Google Research


(Illustrator: Theodor Hosemann)

Munchausen Reinforcement Learning

Matthieu Geist (Google)
Joint work with Nino Vieillard (Google, Univ. Lorraine) and Olivier Pietquin (Google)

Bootstrapping values is ubiquitous in Reinforcement Learning

- Would the optimal value function be known:

$$\hat{q}(s_t, a_t) \leftarrow \hat{q}(s_t, a_t) + \eta(q_*(s_t, a_t) - \hat{q}(s_t, a_t))$$

- Would the optimal value function be known in the transiting state:

$$\hat{q}(s_t, a_t) \leftarrow \hat{q}(s_t, a_t) + \eta(r(s_t, a_t) + \gamma \max q_*(s_{t+1}, \cdot) - \hat{q}(s_t, a_t))$$

- But it is unknown, replace it by the current estimate:

$$\hat{q}(s_t, a_t) \leftarrow \hat{q}(s_t, a_t) + \eta(r(s_t, a_t) + \gamma \max \hat{q}(s_{t+1}, \cdot) - \hat{q}(s_t, a_t))$$

- This is bootstrapping:
    - Gives q-learning here
    - Bootstrapping the value is ubiquitous in RL… but what about other quantities?

Bootstrapping the policy

- Core idea: augment the reward with the log-policy

$$r(s_t, a_t) \rightarrow r(s_t, a_t) + \alpha \ln \hat{\pi}(a_t|s_t)$$

- Rational
  - Assume that the optimal policy is known, $\ln \pi_*(a|s) = \begin{cases} 0 \text{ if } a \text{ is optimal} \\ -\infty \text{ else} \end{cases}$
  - Very strong learning signal!
  - But it is unknown, replace it by the estimated policy
- Munchausen Reinforcement Learning:
  - Augment the reward with the scaled log-policy (assuming a stochastic policy)
  - Different from MaxEnt RL, that subtracts the scaled log-policy
  - Named as a reference to Baron Munchausen, who pulls himself out of a swamp by pulling on his own hair

Case study: DQN

- Let's modify DQN with the Munchausen term to get Munchausen-DQN
- We'll only modify the regression target of DQN:

$$\hat{q}_{\text{dqn}}(r_t, s_{t+1}) = r_t + \gamma \sum_{a' \in \mathcal{A}} \pi_{\bar{\theta}}(a'|s_{t+1}) q_{\bar{\theta}}(s_{t+1}, a') \text{ with } \pi_{\bar{\theta}} \in \mathcal{G}(q_{\bar{\theta}})$$

- We need a stochastic policy, so just add some entropy regularization:

$$\hat{q}_{\text{s-dqn}}(r_t, s_{t+1}) = r_t + \gamma \sum_{a' \in \mathcal{A}} \pi_{\bar{\theta}}(a'|s_{t+1}) \Big(q_{\bar{\theta}}(s_{t+1}, a') - \tau \ln \pi_{\bar{\theta}}(a'|s_{t+1}) \Big) \text{ with } \pi_{\bar{\theta}} = \text{softmax}(\frac{q_{\bar{\theta}}}{\tau})$$

