Online Limited Memory Neural-Linear Bandits with Likelihood Matching

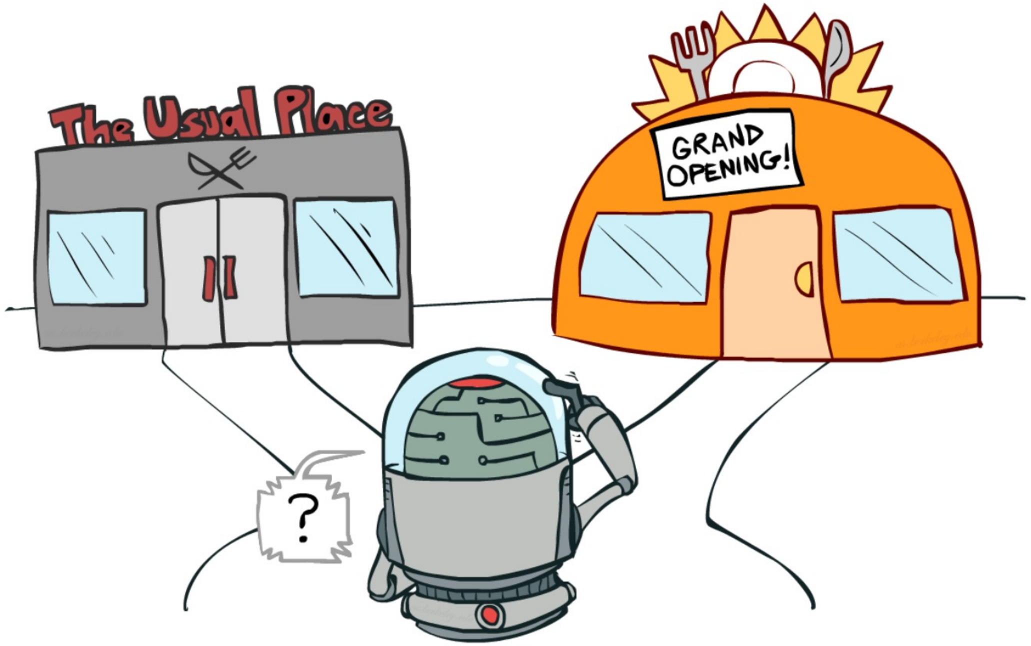
Ofir Nabati[1], Tom Zahavy[1,2] and Shie Mannor[1,3]

ICML 2021

[1]Technion, Israel Institute of Technology
[2]DeepMind
[3]Nvidia Research

Exploration with Neural Networks

- Dropout
- Bootstraping
- ϵ-greedy
- Monte Carlo methods
- Direct Noise Injection
- Variational Auto Encoders
- Neural Linear + Memory constraints!



Contextual Linear Bandits

- Every round we get a context $b(t)$
- We choose an action.
- Get a reward $r_i(t)$
- The expected reward for each action is a linear function

Goal: receive the highest total reward after T rounds.

$$\mathbb{E}[r_i(t)|b(t)] = b(t)^\top \mu_i \,, \qquad i = 1,2,3, \dots, N$$

Thompson Sampling (TS)

Algorithm 1 TS for linear contextual bandits

$\forall i \in [1,..,N]$, set $\Phi_i = 0$, $\Phi_i^0 = I_d$, $\hat{\mu}_i = 0_d$, $\psi_i = 0_d$

for $t = 1, 2, \ldots,$ **do**

  $\forall i \in [1,..,N]$, sample $\tilde{\mu}_i \sim N(\hat{\mu}_i, \nu^2(\Phi_i^0 + \Phi_i)^{-1})$ Posterior sampling

  Play arm $a(t) := \text{argmax}_i b(t)^\top \tilde{\mu}_i$

  Observe reward r_t

  Posterior update:

  Covariance → $\Phi_{a(t)} = \Phi_{a(t)} + b(t)b(t)^\top$

  $\psi_{a(t)} = \psi_{a(t)} + b(t)r_t$

  Mean → $\hat{\mu}_{a(t)} = (\Phi_i^0 + \Phi_{a(t)})^{-1}\psi_{a(t)}$

  Posterior parameters

end for

Agrawal, Shipra, and Navin Goyal. "*Thompson sampling for contextual bandits with linear payoffs.*" *International Conference on Machine Learning*. PMLR, 2013.

Neural Linear Bandits

- Linear exploration policy (TS) on top of the last hidden layer of a neural network

$$\phi(t) = LastNetworkLayer(b(t))$$

- Network is trained in phases to predict rewards.
- State-of-the-art method.
- Assumption: $\mathbb{E}[r_i(t)|\phi\ (t)] = \phi(t)^\top \mu_i$
- Every time the representation is changed, recompute the posterior.
- Memory is unlimited.
- Priors are fixed: $\Phi^0 = I, \;\; \mu^0 = 0$

Riquelme, C., Tucker, G., and Snoek, J. "Deep Bayesian bandits showdown." *International Conference on Learning Representations*, 2018.

