A. DICE derivations

In (Foerster et al., 2018), the surrogate function is given by:

$$J^{\text{DICE}} = \sum_{t=0}^{H-1} \left(\prod_{t'=0}^{t} \frac{\pi_\theta(a_{t'}|s_{t'})}{\perp(\pi_\theta(a_{t'}|s_{t'}))} \right) r(s_t, a_t), \tag{1}$$

where $\perp$ denotes the 'stop gradient' operation. In expectation, the Monte Carlo estimate of the Hessian $\mathbb{E}_{\boldsymbol{\tau} \sim P_{\mathcal{T}}(\boldsymbol{\tau}|\theta)}[\nabla^2_\theta J^{\text{DICE}}]$ is equivalent to the Hessian of the inner objective:

$$\nabla_\theta J^{\text{DICE}}$$

$$= \sum_{t=0}^{H-1} \nabla_\theta \left(\prod_{t'=0}^{t} \frac{\pi_\theta(a_{t'}|s_{t'})}{\perp(\pi_\theta(a_{t'}|s_{t'}))} \right) r(s_t, a_t)$$

$$= \sum_{t=0}^{H-1} \left(\prod_{t'=0}^{t} \frac{\pi_\theta(a_{t'}|s_{t'})}{\perp(\pi_\theta(a_{t'}|s_{t'}))} \right) \left(\sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \right)$$

$$r(s_t, a_t) \rightarrow \sum_{t=0}^{H-1} \left(\sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \right) r(s_t, a_t) \tag{2}$$

Once again we take the derivative in order to obtain the Hessian of J^{DICE}:

$$\nabla^2_\theta J^{\text{DICE}} \tag{3}$$

$$= \sum_{t=0}^{H-1} \nabla_\theta \left(\prod_{t'=0}^{t} \frac{\pi_\theta(a_{t'}|s_{t'})}{\perp(\pi_\theta(a_{t'}|s_{t'}))} \right) \left(\sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \right) \tag{4}$$

$$r(s_t, a_t) + \left(\prod_{t'=0}^{t} \frac{\pi_\theta(a_{t'}|s_{t'})}{\perp(\pi_\theta(a_{t'}|s_{t'}))} \right) \tag{5}$$

$$\nabla_\theta \left(\sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \right) r(s_t, a_t) \tag{6}$$

$$\rightarrow \sum_{t=0}^{H-1} \left(\sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \right) \left(\sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \right)^\top \tag{7}$$

$$r(s_t, a_t) + \left(\sum_{t'=0}^{t} \nabla^2_\theta \log \pi_\theta(a_{t'}|s_{t'}) \right) r(s_t, a_t) \tag{8}$$

It is easy to show that

$$\mathbb{E}_{\boldsymbol{\tau} \sim P_{\mathcal{T}}(\boldsymbol{\tau}|\theta)}[\nabla^2_\theta J^{\text{DICE}}] \tag{9}$$

$$= \mathbb{E}_{\boldsymbol{\tau} \sim P_{\mathcal{T}}(\boldsymbol{\tau}|\theta)} \Bigg[\sum_{t=0}^{H-1} \left(\sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \right) \tag{10}$$

$$\left(\sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \right)^\top r(s_t, a_t) \tag{11}$$

$$+ \left(\sum_{t'=0}^{t} \nabla^2_\theta \log \pi_\theta(a_{t'}|s_{t'}) \right) r(s_t, a_t) \Bigg] \tag{12}$$

$$= \nabla^2_\theta J_{\text{inner}} \tag{13}$$

B. LVC derivations

LVC (Rothfuss et al., 2019) is a method that trade-off bias for lower variance estimation of gradient.

$$\nabla^2_\theta J^{\text{LVC}} \tag{14}$$

$$= \nabla_\theta \sum_{t=0}^{H-1} \frac{\pi_\theta(a_t|s_t)}{\perp(\pi_\theta(a_t|s_t))} \nabla_\theta \log \pi_\theta(a_t|s_t) \left(\sum_{t'=t}^{H-1} r(s_{t'}, a_{t'}) \right) \tag{15}$$

$$= \sum_{t=0}^{H-1} \frac{\pi_\theta(a_t|s_t)}{\perp(\pi_\theta(a_t|s_t))} \nabla_\theta \log \pi_\theta(a_t|s_t) \nabla_\theta \log \pi_\theta(a_t|s_t)^\top \tag{16}$$