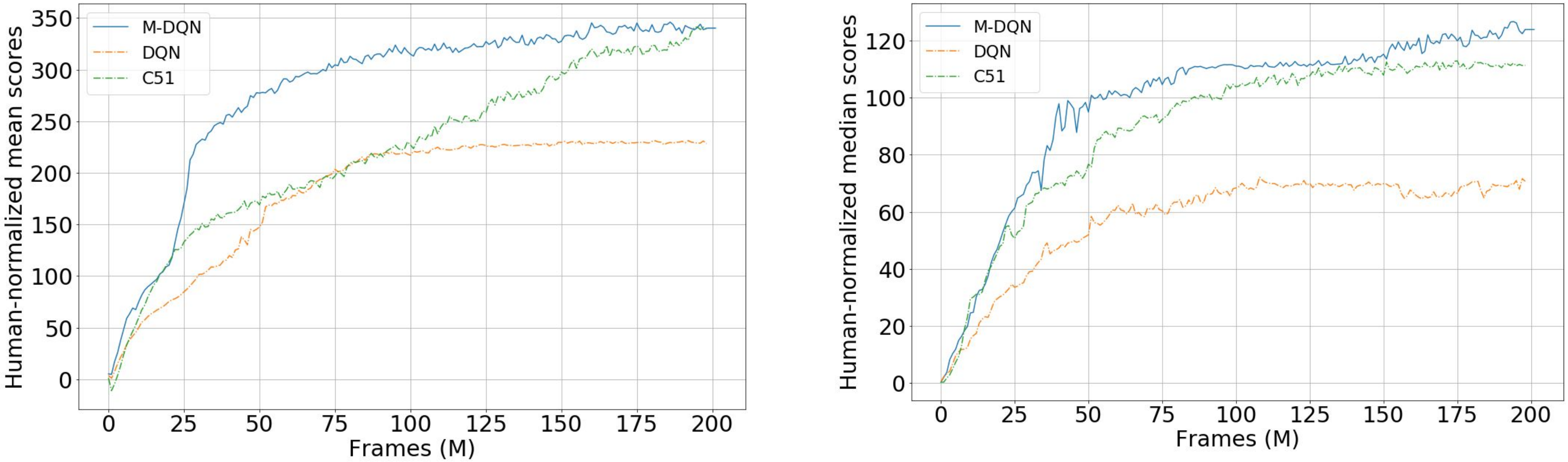
- Then, we just have to add the Munchausen term ($\pi_{\bar{\theta}}$ as above):

$$\hat{q}_{\text{m-dqn}}(r_t, s_{t+1}) = r_t + \alpha \tau \ln \pi_{\bar{\theta}}(a_t|s_t) + \gamma \sum_{a' \in \mathcal{A}} \pi_{\bar{\theta}}(a'|s_{t+1}) \Big(q_{\bar{\theta}}(s_{t+1}, a') - \tau \ln \pi_{\bar{\theta}}(a'|s_{t+1}) \Big)$$

- (notice that the log-policy terms have different signs)
- That's it!

Case study: DQN

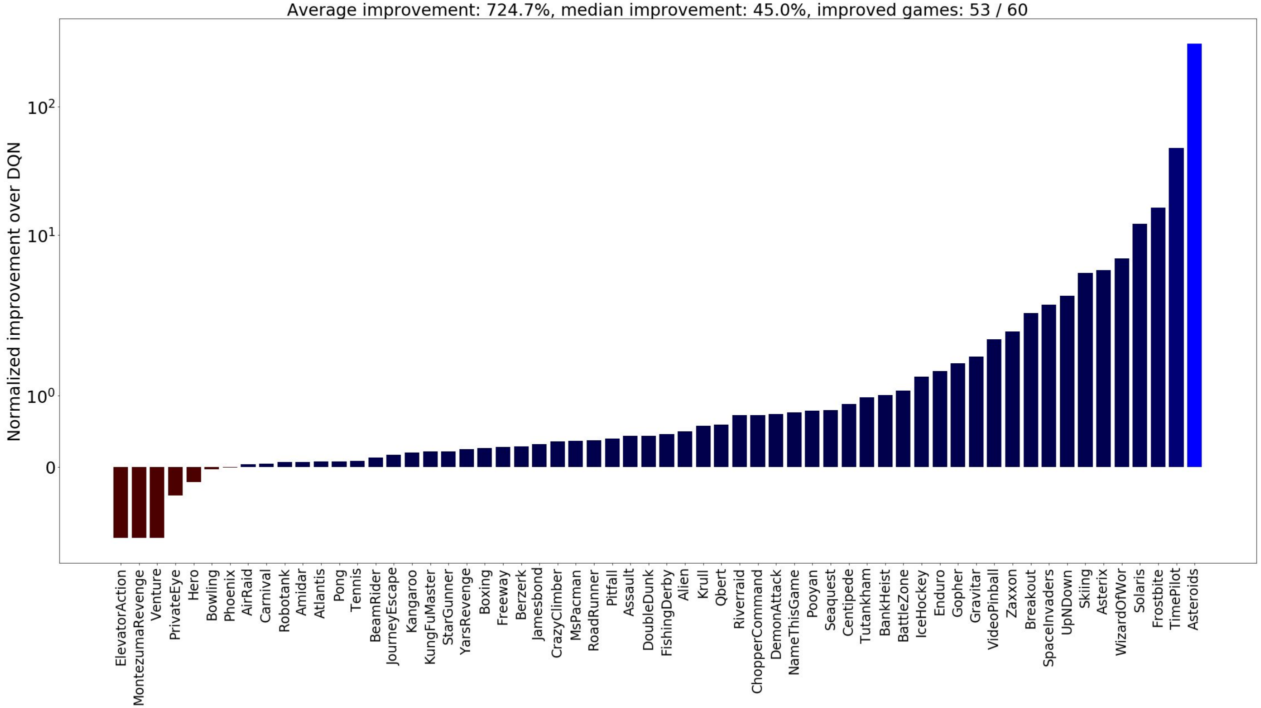
- How good is Munchausen-DQN compared to DQN?
  - Aggregated results on the 60 Atari games of ALE, with also C51



