We thank the reviewers for their comments and suggestions and for finding our theoretical considerations insightful and solid [**R1**, **R3**, **R4**], our experimental results valuable [**R2**, **R3**, **R4**] and our paper well-written [**R1**, **R3**, **R4**].

About multi-step rollouts We appreciate the suggestion of **R1** and **R3** on multi-step rollouts. We are curious as well but we opted to leave it to future work: the goal of the paper is to show that model-based methods can be used to directly learn the action-gradient of the value function, and we intentionally designed the simplest algorithm possible to show this point. Using multi-step methods for computing either the gradient of the TD-target or the value gradient itself is orthogonal to what we propose in MAGE, and would introduce additional complexity and hyperparameters (e.g., rollout length); we believe this would distract the reader from the main contribution of the paper. Nonetheless, we agree that combinations of MAGE and these approaches can be particularly fruitful.

R1: $*$ The setup of MAGE follows common practices in (MB)RL: ensembles of models and the swish activation have been used since PETS [7], and, more recently, in (SotA) MBPO [20], whose training scheme is very similar to ours; the Huber loss has been shown to improve stability in DQN (2015). Those choices are known to be effective, but any reasonable setting is enough for MAGE to perform well (see, e.g., the robustness of MAGE to the choice of λ in Fig. 7 in Appendix B.3 and the additional experiment with different model architectures in Fig. R1). Therefore, we did not try to cherry-pick the settings nor did any overly extensive hyperparameter search.

$*$ Concerning the experiments on the unknown reward function setting, notice that Fig. 6 (in Appendix B.2) contains results for all the tasks. We only included a single environment (Pusher-v2) in the main paper in order to save space.

$*$ We indeed show the asymptotic performance on HalfCheetah-v2 only since our focus is on sample-efficiency; nonetheless, we plan to include a larger number of steps (also for Swimmer-v2) in the final version of the paper.

$*$ We will include the suggested references into the paper. See also **About multi-step rollouts**.

R2: **(1)** The reviewer suggests that the paper should first "show that minimizing the TD-error is not a good choice". Notice, however, that despite being commonly used and thought of as "intuitive", the minimization of the TD-error in DPGs is just a heuristic and lacks proper theoretical justification ever since its first use in [42]. In contrast, our Proposition 3.1 inspires a reliable theoretically-grounded objective for learning a critic. Furthermore, Fig. 3 shows indeed that minimizing the TD-error can lead to a critic being far away from the ideal one.
(2) We find no contradiction in our line of reasoning: we state that, despite the objective suggested by Proposition 3.1 is theoretically sound, there are practical issues that make the optimization problem hard. As written in the paper, we interpret the TD-error as a regularization term with almost no additional overhead, but we do not exclude the existence of other practical strategies for minimizing the bound in Proposition 3.1 without the use of the raw TD-error. We agree that in MAGE the minimization of the TD-error is also important but we clearly showed with our experiments that the *overall* objective is better than the TD-error alone.
(3) We did not write that "model-based RL has no advantage in terms of sample-efficiency than model-free RL". Instead, we said that this advantage is not *intrinsic* and it is instead "highly environment and algorithm-dependent". We find this statement uncontroversial in the light of other recent works in MBRL (see, e.g., [20,52]).

R3: $*$ We carefully positioned our approach in the literature and discussed similarities and differences w.r.t. most of the works listed by R3. In particular, we directly cite Balduzzi et al. (2015) for proposing a solution to the problem of value gradient learning. Notice, however, that MAGE provides a completely different and more direct way to address it. Moreover, in our Related Works section, we acknowledge Werbos who pioneered (in 1977!) value gradient learning, preceding Balduzzi et al's, Fairbank's and Weber et al's work by decades. Having said that, notice that these works lack our theoretical motivation and belong to significantly different algorithmic frameworks. Likewise, Sobolev training and related concepts in value gradients mainly share *technical tools* with MAGE, but they have different algorithmic implications. Therefore, we are truly downhearted by the assertion regarding the alleged limited novelty of MAGE.

$*$ Concerning the gradient stopping, we indeed use it: see Algorithm 1, in which target parameters $\bar{\phi}$ are employed during the computation of $\widehat{\delta}$, a convenient notation for the gradient-stopping operation on the target $r(s,a)+\gamma Q_{\bar{\phi}}(\widehat{s},a')$ and its gradient w.r.t. the action a. Nonetheless, as suggested, we will make the gradient stopping more explicit.

$*$ See **About multi-step rollouts**.

R4: We agree with the reviewer that the quality of the learned model is of paramount importance for most MBRL algorithms, whose performance, generally, deteriorates when the model is not enough expressive for a given task. Thus, as suggested, we performed an additional experiment to investigate how the performance of MAGE (compared to the Dyna-TD3 baseline) is affected by the use of less powerful models. We evaluated two versions of the model with reduced capacity: (i) only 2 members in the ensemble, 4 hidden layers and 256 units per layer (-small suffix); (ii) no ensemble (a single model), 2 hidden layers, 256 units per layer (-smaller suffix). Recall that the original setting involves a more powerful model with 8 members in the ensemble, 4 hidden layers, 512 units per layer (no suffix). The results, reported in Fig. R1, show that MAGE is robust to the presence of a misspecified model: while a simpler but still quite capable model does no harm to MAGE, a significantly smaller model has a reasonable impact on the obtained average return. We will report results for all environments in the final version.

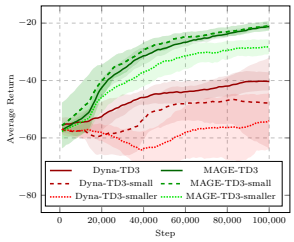

Figure R1: Pusher-v2 (5 runs, 95% c.i.).