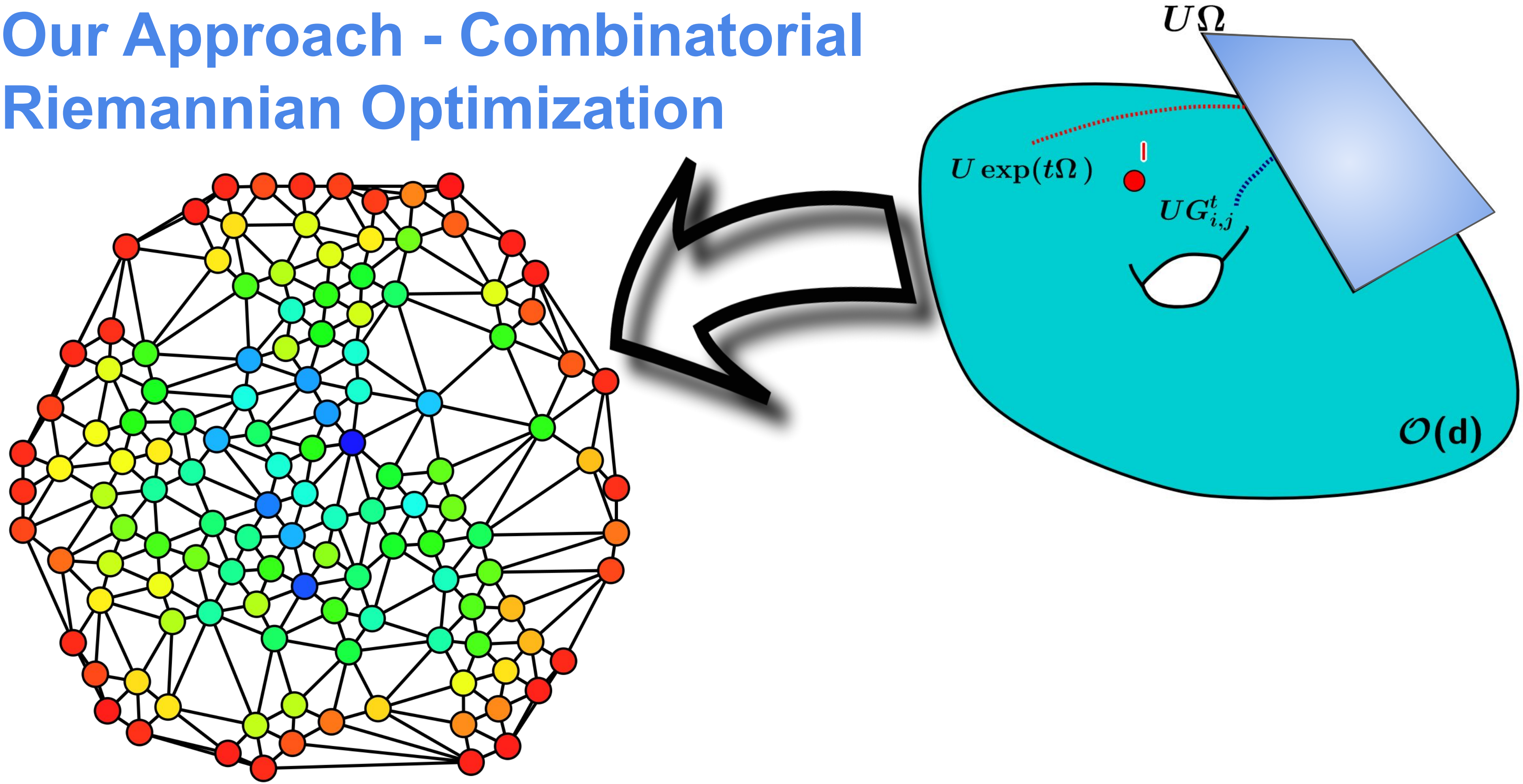
Our Approach - Combinatorial Riemannian Optimization



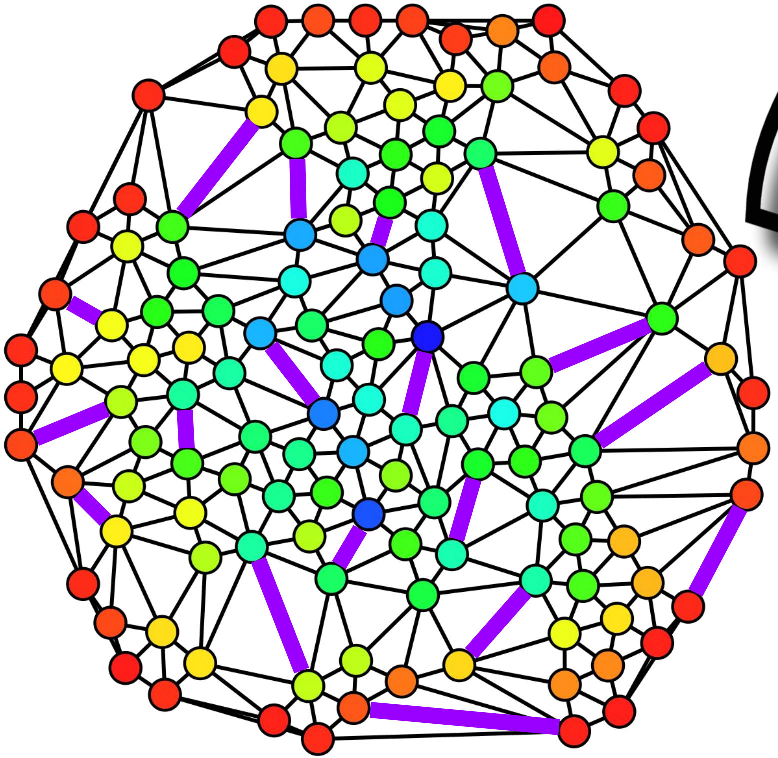
Our Approach - Combinatorial Riemannian Optimization