Google Research

Case study: DQN

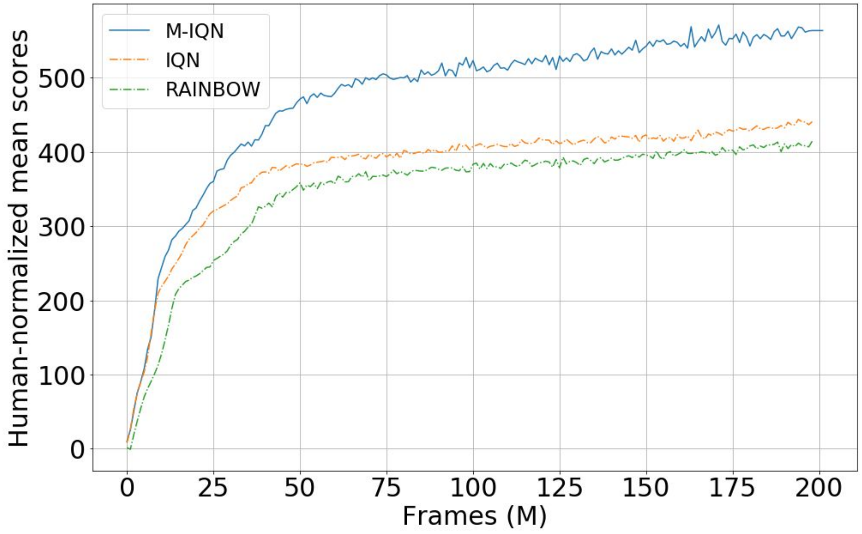
- How good is Munchausen-DQN compared to DQN?
    - Per game improvement

