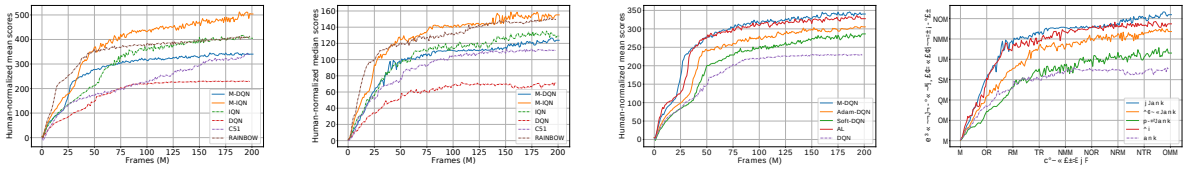

Figure 1: Left to right: all mean, all median, ablation mean, ablation median.

General. Shortly after submission, we found an unintentional inconsistency in our evaluation protocol. Our M-Agents and their ablations were evaluated in a slightly different Atari setting from the others (episode terminated on life loss rather than game over), favoring them on some games. We reran experiments with the correct settings and (human-normalized) results are on Fig. 1. Despite some drop in absolute performance (eg, M-DQN outperforms C51 but is a bit less competitive with Rainbow), it does not change the ranking of algorithms (for all metrics) nor our conclusions.

R1. Q1: In Dopamine, IQN doesn't use double Q-learning (`implicit_quantile_agent.py`, ll. 192-196). **Q2:** Studying the homogeneity of the action gap is an interesting research direction that we didn't investigate, thanks for pointing it out. In the limit (= 1), it is homogeneous (because infinite), but with numerical instabilities. We'll add a comparison with log-DQN in Fig. 2, to get at least some empirical insights. **Q3:** We do not see any obvious connections between Log-RL (related to the *h*-transform of Pohlen *et al*.) and M-RL (justified here through the lens of KL regularization), but that's also an interesting research direction. We also think that, even if different, both approaches could be combined to build an even stronger agent. **Q4:** Yes, exactly, we say why in footnote 6 (and this comparison is done in the ablation). We'll say it earlier. **Q5:** AL is indeed better than C51 on some metrics. Yet, this was not observed in [6], probably because the authors use RMS (to compare with the standard DQN) while we use Adam (to compare with M-DQN). Additionally, note that AL is a special case of M-DQN. We will rephrase appropriately.

R2. Parameters: To choose the parameters, we did a sensitivity analysis on a subset of games, that we'll add to the Appx. Yet, note that the parameters were not selected through a simple grid search. Values of are on par with the ones provided by the analysis in [30], and is the same as the one found optimal in [6]. In terms of "easiness to tune", our empirical findings suggest that the most sensitive parameter is , while it is rather easy to find working values for and I_0. **ALE:** our results are indeed not comparable with the original DQN ones, as discussed ll.217-231. Unfortunately, the authors used a proprietary version of ALE, so exact comparison to their results is not possible. **FQF:** thanks, we missed this paper. However, FQF does not use sticky actions and comparison is thus not straightforward. Our method readily applies to FQF (as for IQN), and we will try to add M-FQF results in the paper. If it is not possible, we will reformulate to soften the state-of-the-art claim. We believe this does not hinder the relevance of our approach, as M-IQN still outperforms Rainbow. **Action-gap:** by "quantifying", we mean "analytically quantifying": Bellemare et al. show that the action gap increases but not by how much, while we derive an actual value for the increase (Thm. 2).

R4. We trust that R4 has deeply misunderstood our contribution. Indeed, they state that "M-RL applies an entropy regularizer to the reward signal", and all the following (mostly negative) comments rely heavily on this statement. This is just wrong and we start the paper by stating otherwise (l.21-24, " We insist right away that this is different from maxent RL, that *subtracts* the scaled log-policy to *all* rewards [...]" while we *add* it to the *immediate* reward). **On novelty:** We strongly disagree, our contribution is not simply an instance of entropy-reg RL, as notably thoroughly discussed in Sec. 3 and related Appx. M-DQN *is different* from Soft-DQN, one just has to compare Eqs. (2) and (3) (paying attention to the *signs* of the different log terms). We're also absolutely certain that neither our algorithms nor our analysis are covered by Neu et al. **Most theoretical results are previously known.** Again, we strongly disagree. We're very clear about what our contributions are, and Thm. 1 and 2 are new (we could reevaluate this claim if a ref was provided). **Hard to believe** (about the theoretical result): we don't ask to believe our claim, as we provide the proofs, see Appx A.2 and Cor. 1 and 2 in Appx. A.3. Shortly, the implicit KL regularization avoids the error in the greedy step, that cannot be avoided when the KL regularization is explicit (the only case considered in [30]). **Clipping:** Obviously, we do not claim to be the first to clip a log term, but we provide this kind of details for the sake of reproducibility.

R5. Linear case: in this setting M-VI is strictly equivalent to MD-VI (see Thm. 1, noticing that in the linear case there is no error in the greedy step), so we refer to [30] for a study on tabular MDPs. Notice that this equivalence is lost with a non-linear paremeterization (due to the necessary error in the greedy step with explicit KL regularization), hence the interest of the Munchausen principle for *deep* RL. **Comparison to Soft-DQN is missing:** no, it's not, it is provided in Fig. 3 (ablation). **Unsure about AL fairly compared**: we use = 0:9 for both M-DQN and AL, and this choice is consistent with the AL paper, so we think the comparison to be fair. Also, notice that AL is a special case of M-DQN. **Montezuma**: this is a hard exploration game, and our method is not designed to favour exploration (there is no signal to reinforce, a discussion to a related issue is provided in Appx. B.2). **M-DQN** is defined l.31. **Eq. l.82:** this is exactly DQN, as is defined to be greedy wrt q . **Bootstrap:** yes, we totally agree with this, it is actually the reason for the name "Munchausen" (l.18-26).