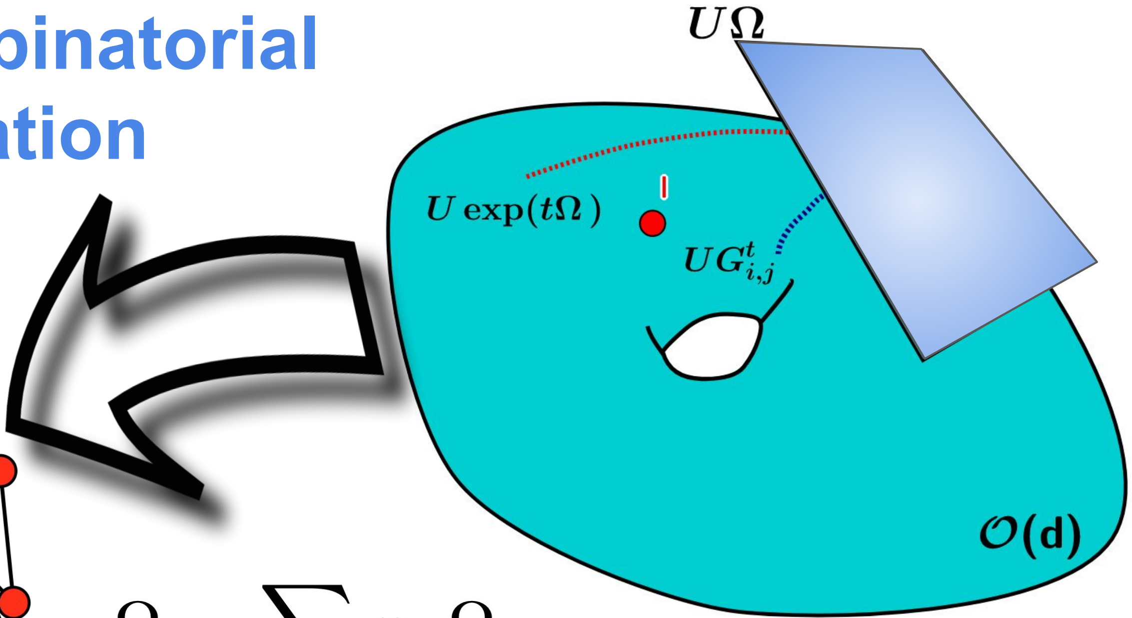


Our Approach - Combinatorial Riemannian Optimization



Our Approach - Combinatorial Riemannian Optimization





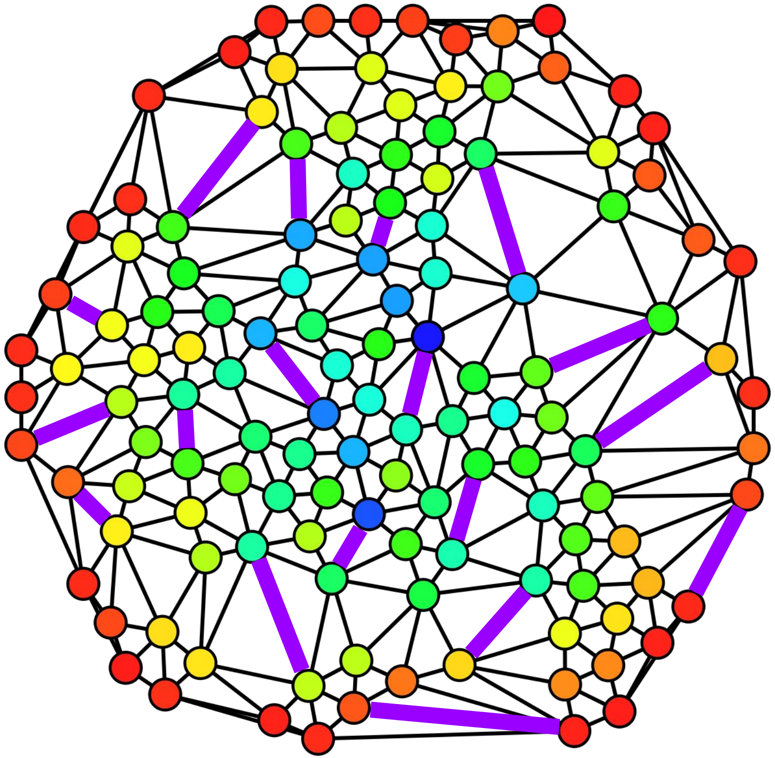
$$\Omega = \sum_{T \in \mathcal{T}} p_T \Omega_T$$