$$\left(\sum_{t'=t}^{H-1} r(s_{t'}, a_{t'}) \right) + \frac{\pi_\theta(a_t|s_t)}{\perp(\pi_\theta(a_t|s_t))} \nabla^2_\theta \log \pi_\theta(a_t|s_t) \tag{17}$$

$$\left(\sum_{t'=t}^{H-1} r(s_{t'}, a_{t'}) \right) \tag{18}$$

$$\rightarrow \sum_{t=0}^{H-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \nabla_\theta \log \pi_\theta(a_t|s_t)^\top \tag{19}$$

$$\left(\sum_{t'=t}^{H-1} r(s_{t'}, a_{t'}) \right) + \nabla^2_\theta \log \pi_\theta(a_t|s_t) \left(\sum_{t'=t}^{H-1} r(s_{t'}, a_{t'}) \right) \tag{20}$$

$$= \sum_{t=0}^{H-1} \left(\sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \nabla_\theta \log \pi_\theta(a_t|s_t)^\top \right) r(s_t, a_t) \tag{21}$$

$$+ \left(\sum_{t'=0}^{t} \nabla^2_\theta \log \pi_\theta(a_{t'}|s_{t'}) \right) r(s_t, a_t) \tag{22}$$

It is can be prove that $\mathbb{E}_{\boldsymbol{\tau} \sim P_{\mathcal{T}}(\boldsymbol{\tau}|\theta)} \left[J^{\text{LVC}} \right]$ is a biased approximation of the $\nabla^2_\theta J_{\text{inner}}(\theta)$ but with lower variance, see (Rothfuss et al., 2019) for more details.

$$\mathbb{E}_{\boldsymbol{\tau} \sim P_{\mathcal{T}}(\boldsymbol{\tau}|\theta)} \left[J^{\text{LVC}} \right] \tag{23}$$

$$= \mathbb{E}_{\boldsymbol{\tau} \sim P_{\mathcal{T}}(\boldsymbol{\tau}|\theta)} \Bigg[\sum_{t=0}^{H-1} \Big(\sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(\boldsymbol{a}_{t'}|\boldsymbol{s}_{t'}) \tag{24}$$

$$\nabla_\theta \log \pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t)^\top \Big) r(\boldsymbol{s}_t, \boldsymbol{a}_t) \Bigg] + \tag{25}$$

$$\mathbb{E}_{\boldsymbol{\tau} \sim P_{\mathcal{T}}(\boldsymbol{\tau}|\theta)} \left[\sum_{t=0}^{H-1} \left(\sum_{t'=0}^{t} \nabla^2_\theta \log \pi_\theta(\boldsymbol{a}_{t'}|\boldsymbol{s}_{t'}) \right) r(\boldsymbol{s}_t, \boldsymbol{a}_t) \right] \tag{26}$$

$$\approx \nabla^2_\theta J_{\text{inner}} \tag{27}$$

C. Proof of Theorem 1.

Firstly, we derive first order gradient of surrogate objective function.

$$\nabla_\theta J_i^{\text{TMAML}} \tag{28}$$

$$=\nabla_\theta \sum_{t=0}^{H-1}\left[1-\left(\prod_{t'=0}^{t}\frac{\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i)}{\perp(\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i))}\right)\right]$$

$$\left(1-\frac{\pi_\theta(a_t|s_t,\mathcal{T}_i)}{\perp(\pi_\theta(a_t|s_t,\mathcal{T}_i))}\right)b(s_t,\mathcal{T}_i) \tag{29}$$

$$=\sum_{t=0}^{H-1} b(s_t,\mathcal{T}_i)\Bigg[-\frac{\pi_\theta(a_t|s_t,\mathcal{T}_i)}{\perp(\pi_\theta(a_t|s_t,\mathcal{T}_i))} \tag{30}$$

$$\nabla_\theta \log\pi_\theta(a_t|s_t,\mathcal{T}_i)\left[1-\left(\prod_{t'=0}^{t}\frac{\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i)}{\perp(\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i))}\right)\right] \tag{31}$$

$$-\left(1-\frac{\pi_\theta(a_t|s_t,\mathcal{T}_i)}{\perp(\pi_\theta(a_t|s_t,\mathcal{T}_i))}\right)\left(\prod_{t'=0}^{t}\frac{\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i)}{\perp(\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i))}\right) \tag{32}$$

$$\left(\sum_{t'=0}^{t}\nabla_\theta\log\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i)\right)\Bigg] \tag{33}$$

