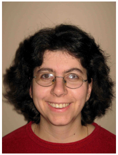
Interference and Generalization in Temporal Difference Learning

Emmanuel Bengio   Joelle Pineau   Doina Precup



ICML 2020

McGill   Mila

Overview

The setting:

- Deep Neural Networks
- Interference: $\rho = \langle \nabla_\theta f(u_1), \nabla_\theta f(u_2) \rangle$
- Data: classification, regression, interactive environments
- Training: supervised vs reinforcement (TD, TD(λ), & PG)

We wish to understand the relation between **interference** and **generalization**, and how **Temporal Difference** affects both.

Key Takeaways

For the **same data**:

- TD tends to induce **unaligned** ($\rho = 0 \pm \epsilon$) representations
- SL tends to induce **aligned** ($\rho > 0$) representations
- increased alignment is correlated with:
    - a **reduced** generalization gap in **TD**
    - an **increased** generalization gap in **SL**
- TD and SL *generalize* differently! Even for RL data
- TD(λ) controls this behaviour ($\lambda = 1$ being $\approx$ SL)

Key Takeaways

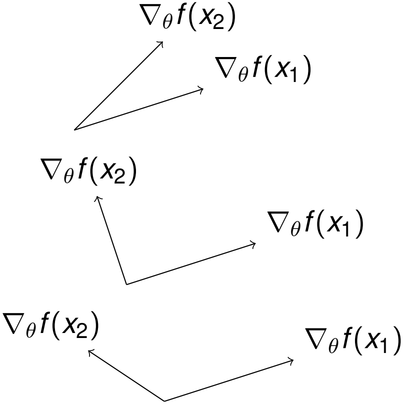
In more intuitive words/conjecture:

For the **same data**:

- TD tends to **memorize** its data
- SL tends to **generalize**
- further training:
  - **breaks** memorized structures in **TD**
  - **creates** memorized structures in **SL** (overfitting)
- TD and SL *generalize* differently! Even for RL data
- TD(λ) controls this behaviour ($\lambda = 1$ being $\approx$ SL)

Interference

- $\rho > 0$
- $\rho = 0$
- $\rho < 0$



$$\Delta f(x_2) = \alpha \nabla_\theta^T f(x_2) \nabla_\theta f(x_1)$$

Interference

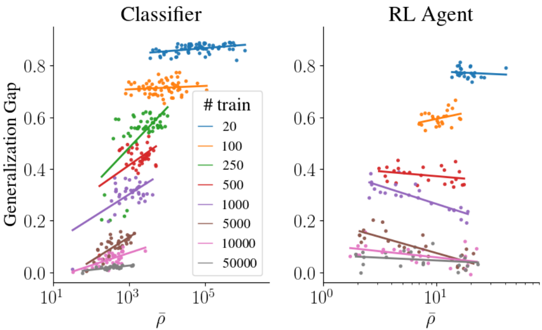
- Taylor expansion:

$$f(x,\theta') = f(x,\theta) + \underbrace{\nabla_\theta f(x)^T(\theta' - \theta)} + (\theta' - \theta)^T \nabla_\theta^2 f(x)(\theta' - \theta) + ...$$

- stiffness (Fort et al., 2019):

$$\text{angle}(\nabla f(x_1), \nabla f(x_2)) = \frac{\nabla f(x_1)^T \nabla f(x_2)}{\|\nabla f(x_1)\| \|\nabla f(x_2)\|}$$