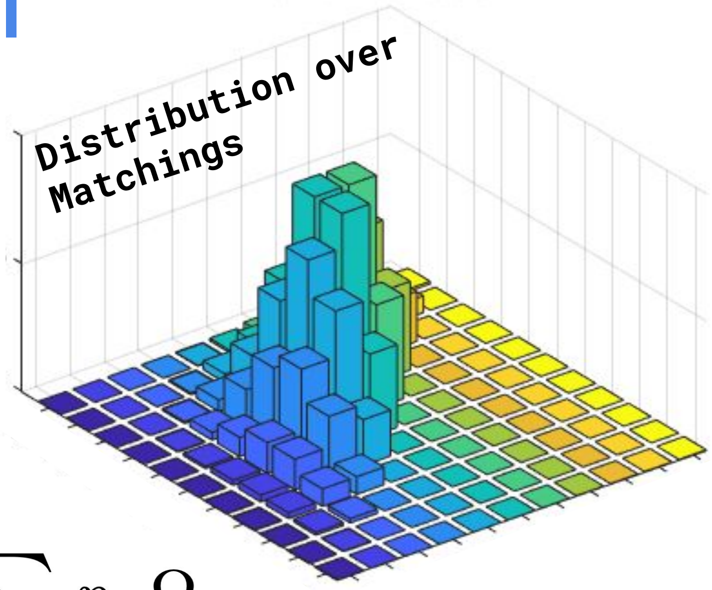
$$\mathbf{X}_{i+1} = \prod_{k=1,\ldots,l} \mathbf{G}^{\theta_k}_{i_k, j_k} \mathbf{X}_i$$

Learned Givens rotations

matching [k-clustering for k=1]

Our Approach - Combinatorial Riemannian Optimization





$$\Omega = \sum_{T \in \mathcal{T}} p_T \Omega_T$$

$$\mathbf{X}_{i+1} = \prod_{k=1,\ldots,l} \mathbf{G}^{\theta_k}_{i_k, j_k} \mathbf{X}_i$$

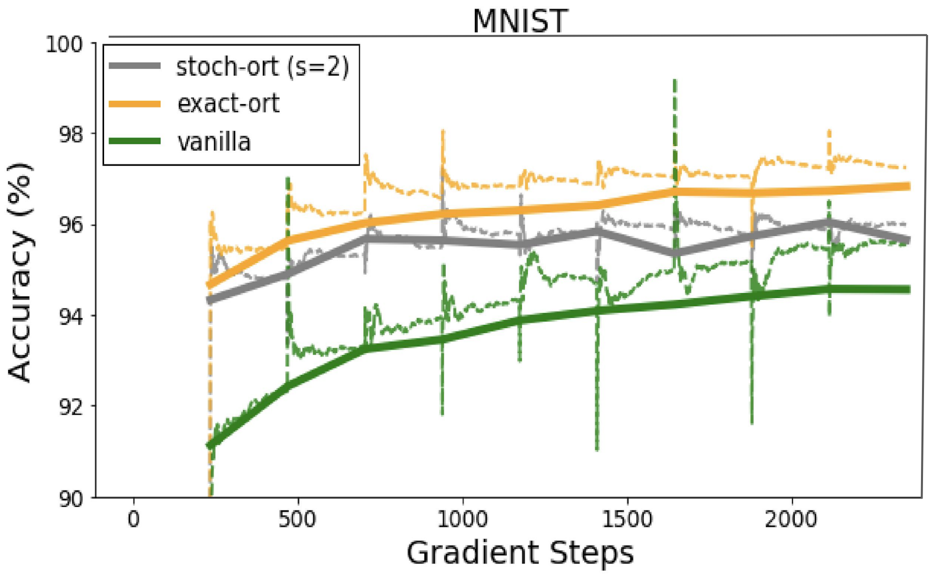
Learned Givens rotations

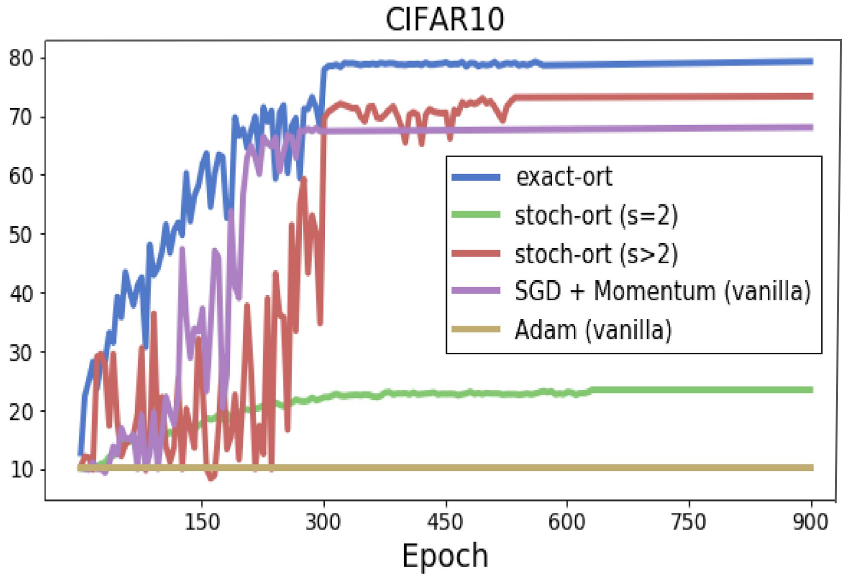
matching [k-clustering for k=1]

