

# Exploration and uncertainty

## A3.1: Implement and test actor-critics

(5 points)

In this exercise you will implement the vanilla actor-critic algorithm, an off-policy extension and finally PPO. Your implementations will be based on a REINFORCE implementation from the Jupyter notebook `ac.ipynb`, that can be easily extended. Please don't forget that *all* of your answers have to be submitted in *one* PDF file, including this one.

- (a) [1 point] Read carefully through the provided code (only few classes have changed from exercise sheet 2) and run the given REINFORCE algorithm on the `Cartpole-v1` environment for 2 million (2M) steps. You can stop episodes after 200 steps. This can take 10-20 minutes.
- (b) [1 point] Extend the `ReinforceLearner` class in the given Jupyter Notebook with a value function as bias (see slide 6 of Lecture 6). Implement two target-definitions for the value function, selected by the `'value_targets'` parameter: `'returns'` uses the returns  $R_t$  that are stored in the mini-batch, whereas `'td'` uses the TD-error. Make sure that the bias is ignored when the given parameter `'advantage_bias'` is **False**. Test your implementation as above on the environment `Cartpole-v1` for 2M steps with the default parameters (using `'returns'` value targets).

*Hint:* The first heads of the model are interpreted as logits of a softmax policy, and the last head of the model is interpreted as the value function.

- (c) [1 point] Extend the `BiasedReinforceLearner` class in your implementation with an advantage function that uses bootstrapping (replaces  $R_t$  with  $r_t + \gamma v_\phi(s_{t+1})$ , see slide 6 of Lecture 6). Make sure the original behavior is maintained when the parameter `advantage_bootstrap` is **False**. Test your implementation as above on the environment `Cartpole-v1` for 2M steps with the default parameters.
- (d) [1 point] Run your `ActorCriticLearner` with 80 off-policy iterations (by setting the parameter `params['offpolicy_iterations'] = 80`). Extend your implementation to the class `OffpolicyActorCriticLearner`, that uses on-policy gradients in the first and off-policy gradients in all following iterations ( $\mathcal{L}_\mu$  on slide 15 of Lecture 6). Test your implementation as above on the environment `Cartpole-v1` with the default parameters, but only for 500k steps.

*Hint:* `ReinforceLearner` has an attribute `old_pi` which is set to **None** at the beginning of `train()`. You can save the on-policy probabilities of the initial policy here to use them in the ratios of the off-policy loss.

- (e) [1 point] Now extend `OffpolicyActorCriticLearner` by adding PPO clipping to the off-policy loss ( $\mathcal{L}_\mu^{\text{clip}}$  on slide 18 of Lecture 6). Test your implementation as above on the environment `Cartpole-v1` with the default parameters, but only for 500k steps.

**Solution** The solution can be found in the Jupyter notebook `ac_solution.ipynb`

**A3.2: Policy optimization under constraints****(3 points)**

This exercise assumes a multi-armed bandit, where the player is choosing between arms  $a \in \mathcal{A} \subset \mathbb{N}$  to get a randomized reward  $r \in \mathbb{R}$  drawn from  $p(r|a)$ . Let  $Q(a) := \mathbb{E}[r | r \sim p(\cdot|a)]$  denote the expected reward of action  $a \in \mathcal{A}$ . Prove analytically that the solution to the optimization problem

$$\max_{\pi} \mathbb{E}[r | r \sim p(\cdot|a), a \sim \pi(a)] \quad \text{s.t.} \quad \mathcal{H}(\pi) \geq \epsilon,$$

where  $\pi$  is a probability distribution over  $\mathcal{A}$  and  $\mathcal{H}(\pi) = -\sum_a \pi(a) \ln \pi(a)$  the corresponding entropy, is the softmax of the expected reward (with some inverse temperature  $\beta$ ):

$$\pi(a) = \frac{\exp(\beta Q(a))}{\sum_{a'} \exp(\beta Q(a'))}, \quad \forall a \in \mathcal{A}.$$

*Bonus-question:* While there exists no analytical solution for the inverse temperature  $\beta$ , can you derive an iterative approach that improves the optimization problem for a given  $\beta$  and target entropy  $\epsilon$ ?