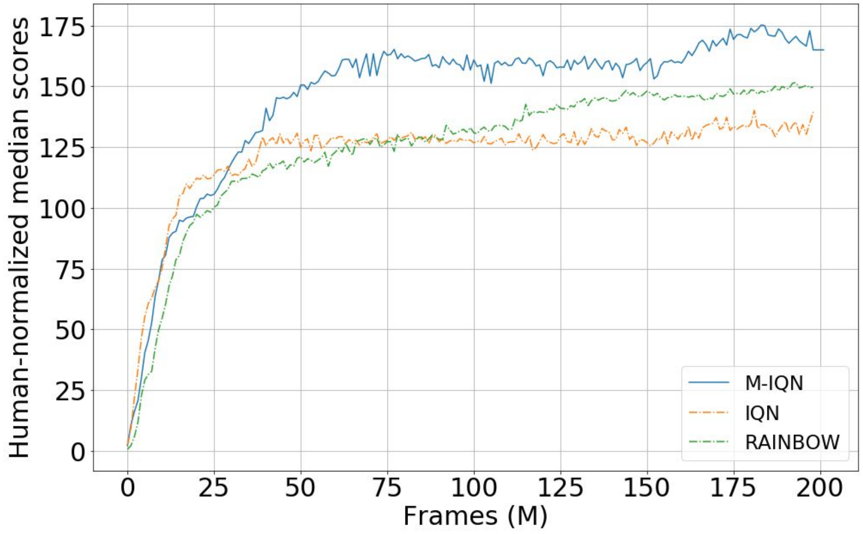


Google Research

Case study: IQN

- This is a general approach. As an example, we apply it to IQN
- Munchausen-IQN vs IQN, aggregated results over 60 games





What happens under the hood?

Two main things:

- **Implicit KL regularization**:
  - Performs KL regularization without error in the greedy step
  - Very strong performance bound, that applies in the deep learning setting

- **Increase of the action gap**:
  - Munchausen generalizes advantage learning
  - For Munchausen, we can quantify analytically the increase of the action-gap

Implicit KL regularization

$$\begin{cases} \pi_{k+1} = \operatorname{argmax}_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \langle \pi, q_k \rangle + \tau \mathcal{H}(\pi) \\ q_{k+1} = r + \alpha\tau \ln \pi_{k+1} + \gamma P \langle \pi_{k+1}, q_k - \tau \ln \pi_{k+1} \rangle + \epsilon_{k+1}. \end{cases}$$

Abstraction of M-DQN

- The solution to the greedy step is the policy being softmax over q-values
  - Can be computed analytically, even with neural nets
- The evaluation equation is the M-DQN update
  - The error term is the difference between the actual update and the ideal one

Implicit KL regularization

Abstraction of explicit KL-regularized RL

- Analysed in "Leverage the Average: an analysis of regularization in RL"
  - Strong bounds
  - Abstracts TRPO, MPO, and more
- The solution to the greedy step is $\pi_{k+1} \propto \pi_k^{\alpha} \exp(q_k/\tau)$
  - Could be computed analytically for a linear parameterization
  - Cannot be computed analytically for a nonlinear one (neural network!)
  - Requires an actor, so there's error in the greedy step, breaks the analysis

$$\begin{cases} \pi_{k+1} = \operatorname{argmax}_{\pi \in \Delta_{\mathcal{A}}^{S}} \langle \pi, q'_k \rangle - \alpha\tau \, \mathrm{KL}(\pi || \pi_k) + (1-\alpha)\tau \mathcal{H}(\pi) \\ q'_{k+1} = r + \gamma P(\langle \pi_{k+1}, q'_k \rangle - \alpha\tau \, \mathrm{KL}(\pi_{k+1} || \pi_k) + (1-\alpha)\tau \mathcal{H}(\pi_{k+1})) + \epsilon_{k+1} \end{cases}$$

Implicit KL regularization

$$\begin{cases} \pi_{k+1} = \operatorname{argmax}_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \langle \pi, q_k \rangle + \tau \mathcal{H}(\pi) \\ q_{k+1} = r + \alpha \tau \ln \pi_{k+1} + \gamma P \langle \pi_{k+1}, q_k - \tau \ln \pi_{k+1} \rangle + \epsilon_{k+1}. \end{cases}$$

$$\Updownarrow \qquad (q'_k \triangleq q_k - \tau \ln \pi_k)$$

$$\begin{cases} \pi_{k+1} = \operatorname{argmax}_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \langle \pi, q'_k \rangle - \alpha \tau \operatorname{KL}(\pi || \pi_k) + (1-\alpha) \tau \mathcal{H}(\pi) \\ q'_{k+1} = r + \gamma P (\langle \pi_{k+1}, q'_k \rangle - \alpha \tau \operatorname{KL}(\pi_{k+1} || \pi_k) + (1-\alpha) \tau \mathcal{H}(\pi_{k+1})) + \epsilon_{k+1} \end{cases}$$