Our Algorithm in Action

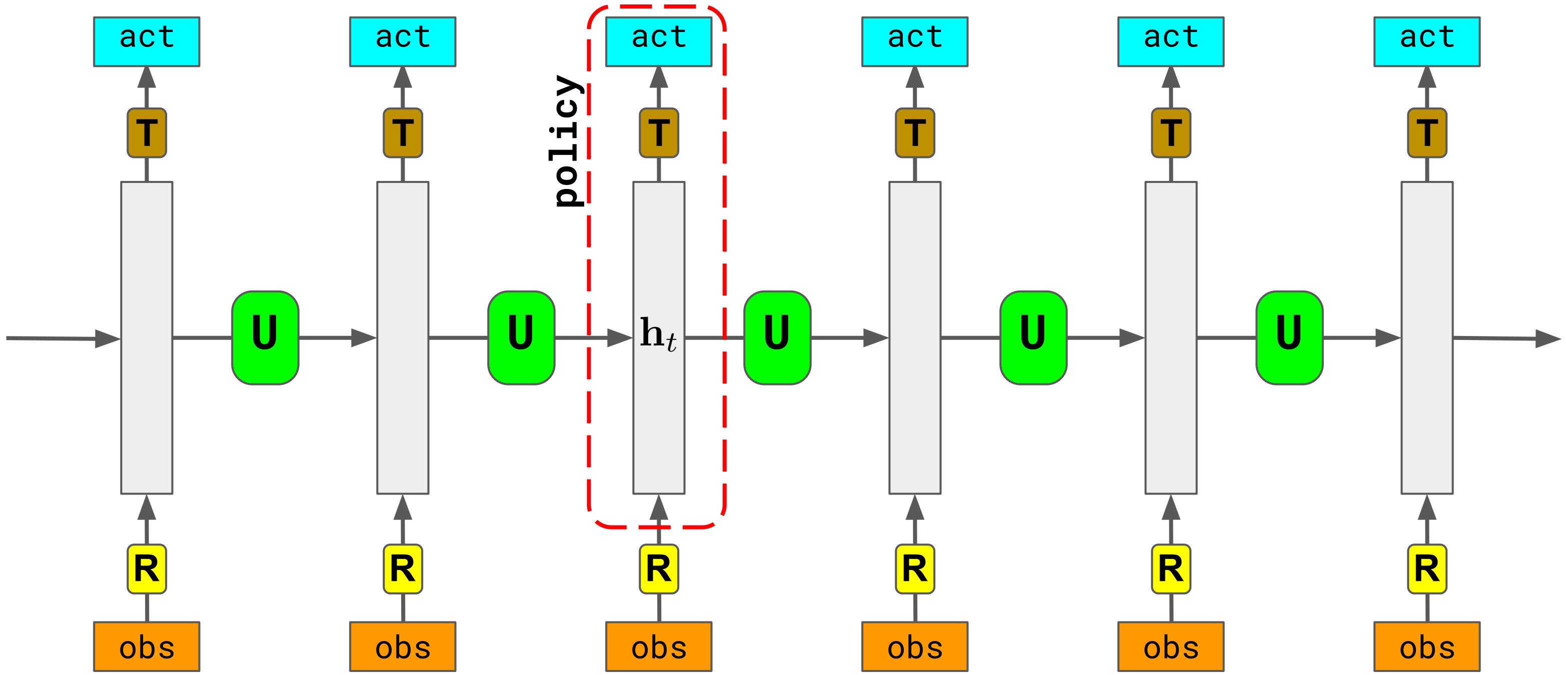
Average number of FLOPS in 10K

	ES-200	ES-400	MNIST	CIFAR
exact-ort	> 1600	> 12800	> 200	> 49K
stoch-ort ($s = 2$)	$<$ **68**	$<$ **272**	$<$ **8.5**	< 2K
ortinit-vanRNN ($p = 20$)	> 84	> 656	N/A	N/A
ortinit-vanRNN ($p = 10$)	> 164	> 1296	N/A	N/A
ortinit-vanRNN ($p = 8$)	> 204	> 1616	N/A	N/A
ortinit-vanRNN ($p = 5$)	> 324	> 2576	N/A	N/A
ortinit-vanRNN ($p = 4$)	> 404	> 3216	N/A	N/A
stoch-ort ($s = r^*$)	N/A	N/A	< 43	$<$ **14K**