Limited Memory Case: Catastrophic Forgetting

- Memory size is <u>limited.</u>
- Each representation update, there is an information loss.
- This causes performance degradation.

Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." *Proceedings of the national academy of sciences 114.13 (2017): 3521-3526*.

The Big Quesiton:

How to solve representation drift without suffering from catastrophic forgetting?

Our Solution:

Limited Memory Neural Bandits with Likelihood Matching (LiM2)

Likelihood Matching

- We want to preserve past information before the update.
- We store the information at the posterior's priors Φ_i^0 and μ_i^0 under the new representation.

This is done by matching the likelihood of the reward before and after the updates:

$$\textit{Find priors } \Phi_i^0 \textit{ and } \hat{\mu}_i^0 \textit{ such that } \forall b_j \in \textit{Memory with action } i:$$

Variance matching: $\underbrace{\phi_j^{old}(t)^\top \left(\Phi_i^{old}\right)^{-1} \phi_j^{old}(t)}_{s_{i,j}^2} = \phi_j^{new}(t)^\top \left(\Phi_i^0\right)^{-1} \phi_j^{new}(t)$

Mean matching: $\phi_j^{old}(t)^\top \hat{\mu}_i^{old} = \phi_j^{new}(t)^\top \hat{\mu}_i^0$

Likelihood Matching

Find priors Φ_i^0 *and* $\hat{\mu}_i^0$ *such that* $\forall b_j \in$ *Memory with action* i:

Variance matching: $\underbrace{\phi_j^{old}(t)^\top \left(\Phi_i^{old}\right)^{-1} \phi_j^{old}(t)}_{s_{j,i}^2} = \phi_j^{new}(t)^\top \left(\Phi_i^0\right)^{-1} \phi_j^{new}(t)$

Mean matching: $\phi_j^{old}(t)^\top \hat{\mu}_i^{old} = \phi_j^{new}(t)^\top \hat{\mu}_i^0$

Computing Φ_i^0 via SDP:

$$\underset{(\Phi_i^0)^{-1}}{\text{minimize}} \sum_{j=1}^{n_i} \left(\text{Trace}(X_{j,i}^\top (\Phi_i^0)^{-1}) - s_{j,i}^2\right)^2$$

$$\text{subject to} \quad (\Phi_i^0)^{-1} \succeq 0.$$

where $X_{j,i} \triangleq \phi_j \phi_j^\top$

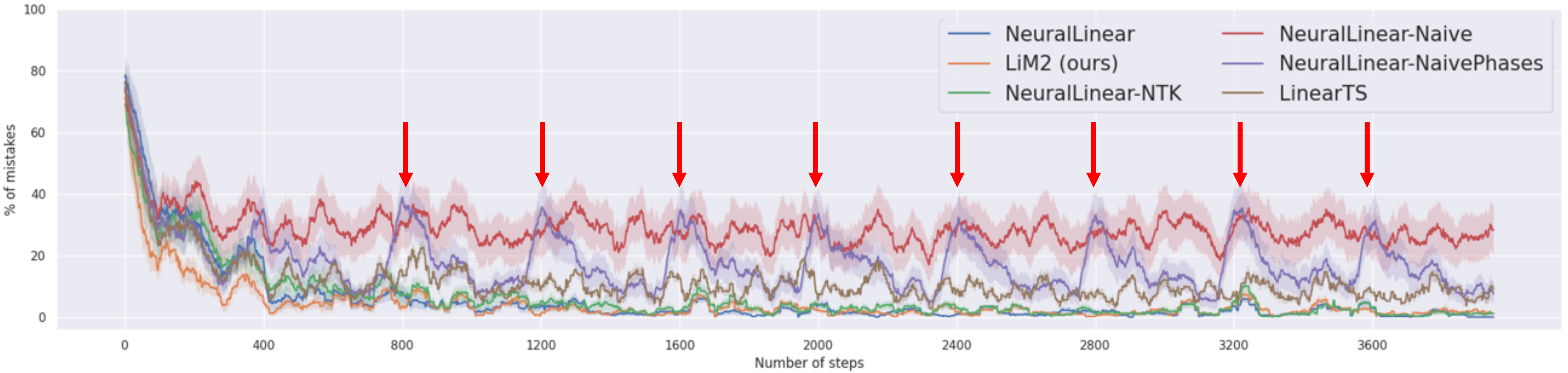
Computing $\hat{\mu}_i^0$: taking the weights of the last layer makes a good prior.