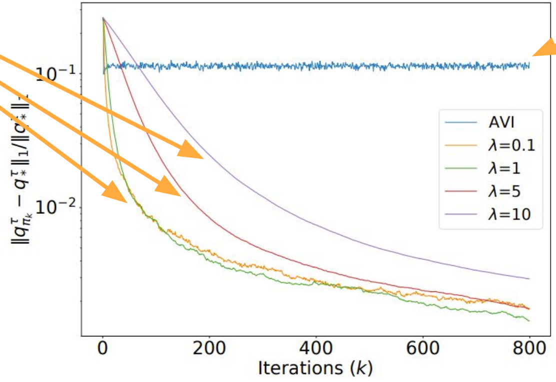
Implicit KL regularization

- As a consequence, a strong performance bound applies to M-DQN
  - (more bounds, more general, in the paper)

Munchausen-DQN vs DQN

$$\|q_* - q_{\pi_k}\|_\infty \leq \frac{2}{1-\gamma}\left\|\frac{1}{k}\sum_{j=1}^{k}\epsilon_j\right\|_\infty + \frac{4}{(1-\gamma)^2}\frac{r_{\max} + \tau \ln|\mathcal{A}|}{k}$$

$$\|q_* - q_{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\left((1-\gamma)\sum_{j=1}^{k}\gamma^{k-j}\|\epsilon_j\|_\infty\right) + \frac{2}{1-\gamma}\gamma^k v_{\max}$$



Increasing the action gap

- Recall the M-DQN regression target ($\pi_{\bar\theta} = \mathrm{softmax}(\frac{q_{\bar\theta}}{\tau})$)

$$\hat{q}_{\text{m-dqn}}(r_t, s_{t+1}) = r_t + \alpha\tau \ln \pi_{\bar\theta}(a_t|s_t) + \gamma \sum_{a' \in \mathcal{A}} \pi_{\bar\theta}(a'|s_{t+1}) \Big(q_{\bar\theta}(s_{t+1}, a') - \tau \ln \pi_{\bar\theta}(a'|s_{t+1}) \Big)$$

- Rewrite the Munchausen term

$$\tau \ln \pi_{\bar\theta}(a|s) = \tau \ln \left(\frac{\exp \frac{q_{\bar\theta}(s,a)}{\tau}}{\sum_{a'} \exp \frac{q_{\bar\theta}(s,a')}{\tau}} \right) = q_{\bar\theta}(s,a) - \tau \ln \left(\sum_{a'} \exp \frac{q_{\bar\theta}(s,a)}{\tau} \right)$$

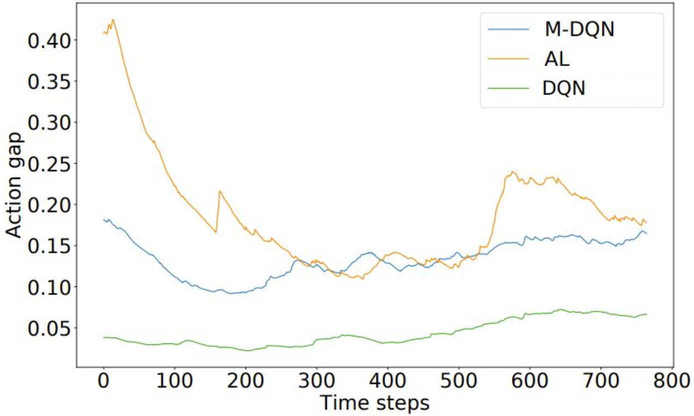
  - Softmax = smoothed argmax (recovered as the temperature goes to zero)
  - Log-sum-exp = smoothed max (idem)
- As the temperature goes to zero, the target becomes the one of advantage learning

$$\hat{q}_{\text{m-dqn}}(r_t, s_{t+1}) \overset{\tau \to 0}{=} r_t + \alpha(q_{\bar\theta}(s_t, a_t) - \max_a q_{\bar\theta}(s_t, a)) + \gamma \max_{a'} q_{\bar\theta}(s_{t+1}, a')$$