Learning Recurrent Reinforcement Learning Policies

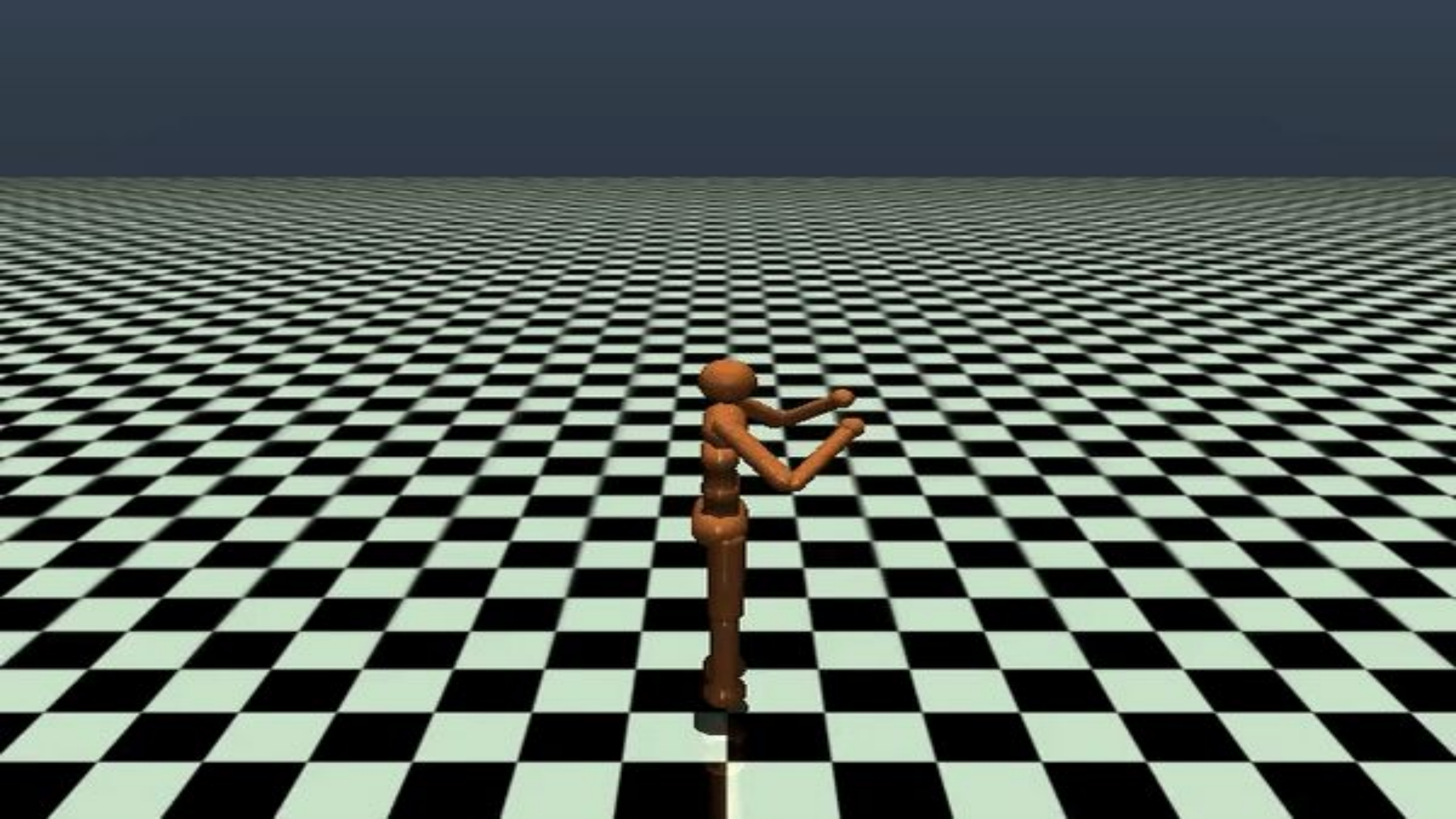


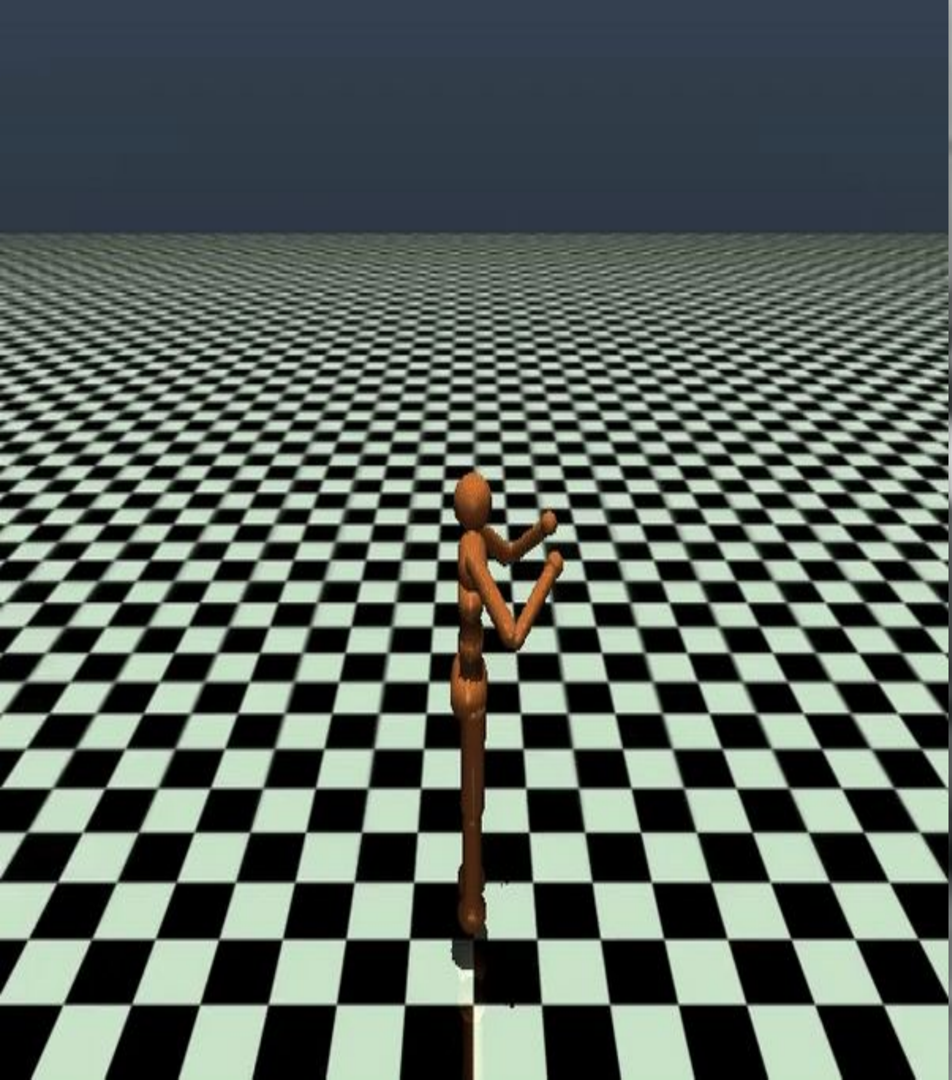
Tested methods:

- affine policies
- vanilla RNNs
- vanilla RNNs with orthogonal initialization
- ortho-RNNs - exact Riemannian optimization
- ortho-RNNs - stochastic optimization

Affine doesn't fly...

well...even doesn't really walk...