Once again we take the derivative of Equation (33) in order to obtain the Hessian of J_i^{TMAML},

$$\nabla_\theta^2 J_i^{\text{TMAML}} \tag{34}$$

$$=\sum_{t=0}^{H-1} b(s_t,\mathcal{T}_i)\Bigg[-\left(1-\prod_{t'=0}^{t}\frac{\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i)}{\perp(\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i))}\right) \tag{35}$$

$$\nabla_\theta^2\frac{\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i)}{\perp(\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i))}-\left(1-\frac{\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i)}{\perp(\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i))}\right) \tag{36}$$

$$\nabla_\theta^2\prod_{t'=0}^{t}\frac{\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i)}{\perp(\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i))} \tag{37}$$

$$+2\frac{\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i)}{\perp(\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i))}\nabla_\theta\log\pi(a_t|s_t,\mathcal{T}_i) \tag{38}$$

$$\prod_{t'=0}^{t}\frac{\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i)}{\perp(\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i))}\Bigg] \tag{39}$$

$$\to 2\sum_{t=0}^{H-1}\nabla_\theta\log\pi(a_t|s_t,\mathcal{T}_i) \tag{40}$$

$$\left[\sum_{t'=t}^{H-1}\nabla_\theta\log\pi(a_{t'}|s_{t'},\mathcal{T}_i)b(s_{t'},\mathcal{T}_i)\right] \tag{41}$$

In expectation, $\mathbb{E}_{\boldsymbol{\tau}\sim P_\mathcal{T}(\boldsymbol{\tau}|\theta)}[\nabla_\theta^2 J_i^{\text{TMAML}}]$ is zero, to show this, we use the form of $\nabla_\theta^2 J_i^{\text{TMAML}}$ just derived, for each t, we have

$$\mathbb{E}_{\boldsymbol{\tau}\sim P_\mathcal{T}(\boldsymbol{\tau}|\theta)}\Bigg[\nabla_\theta\log\pi(a_t|s_t)\Bigg[\sum_{t'=t+1}^{H-1}b(s_{t'},\mathcal{T}_i)$$

$$\nabla_\theta\log\pi(a_{t'}|s_{t'},\mathcal{T}_i)\Bigg]\Bigg]=\mathbb{E}_{\boldsymbol{\tau}\sim P_\mathcal{T}(\boldsymbol{\tau}|\theta)}\Bigg[\nabla_\theta\log\pi(a_t|s_t,\mathcal{T}_i)$$

$$\left[\sum_{t'=t+1}^{H-1}b(s_{t'},\mathcal{T}_i)\mathbb{E}_{a_{t'},s_{t'}|a_t,s_t,\mathcal{T}_i}\left[\nabla_\theta\log\pi(a_{t'}|s_{t'},\mathcal{T}_i)\right]\right]\Bigg]$$

$=0$ (score function expectation equals 0) ,

Base on this, sum over all time step t will get the expectation $\nabla_\theta^2 J_i^{\text{TMAML}}$ equals 0,

$$\mathbb{E}_{\boldsymbol{\tau}\sim P_\mathcal{T}(\boldsymbol{\tau}|\theta)}[\nabla_\theta^2 J_i^{\text{TMAML}}]=0$$

D. Proof of Theorem 2.

Here we show how J_i^{TMAML} can introduce control variate into the Hessian estimation. The intuition is getting a sum of reward in the Hessian formula instead of the sum of gradient which is impossible to introduce control variates. In order to do this, we first decompose the expectation of J_i^{DICE} into the desired form.

$$\nabla_\theta^2\left(J_i^{\text{DICE}}+J_i^{\text{TMAML}}\right) \tag{42}$$

$$=\nabla_\theta^2 J_i^{\text{TMAML}}+ \tag{43}$$

$$\sum_{t=0}^{H-1}\nabla_\theta\left(\prod_{t'=0}^{t}\frac{\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i)}{\perp(\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i))}\right)r(s_t,a_t,\mathcal{T}_i) \tag{44}$$

$$=\nabla_\theta^2 J_i^{\text{TMAML}}+ \tag{45}$$

$$\sum_{t=0}^{H-1}\left(\prod_{t'=0}^{t}\frac{\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i)}{\perp(\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i))}\right)r(s_t,a_t,\mathcal{T}_i) \tag{46}$$

$$\Bigg[\left(\sum_{t'=t}^{H-1}\nabla_\theta\log\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i)\right)^2 \tag{47}$$

$$+\sum_{t'=t}^{H-1}\nabla_\theta^2\log\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i)\Bigg] \tag{48}$$

$$\tag{49}$$

Then we substitute $\nabla_\theta^2 J_i^{\text{TMAML}}$ with Equation (34), we can derive,