Solving the SDP

$$\underset{(\Phi_i^0)^{-1}}{\text{minimize}} \sum_{j=1}^{n_i} \left(\text{Trace}(X_{j,i}^\top (\Phi_i^0)^{-1}) - s_{j,i}^2\right)^2$$
$$\text{subject to} \quad (\Phi_i^0)^{-1} \succeq 0.$$

- Computationally prohibitive.
- We solve the SDP by applying stochastic gradient decent (SGD).
- Project the covariance matrix back to PSD space by eigenvalues thresholding.
- We can use the same batch for network training and likelihood matching!
- Online mode - applying only one iteration each round.

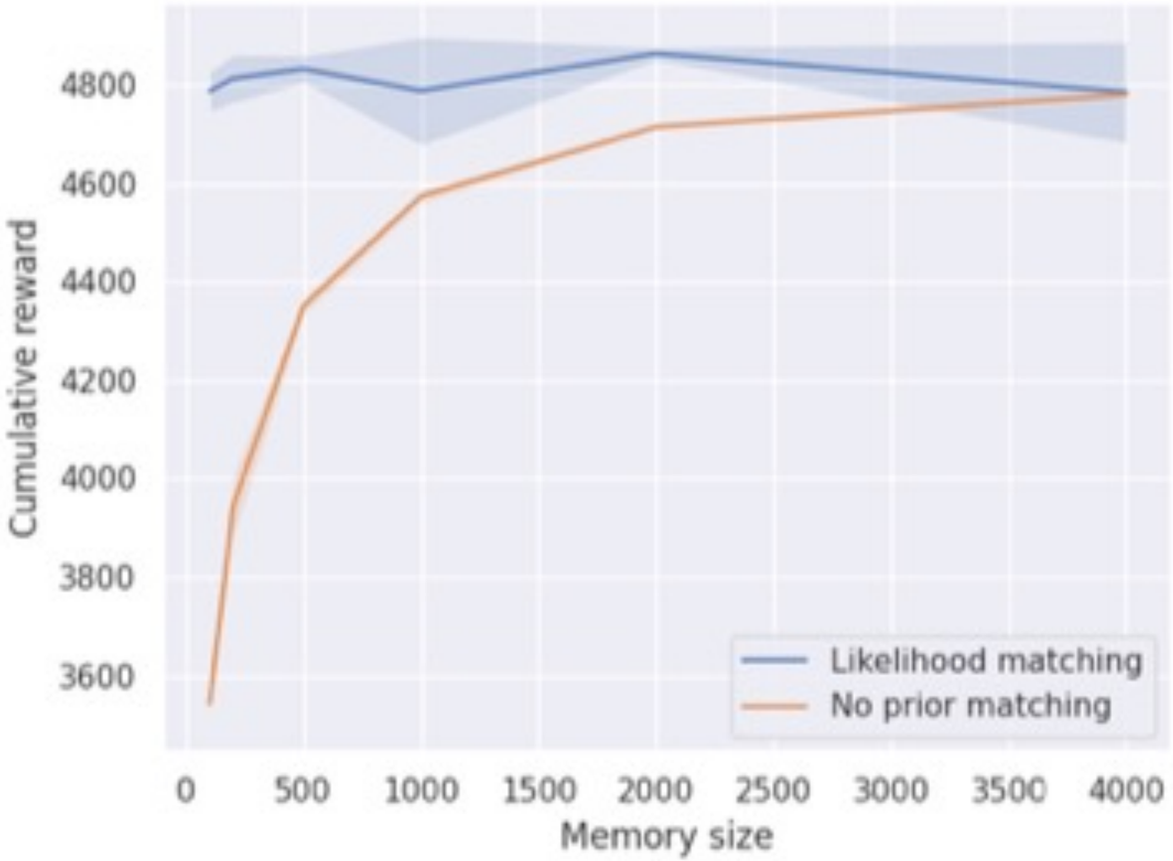
Results - Catastrophic Forgetting



- LiM2 eliminates catastrophic forgetting.
- Naive approach suffers from degradation each network update.

Results – Memory Size

- Naive approach does not cope well with limited memory.
- LiM2 is robust to memory size.