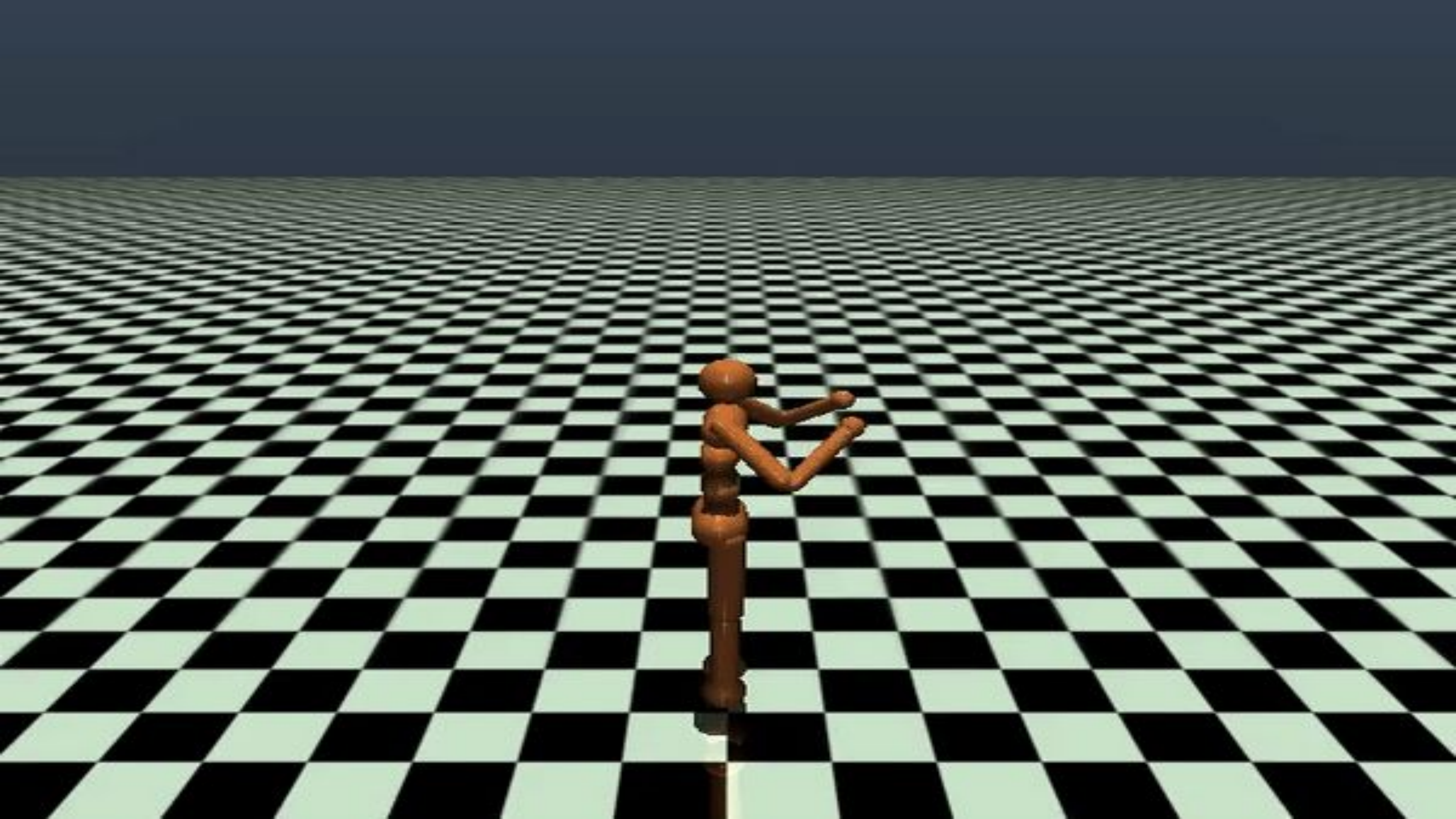



Vanilla RNN

aka in the search for some spin-dance...