*Hint:* Try to formulate the above optimization problem with the method of Lagrangian multipliers, which you can read up on at [https://en.wikipedia.org/wiki/Lagrange\\_multiplier](https://en.wikipedia.org/wiki/Lagrange_multiplier). You may ignore the implicit constraints  $\pi(a) \geq 0, \forall a \in \mathcal{A}$ , at first and check whether the solution adheres to them later.

**Solution** That  $\pi(a)$  is a probability distribution over  $\mathcal{A}$  induces the additional constraints  $\pi(a) \geq 0, \forall a \in \mathcal{A}$ , and  $\sum_a \pi(a) = 1$ , which yields the Lagrange function (ignoring  $\pi(a) \geq 0, \forall a \in \mathcal{A}$ , for now):

$$L(\pi, \lambda, \eta) = \sum_a \pi(a) Q(a) + \lambda(\mathcal{H}(\pi) - \epsilon) + \eta \left( \sum_a \pi(a) - 1 \right), \quad \lambda \geq 0.$$

Note that the entropy of  $\pi$  is defined as  $\mathcal{H}(\pi) = -\sum_a \pi(a) \ln \pi(a)$ . Setting the derivative of  $L$  (w.r.t. the probability  $\pi(a)$  to choose action  $a \in \mathcal{A}$ ) to zero yields with the substitution  $\beta := \frac{1}{\lambda}$  the solution

$$\frac{\partial}{\partial \pi(a)} L(\pi) = Q(a) - \lambda \ln \pi(a) - \lambda + \eta \stackrel{!}{=} 0 \quad \Leftrightarrow \quad \pi(a) = \frac{\exp(\beta Q(a))}{\exp(\frac{\eta}{\lambda} - 1)}.$$

This always fulfills  $\pi(a) \geq 0, \forall a \in \mathcal{A}$ . Furthermore, substituting this into the derivative w.r.t.  $\eta$  yields:

$$\frac{\partial}{\partial \eta} L(\eta) = \sum_a \pi(a) - 1 \stackrel{!}{=} 0 \quad \Leftrightarrow \quad \sum_a \pi(a) = 1 \quad \Leftrightarrow \quad \exp(\frac{\eta}{\lambda} - 1) = \left( \sum_a \exp(\beta Q(a)) \right)^{-1}.$$

Substituting this back into the analytical solution for  $\pi(a)$  yields the softmax of the scaled value:

$$\pi(a) = \frac{\exp(\beta Q(a))}{\sum_{a'} \exp(\beta Q(a'))}.$$

*Bonus-question:* Lastly, setting the derivative of  $L$  w.r.t.  $\lambda$  to zero yields

$$\epsilon = -\sum_a \pi(a) \ln \pi(a) = -\beta \sum_a \pi(a) Q(a) + \ln \left( \sum_{a'} \exp(\beta Q(a')) \right).$$

Obviously there exist no analytical solution for  $\beta$ , but one can derive the iterative update:

$$\beta \leftarrow \frac{\ln \left( \sum_{a'} \exp(\beta Q(a')) \right) - \epsilon}{\sum_a \pi(a) Q(a)}, \quad \text{if } \sum_a \pi(a) Q(a) \neq 0.$$

**A3.3: TD(1) is Monte-Carlo sampling****(3 points)**

Let  $\{s_t, a_t, r_t\}_{t=0}^{\infty}$  denote an infinite trajectory sampled from some Markov chain. Prove analytically that in the limit  $\lambda \rightarrow 1$  the TD( $\lambda$ ) targets  $V_{\lambda}^{\pi}(s)$  are identical to the Monte-Carlo targets  $V_{MC}^{\pi}(s)$ .

$$V_{\lambda}^{\pi}(s_0) = (1-\lambda) \sum_{n=0}^{\infty} \lambda^n V_{n+1}^{\pi}(s_0), \quad V_n^{\pi}(s_0) = \sum_{t=0}^{n-1} \gamma^t r_t + \gamma^n V^{\pi}(s_n), \quad V_{MC}^{\pi}(s_0) = \lim_{n \rightarrow \infty} V_n^{\pi}(s_0).$$

*Hint:* You can assume that  $\lim_{\lambda \rightarrow 1} \lim_{m \rightarrow \infty} \lambda^m x_m = 0$ , for any bounded sequence  $-\infty < x_m < \infty, \forall m \geq 0$ .