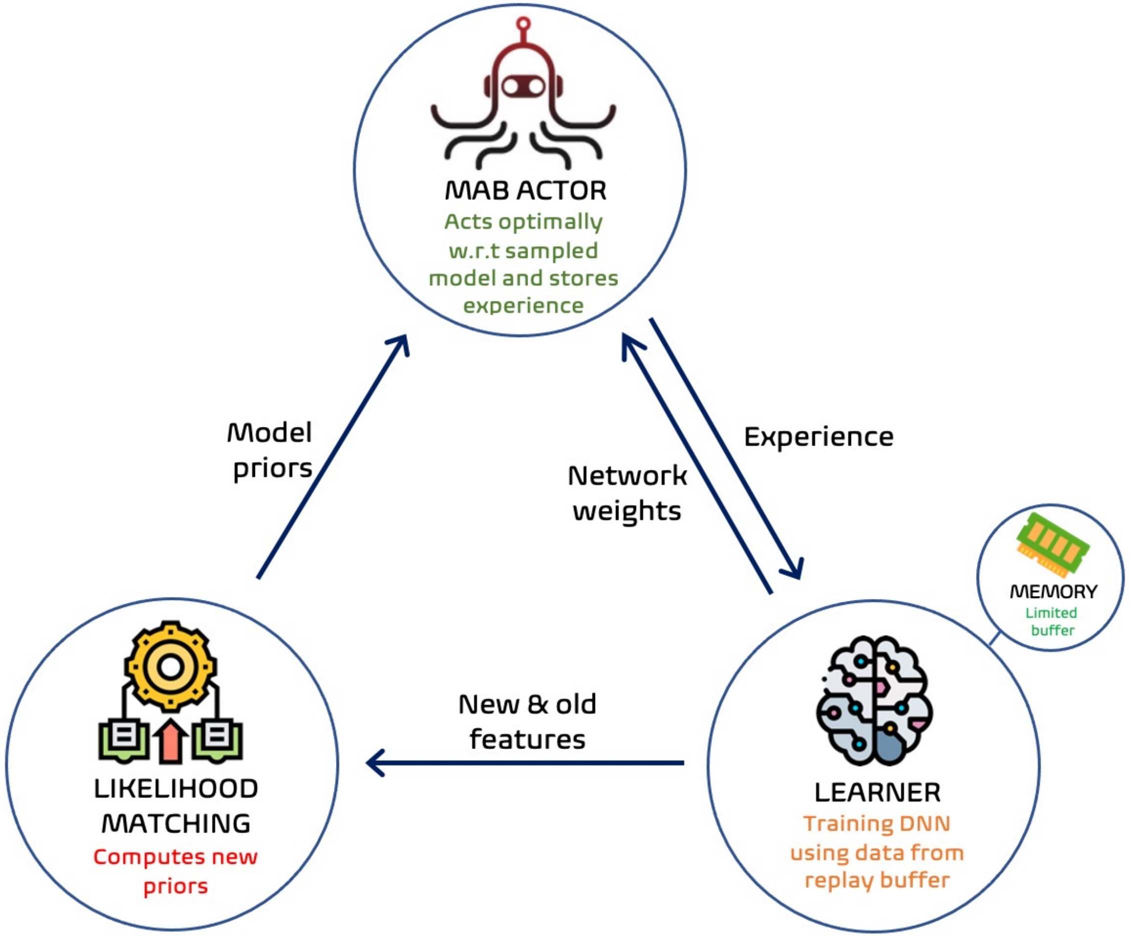


Results – Real Datasets

			Full memory		Limited memory			NTK based		
Name	d	A	LinearTS	NeuralLinear	LiM2 (Ours)	NeuralLinear-MM	NeuralLinear-Naive	NeuralUCB	NeuralTS	NeuralLinear-NTK
Mushroom	117	2	**1.000**	0.985	0.945	0.719	0.730	0.521	0.521	0.941
Financial	21	8	0.997	0.946	**1.000**	0.743	0.723	0.292	0.228	0.959
Jester	32	8	**1.000**	0.784	0.819	0.287	0.234	0.546	0.546	0.768
Adult	88	2	0.977	0.974	**1.000**	0.638	0.634	0.822	0.823	0.966
Covertype	54	7	**1.000**	0.902	0.892	0.679	0.693	0.514	0.517	0.887
Census	377	9	0.548	0.860	**1.000**	0.679	0.686	0.644	0.603	0.863
Statlog	9	7	0.912	0.978	**1.000**	0.933	0.916	0.818	0.885	0.976
Epileptic	178	5	0.282	**1.000**	0.684	0.562	0.504	0.019	0.020	0.589
Smartphones	561	6	0.649	0.970	**1.000**	0.521	0.515	0.396	0.670	0.965
Scania Trucks	170	2	0.181	0.672	0.745	-0.344	-0.050	0.988	**1.000**	0.259
Amazon	7K	5	-	0.986	**1.000**	0.873	0.879	-	-	0.981
Average			0.755	0.914	**0.917**	0.572	0.588	0.556	0.581	0.832
Median			0.945	0.970	**1.000**	0.679	0.686	0.534	0.575	0.941

Conclusions

- In order to use limited memory without suffering from catastrophic forgetting – LiM2 provides a good robust solution.
- No significant additional computational burden.
- LiM2 enables to operate online.



Thank you!

Contact mail: ofirnabati@gmail.com

For more information see our paper