$$\nabla_\theta^2\left(J_i^{\text{DICE}}+J_i^{\text{TMAML}}\right)$$

$$\to\sum_{t=0}^{H-1}\left[\nabla_\theta^2\log\pi_\theta(a_t|s_t,\mathcal{T}_i)\left(\sum_{t'=0}^{t}r(a_{t'},s_{t'},\mathcal{T}_i)\right)\right]$$

$$+2\sum_{t=0}^{H-1}\Bigg[\nabla\log\pi_\theta(a_t|s_t,\mathcal{T}_i)^\top\left(\sum_{t'=t}^{H-1}\nabla\log\pi_\theta(a_{t'}|s_{t'},\mathcal{T}_i)\right.$$

$$\left.\left(\sum_{k=t'}^{H-1}r(a_k,s_k)-b(s_t,\mathcal{T}_i)\right)\right)\Bigg],$$

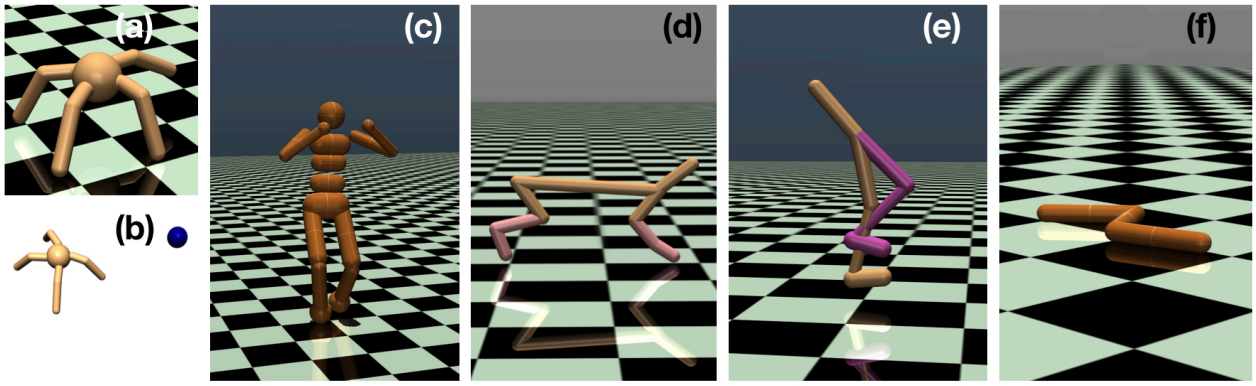

Figure 1. Meta-RL environments based on Mujoco (Todorov et al., 2012) and OpenAI Gym (Brockman et al., 2016): *(a)* Ant chasing random goal location, the tasks are generated by sampling the target positions from the uniform distribution on $[-3, 3]^2$.; *(b)* An example of Ant chasing goal task, where green ball denotes a goal, the goal is randomly sampled at the beginning of each episode; *(c)* Humanoid; *(d)* HalfCheetah; *(e)* Walker; *(f)* Swimmer; We build different tasks including random velocity, random direction, and random goal location on each environment, see experiments in main paper for more details. In order to get high reward in these environments, agents must be able to do learning to learning or meta-learning from past experience.

Hyperparameters	
Policy Hidden Layer Sizes	64^2 (Humanoid: 128^2)
Num Adapt Steps	1 (Ant, rand goal: 2)
Inner Step Size α	0.01
Outer LR β	0.001
Meta CV Inner Step Size α'	0.01
Meta CV Outer LR β'	0.01
Meta-Tasks Per Iteration	40
Num Traj Per Meta-Task	20
Grad Steps Per Iteration	5
Outer Clip Ratio ϵ	0.3
KL Penalty Coefficient η	0.0005
Traj Length	100

Table 1. Hyperparameter settings used in each algorithm

E. Hyperparameters and Environments Setup

In every experiment, each algorithm is tested with 5 random seeds. An example of tested environment is shown in Figure 1. Table 1 contains the hyperparameter settings used for the different algorithms.

References

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Foerster, J., Farquhar, G., Al-Shedivat, M., Rocktäschel, T., Xing, E., and Whiteson, S. DiCE: The infinitely differentiable Monte Carlo estimator. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1529–1538, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Rothfuss, J., Lee, D., Clavera, I., Asfour, T., and Abbeel, P. ProMP: Proximal meta-policy search. In *International Conference on Learning Representations*, 2019.

Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pp. 5026–5033. IEEE, 2012.