*Hint:* You can use the identity  $\sum_{i=0}^m \sum_{j=0}^i A_{ij} = \sum_{j=0}^m \sum_{i=j}^m A_{ij}$ , which can be proven by rearranging summands.

**Solution** Equation (1) uses the second hint, (2) uses the geometric series and (3) the first hint.

$$\begin{aligned} V_{\lambda}^{\pi}(s_0) &= (1-\lambda) \sum_{n=0}^{\infty} \lambda^n V_{n+1}^{\pi}(s_0) = (1-\lambda) \sum_{n=0}^{\infty} \lambda^n \left( \sum_{t=0}^n \gamma^t r_t + \gamma^{n+1} V^{\pi}(s_{n+1}) \right) \\ &= (1-\lambda) \gamma \sum_{n=0}^{\infty} (\lambda \gamma)^n V^{\pi}(s_{n+1}) + (1-\lambda) \lim_{m \rightarrow \infty} \sum_{n=0}^m \sum_{t=0}^n \lambda^n \gamma^t r_t \\ &\stackrel{(1)}{=} (1-\lambda) \gamma \sum_{n=0}^{\infty} (\lambda \gamma)^n V^{\pi}(s_{n+1}) + (1-\lambda) \lim_{m \rightarrow \infty} \sum_{t=0}^m \sum_{n=t}^m \lambda^n \gamma^t r_t \\ &= (1-\lambda) \gamma \sum_{n=0}^{\infty} (\lambda \gamma)^n V^{\pi}(s_{n+1}) + (1-\lambda) \lim_{m \rightarrow \infty} \sum_{t=0}^m \gamma^t r_t \lambda^t \sum_{n=0}^{m-t} \lambda^n \\ &\stackrel{(2)}{=} (1-\lambda) \gamma \sum_{n=0}^{\infty} (\lambda \gamma)^n V^{\pi}(s_{n+1}) + (1-\lambda) \lim_{m \rightarrow \infty} \sum_{t=0}^m (\lambda \gamma)^t r_t \frac{1-\lambda^{m-t+1}}{1-\lambda} \\ &= (1-\lambda) \gamma \sum_{n=0}^{\infty} (\lambda \gamma)^n V^{\pi}(s_{n+1}) + \sum_{t=0}^{\infty} (\lambda \gamma)^t r_t - \lim_{m \rightarrow \infty} \lambda^{m+1} \sum_{t=0}^m \gamma^t r_t \\ \lim_{\lambda \rightarrow 1} V_{\lambda}^{\pi}(s_0) &= \lim_{\lambda \rightarrow 1} \sum_{t=0}^{\infty} \underbrace{(\lambda \gamma)^t}_{\rightarrow \gamma^t} \left( r_t + \gamma \underbrace{(1-\lambda)}_{\rightarrow 0} V^{\pi}(s_{t+1}) \right) - \overbrace{\lim_{\lambda \rightarrow 1} \lim_{m \rightarrow \infty} \lambda^{m+1} \sum_{t=0}^m \gamma^t r_t}^0 \\ &\stackrel{(3)}{=} \sum_{t=0}^{\infty} \gamma^t r_t = \lim_{n \rightarrow \infty} V_n^{\pi}(s_0) = V_{MC}^{\pi}(s_0) \end{aligned}$$

$-\infty < V_{MC}^{\pi}(s_0) < \infty$

**A3.4: Random exploration decays exponentially****(3 points)**

In this exercise we will show how  $\epsilon$ -greedy exploration yields exponentially decaying exploration, which can prevent an RL agent from learning some MDPs in expectation. For simplicity we look at an infinite state-chain  $\mathcal{S} = \mathbb{Z}$ , where we always start in  $s_0 = 0$  and have four actions  $\mathcal{A} = \{L, R, U, D\}$ . The transition model is  $P(s-1|s, L) = P(s|s, U) = P(s|s, D) = P(s+1|s, R) = 1$ . We will explore the MDP by executing a uniform random policy  $\pi(a|s) = \frac{1}{|\mathcal{A}|}, \forall a \in \mathcal{A}, \forall s \in \mathcal{S}$ .