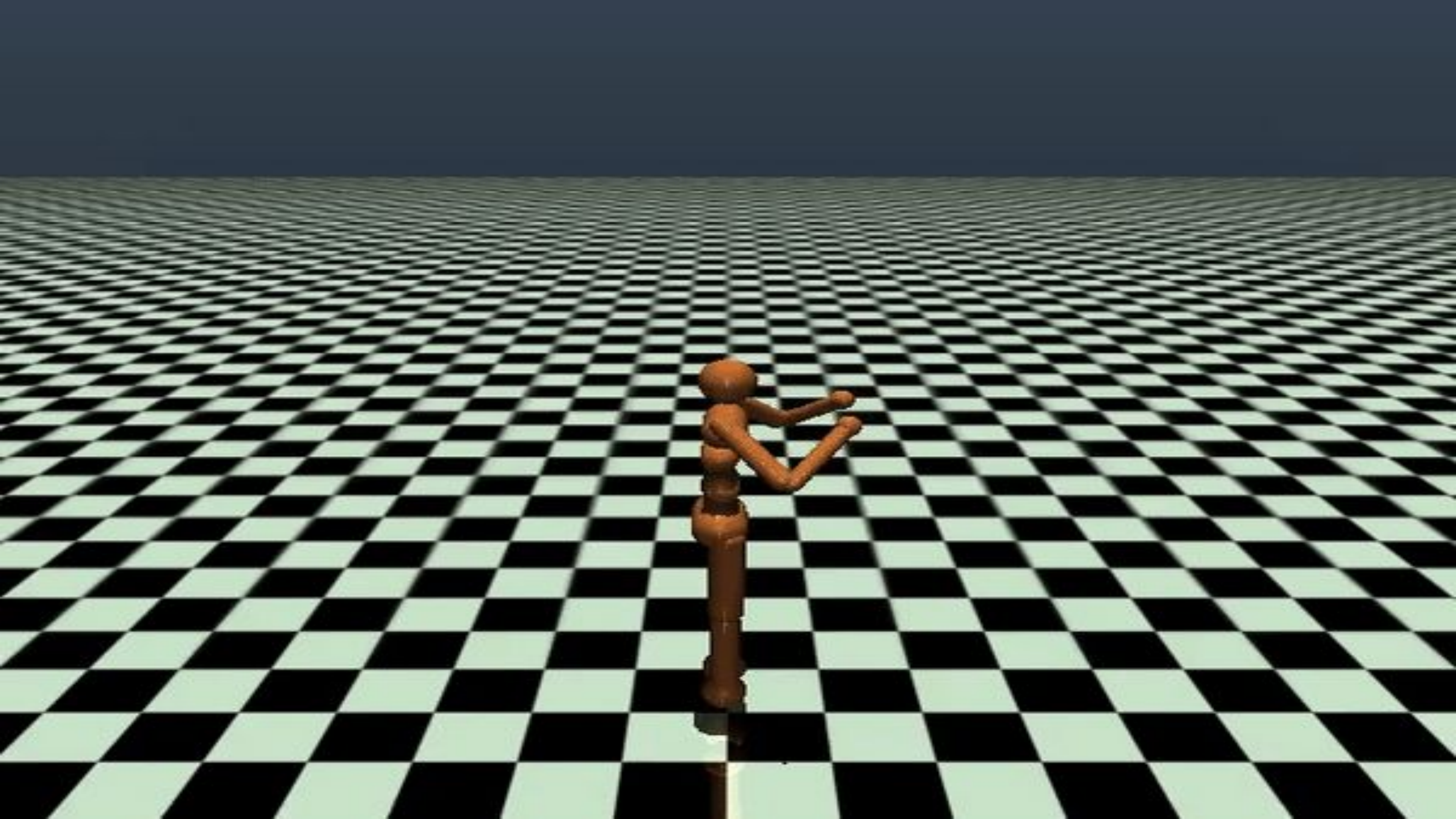
Vanilla RNN-ortho init

no more spins...