Increasing the action gap

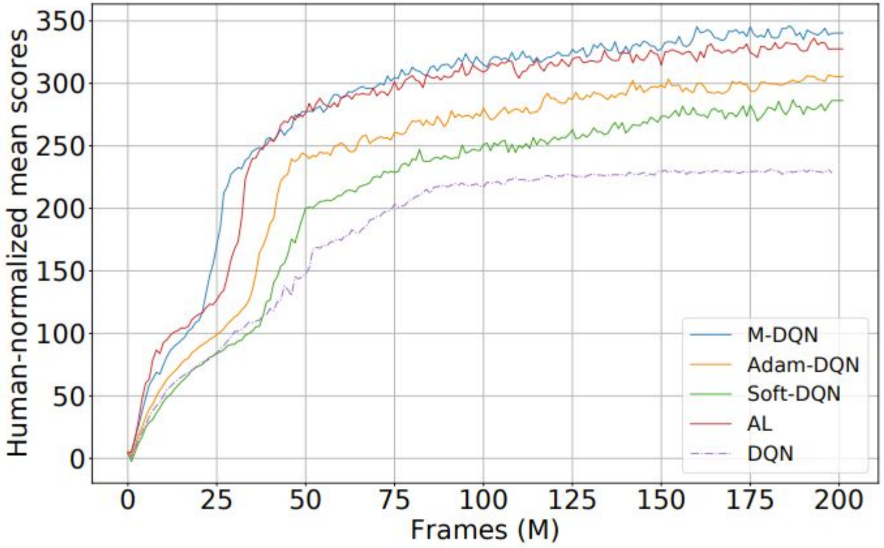
- Define the original action gap as
  - $\text{gap}_*^\tau(s) = \max_a q_*^\tau(s,a) - q_*^\tau(s,\cdot) \in \mathbb{R}_+^{\mathcal{A}}$
- Define the action gap of the kth iteration of Munchausen-VI, without error, as
  - $\text{gap}_k^{\alpha,\tau}(s) = \max_a q_k(s,a) - q_k(s,\cdot) \in \mathbb{R}_+^{\mathcal{A}}$
- We have that

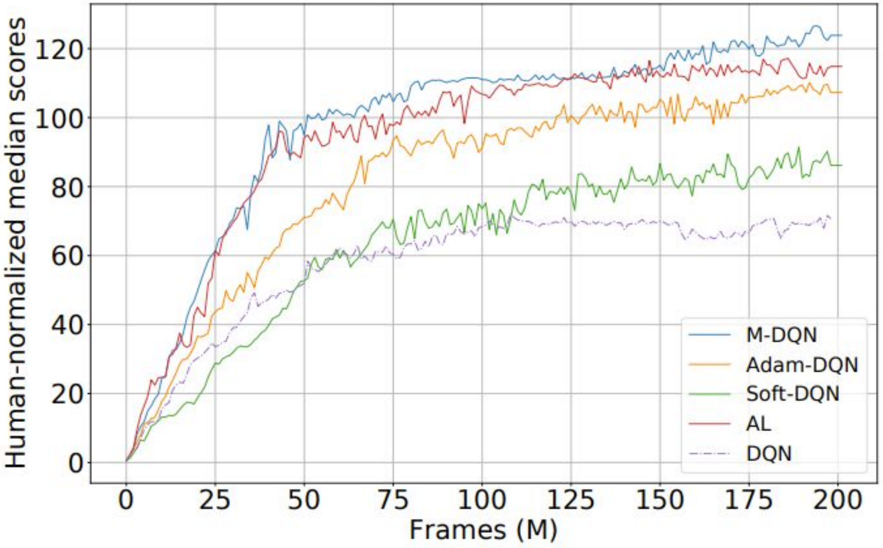
$$\lim_{k\to\infty} \text{gap}_k^{\alpha,\tau}(s) = \frac{1+\alpha}{1-\alpha}\text{gap}_*^{(1-\alpha)\tau}(s)$$