(a) [1 point] Derive the recursive update equation for the state distribution after  $t$  time steps  $\xi_t^{\pi}(s)$ .

*Hint:* think about from which states one can transition into  $s$  and define  $\xi_{t+1}^{\pi}(s)$  in terms of  $\xi_t^{\pi}(s)$ .

(b) [1 point] Prove analytically that  $\xi_t^{\pi}(s) = \frac{1}{4^t} \binom{2t}{t+s}, \forall |s| \leq t$ , otherwise  $\xi_t^{\pi}(s) = 0$ .

*Hint:*  $\binom{n}{k} = \frac{n!}{k!(n-k)!}$  are binomial coefficients.<sup>1</sup> You can use the identity  $\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$ .

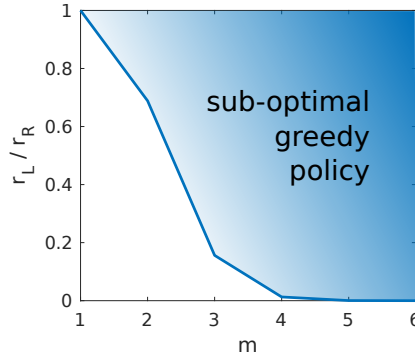
<sup>1</sup>see [https://en.wikipedia.org/wiki/Binomial\\_coefficient](https://en.wikipedia.org/wiki/Binomial_coefficient) for more information and identities.

- (c) [1 point] The above  $\xi_t^\pi(s)$  converges approximately<sup>2</sup> in the limit  $t \rightarrow \infty$  to the normal distribution  $\xi_\infty^\pi(s) = \mathcal{N}(s|0, 1)$ . Let in the following  $\bar{Q}^\pi(s, a) := \mathbb{E}_\pi[\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=0}^{n-1} r_t | \frac{s_0=s}{a_0=a}]$  denote the expected average reward conditioned on state and action. Reaching state  $s = -1$  is rewarded by  $r_L > 0$  and reaching the state  $s = m$  is rewarded by  $r_R > r_L$ . Prove analytically that

$$\bar{Q}^\pi(0, L) > \bar{Q}^\pi(0, R) \quad \Leftrightarrow \quad r_L > r_R \sqrt{2\pi e} \frac{\sinh m}{\sinh 1} \xi_\infty^\pi(m),$$

that is, that the greedy policy based on  $\bar{Q}^\pi$  selects a suboptimal action if  $m$  is too large.

*Hint:* The hyperbolic sine is defined as  $\sinh(m) = \frac{1}{2}(e^m - e^{-m})$ .



### Solution

- (a) You can transition into  $s$  in 4 ways: from  $P(s|s-1, R)$ ,  $P(s|s, U)$ ,  $P(s|s, D)$  and  $P(s, s+1, L)$ .

$$\xi_0^\pi(0) := 1, \quad \xi_{t+1}^\pi(s) := \frac{1}{4}\xi_t^\pi(s-1) + \frac{2}{4}\xi_t^\pi(s) + \frac{1}{4}\xi_t^\pi(s+1), \quad \forall |s| \leq t+1.$$

For  $|s| > t+1$ ,  $\xi_t(s) := 0$  as there is no way to reach those states yet.