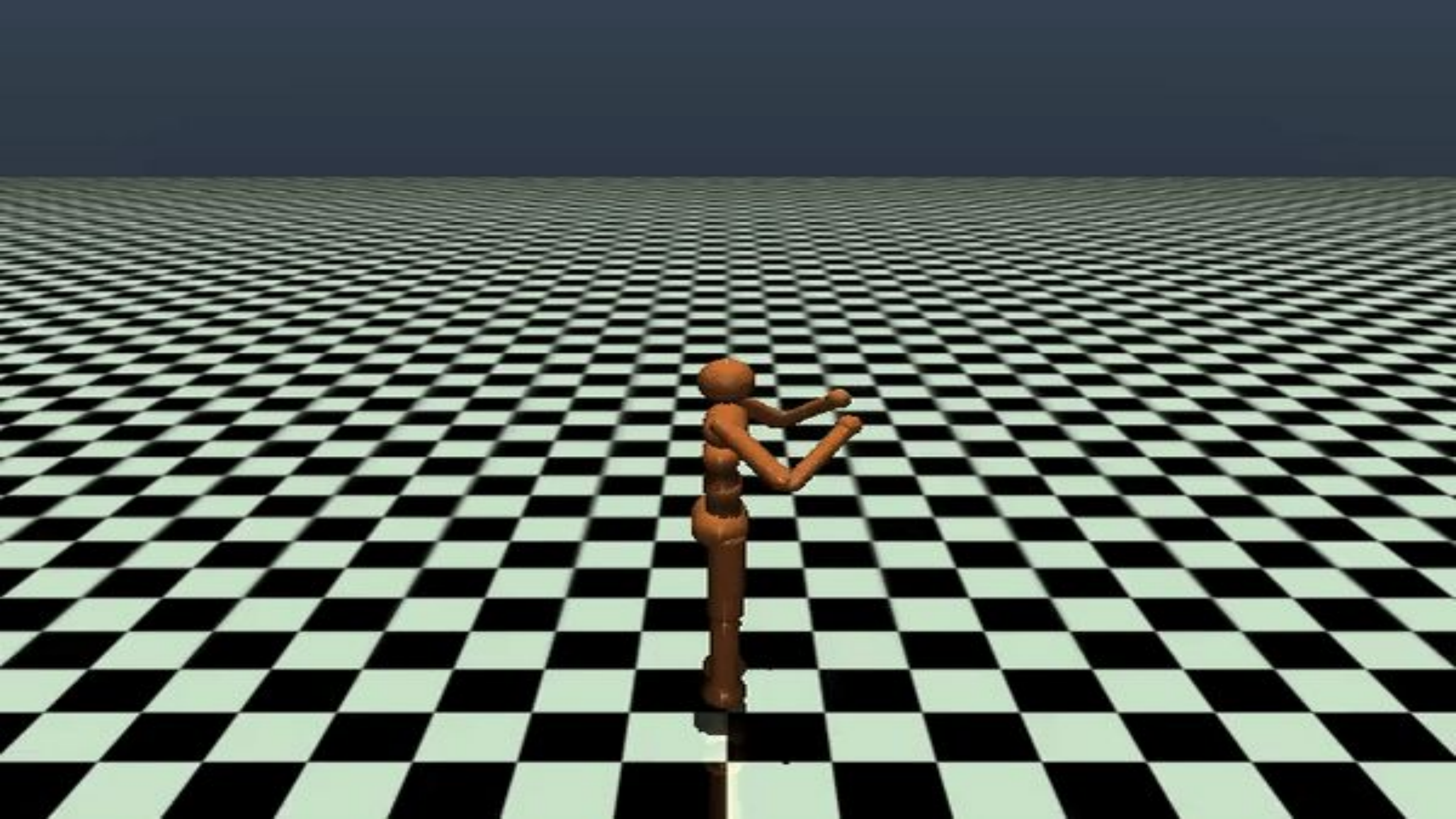
Ortho RNN-nostochastic

getting closer...

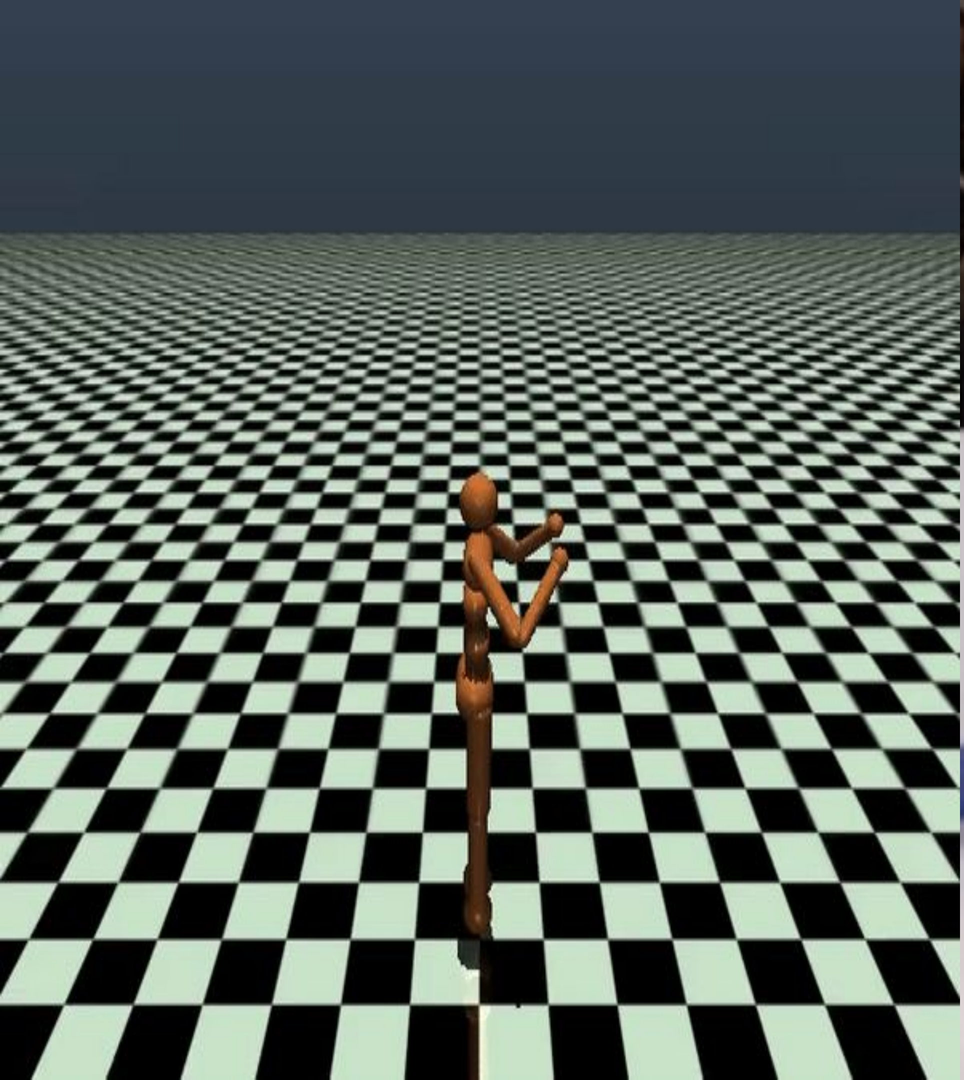


Ortho RNN-stochastic

lets see...



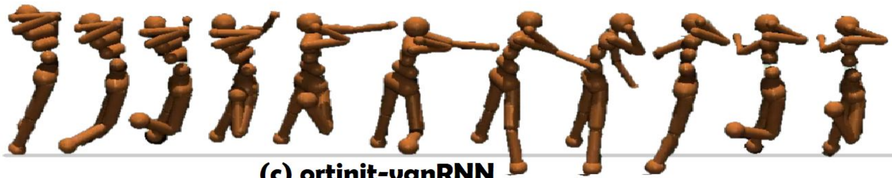
The Pirouette




Prudential
TEAM USA

(a) stoch-ortRNN

(b) exact-ortRNN



(c) ortinit-vanRNN

(d) vanRNN

(e) affine

Thank you for the attention