Experimental study

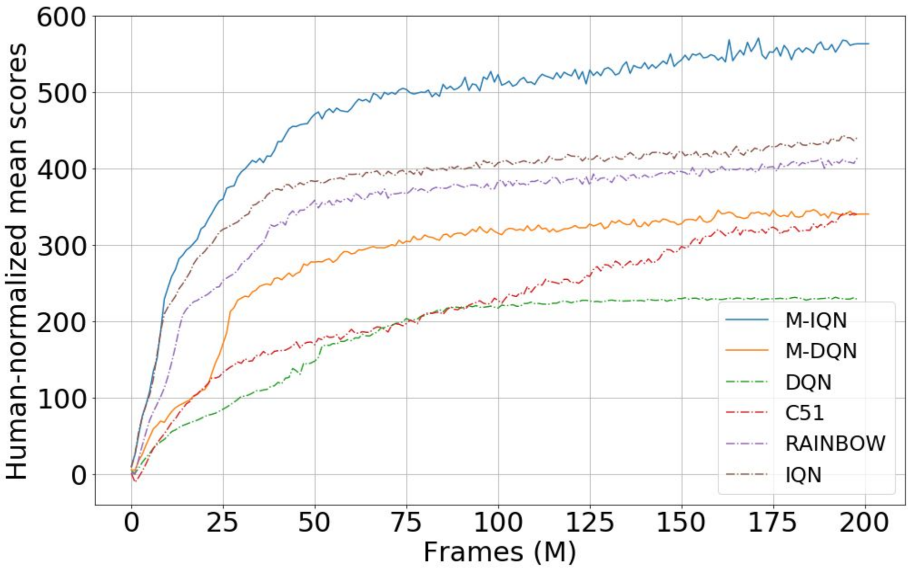
- Ablation

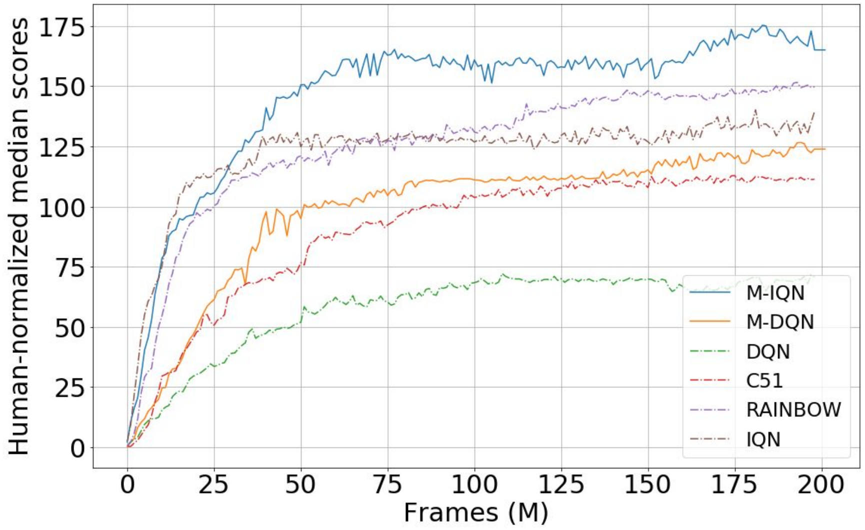




Experimental study

- Vs baselines

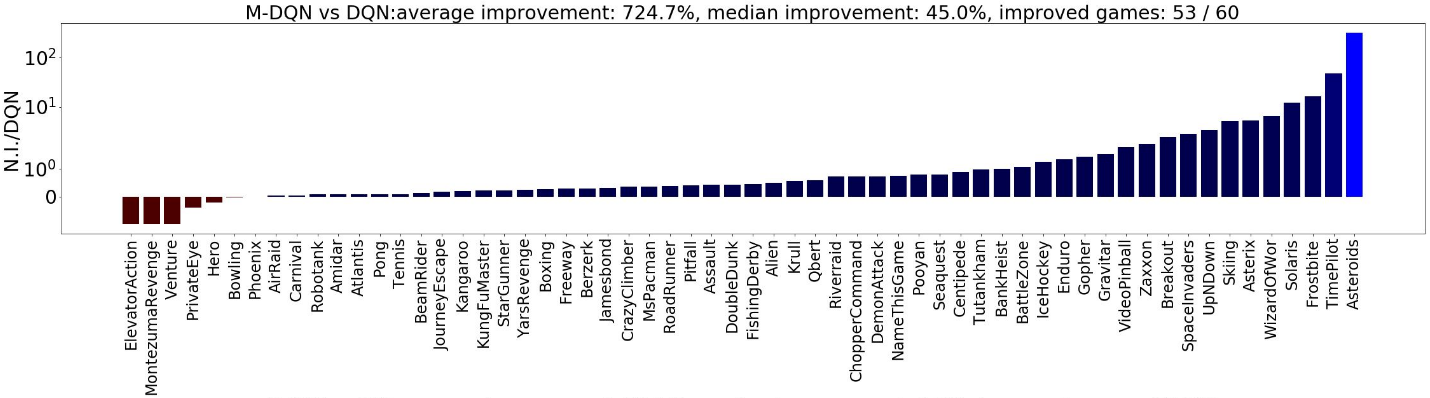


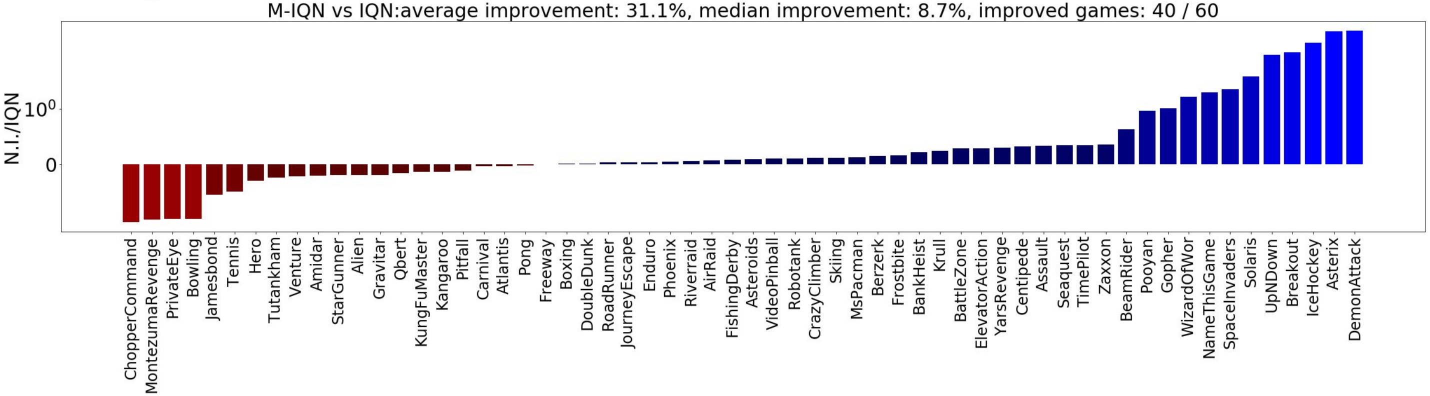


Google Research

Experimental study

- Munchausen improvement (vs modified algorithm)





Google Research

Take home message

- **Munchausen RL** is a very simple idea
  - Augment the reward with the log policy
  - Simple modification of existing agents
- Munchausen RL is theoretically grounded
  - It performs implicit KL regularization
  - It enjoys a very strong performance bound
  - It increases the action gap, asymptotic theoretical quantification
- Munchausen RL works very well
  - M-DQN > C51 > DQN
  - M-IQN > Rainbow > IQN
- More:
  - Paper (Munchausen Reinforcement Learning)
  - Open source code (Google's github)
  - Theoretical analysis relies on our previous work (Leverage the Average)