Classification



Overfitting manifests differently

Supervised Data

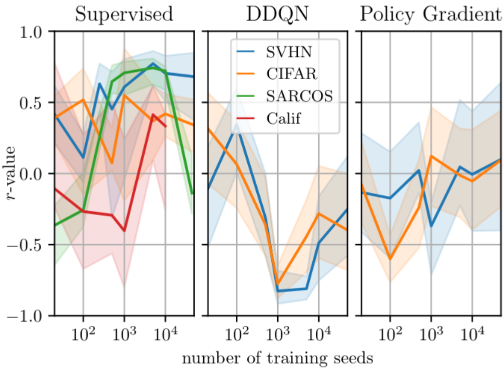
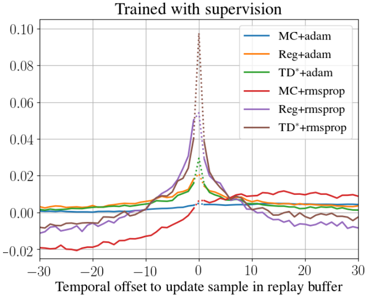
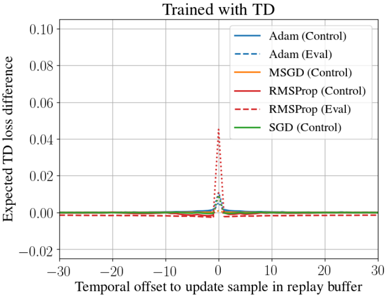


Figure 1. Correlation coefficient r between the (log) function interference $\bar{\rho}$ and the generalization gap, as a function of training set size; shaded regions are bootstrapped 90% confidence intervals. We see different trends for value-based experiments (middle) than for supervised (left) and PG experiments (right).

Atari

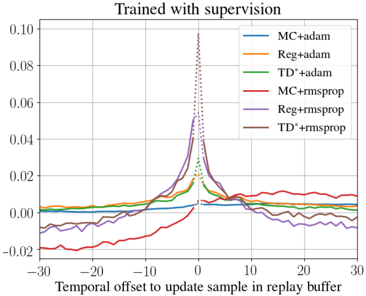
Measuring gain (effective loss interference) for nearby states:

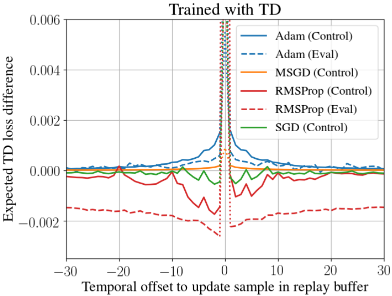




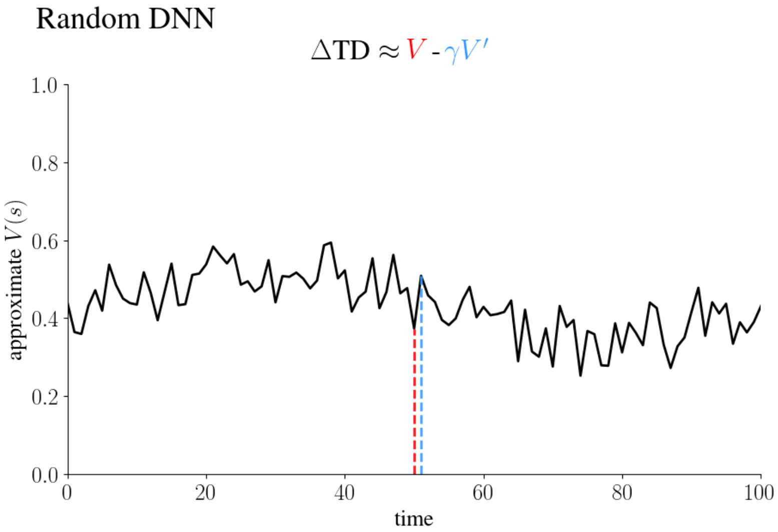
Atari

Measuring gain (effective loss interference) for nearby states:

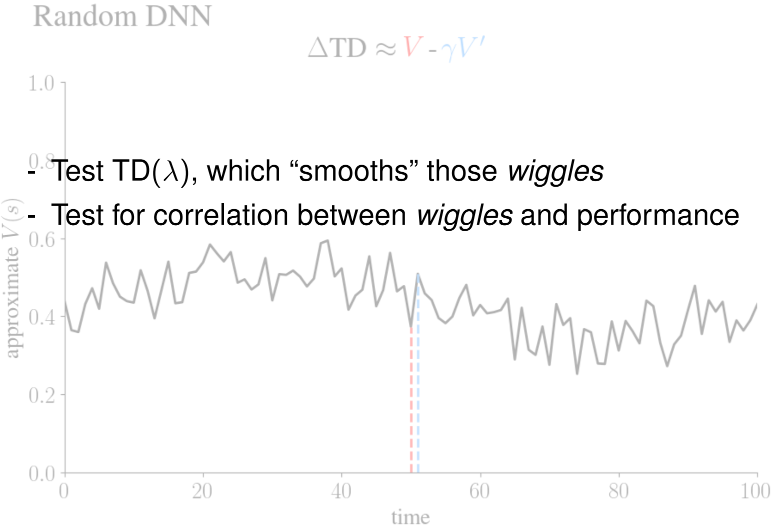




Understanding interference in TD



Understanding interference in TD



- Test TD(λ), which “smooths” those *wiggles*
- Test for correlation between *wiggles* and performance

TD(λ)

TD(λ) smooths the TD target by taking into account (weighed) future predictions:

$$G^{\lambda}(S_t) = (1 - \lambda) \sum_{n=1}^{\infty} \lambda^{n-1} G^{n}(S_t) \tag{1}$$

$$G^{n}(S_t) = \gamma^{n} V(S_{t+n}) + \sum_{j=0}^{n-1} \gamma^{j} R(S_{t+j}) \tag{2}$$

TD(λ)

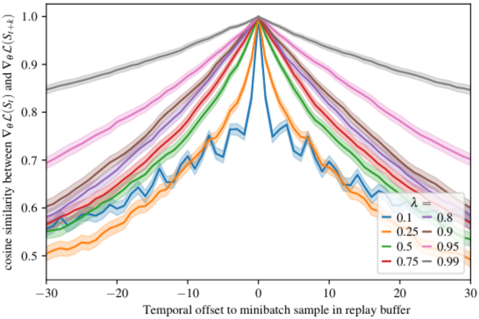
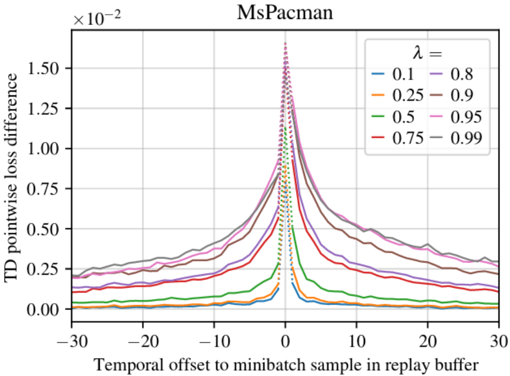


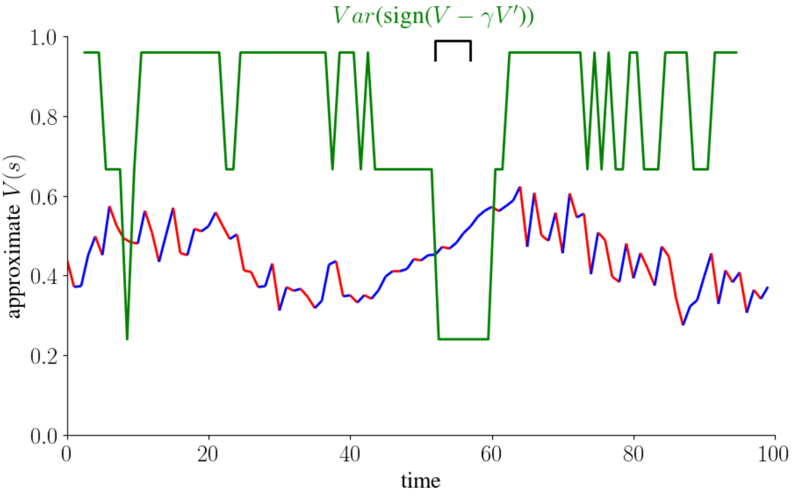
Figure 5. Cosine similarity between gradients at S_t (offset $x = 0$) and the gradients at the neighboring states in the replay buffer (MsPacman). As λ increases, so does the temporal coherence of the gradients.

TD(λ)

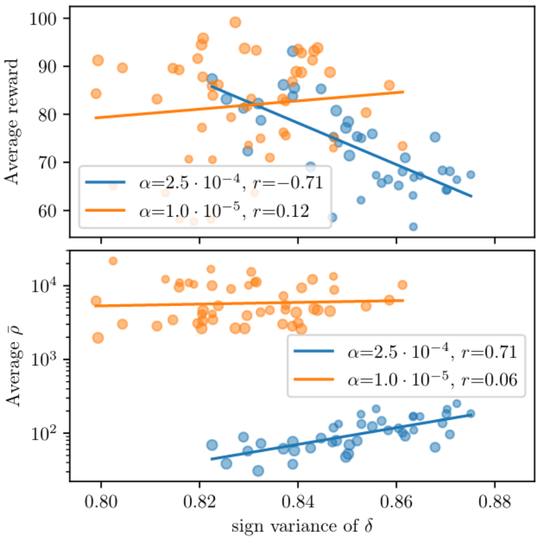
Increasing λ increases how fast the loss decreases (around s_t)



Local prediction variance



Local prediction variance



Interference update decomposition

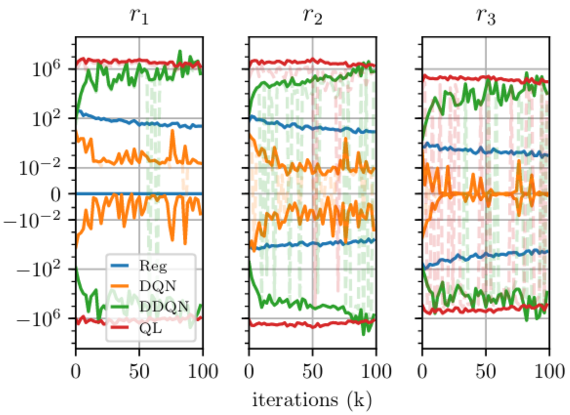
Two extra terms in the TD update's interference time derivative:

$$
\begin{aligned}
\rho'_{reg;AB} &= -\bar{\rho}^2_{AB}\delta^2_B - 2\delta_A\delta_B\bar{\rho}_{AB}\bar{\rho}_{BB} \\
&\quad - \delta_A\delta^2_B\nabla f_B(\bar{H}_A\nabla f_B + \bar{H}_B\nabla f_A) \\
\rho'_{TD;AB} &= -\delta^2_B\bar{\rho}_{AB}(\bar{\rho}_{AB} - \gamma\bar{\rho}_{A'B}) - \delta_A\delta_B\bar{\rho}_{AB}(\bar{\rho}_{BB} - \gamma\bar{\rho}_{B'B}) \\
&\quad - \delta_A\delta^2_B\nabla f_B(\bar{H}_A\nabla f_B + \bar{H}_B\nabla f_A)
\end{aligned}
$$

$\rightarrow$ gradient variance induced by errors in predictions will be much larger for a high-capacity high-variance model

Interference update decomposition

DDQN and QL (no frozen target) have unstable updates, unlike Regression and DQN (frozen target):



:)

Recap & Conclusion

- generalization dynamics in SL and DL $\rightarrow$ different parameterizations.
- in RL tasks, TD doesn't generalize as well as SL (even when the f to approximate is the same)
- find link between the complexity and variance of TD targets and interference
- TD(λ) has generalization potential
- better optimizers for TD might improve things quite a lot!