- (b) The beginning of the induction for  $t = 0$  is easy:  $\xi_0^\pi(0) \stackrel{\text{ind}}{=} \frac{1}{4^0} \binom{0}{0} =: 1$ . ✓

$$\begin{aligned} \xi_{t+1}^\pi(s) &:= \frac{1}{4}\xi_t^\pi(s-1) + \frac{2}{4}\xi_t^\pi(s) + \frac{1}{4}\xi_t^\pi(s+1) \\ &\stackrel{\text{ind}}{=} \frac{1}{4^{t+1}} \left( \underbrace{\binom{2t}{t+s-1} + \binom{2t}{t+s}}_{\binom{2t+1}{t+s}} + \underbrace{\binom{2t}{t+s} + \binom{2t}{t+s+1}}_{\binom{2t+1}{t+s+1}} \right) \\ &= \frac{1}{4^{t+1}} \binom{2(t+1)}{(t+1)+s} \quad \checkmark \end{aligned}$$

There are two cases that might be problematic:  $|s| = t$  and  $|s| = t+1$ , as  $\xi_t^\pi(s) = 0, \forall s > t$ .

$$\xi_{t+1}^\pi(t) := \frac{1}{4}\xi_t^\pi(t-1) + \frac{2}{4}\xi_t^\pi(t) = \frac{1}{4^{t+1}} \left( \underbrace{\binom{2t}{2t-1} + \binom{2t}{2t}}_{\binom{2t+1}{2t}} + \underbrace{\binom{2t}{2t}}_{1 = \binom{2t+1}{2t+1}} \right) = \frac{1}{4^{t+1}} \binom{2(t+1)}{(t+1)+t} \quad \checkmark$$

$$\xi_{t+1}^\pi(t+1) := \frac{1}{4}\xi_t^\pi(t) = \frac{1}{4^{t+1}} \binom{2t}{2t} = \frac{1}{4^{t+1}} \binom{2t+2}{2t+2} = \frac{1}{4^{t+1}} \binom{2(t+1)}{(t+1)+(t+1)}. \quad \checkmark$$

If we want to make sure that  $\xi_t^\pi$  is a probability distribution, we could also check whether it sums to 1 (note that  $\sum_{k=0}^n \binom{n}{k} = 2^n$ ):

$$\sum_{s \in \mathbb{Z}} \xi_t^\pi(s) = \frac{1}{4^t} \sum_{s=-t}^t \binom{2t}{t+s} = \frac{1}{4^t} \sum_{k=0}^{2t} \binom{2t}{k} = \frac{1}{4^t} 2^{2t} = 1. \quad \checkmark$$

<sup>2</sup>Proving this is very involved, see <http://www.math.uchicago.edu/~lawler/reu1> for a similar derivation.

- (c) The steady state distribution is now  $\xi_\infty^\pi(s) = \mathcal{N}(s|0, 1) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}s^2)$ . Note that the state distribution after executing  $a = L$  is  $\xi_\infty^\pi(s + 1)$  and after  $a = R$  is  $\xi_\infty^\pi(s - 1)$ .

$$\begin{aligned}
\bar{Q}^\pi(0, L) &> \bar{Q}^\pi(0, R) \\
r_L \xi_\infty^\pi(0) + r_R \xi_\infty^\pi(m + 1) &> r_L \xi_\infty^\pi(-2) + r_R \xi_\infty^\pi(m - 1) \\
r_L (\exp(0) - \exp(-\frac{1}{2}2^2)) &> r_R (\exp(-\frac{1}{2}(m - 1)^2) - \exp(-\frac{1}{2}(m + 1)^2)) \\
r_L (1 - e^{-2}) &> r_R (\exp(-\frac{1}{2}m^2 + m - \frac{1}{2}) - \exp(-\frac{1}{2}m^2 - m - \frac{1}{2})) \\
r_L e^{-1} (\underbrace{e^1 - e^{-1}}_{2 \sinh(1)}) &> r_R \exp(-\frac{1}{2}m^2) e^{-\frac{1}{2}} (\underbrace{e^m - e^{-m}}_{2 \sinh(m)}) \\
r_L &> r_R \sqrt{2\pi} e^{\frac{\sinh m}{\sinh 1}} \xi_\infty^\pi(m)
\end{aligned}$$

### A3.5: Epistemic uncertainty

(3 points)

There is no general consensus on how epistemic uncertainty should be defined mathematically, only that it represents the “reducible” uncertainty about our prediction. In this exercise you will derive *one* possible definition. Let in the following  $x \in \mathcal{X}$  denote samples drawn i.i.d. from distribution  $\rho(x)$  and  $y \in \mathbb{R}$  denote labels drawn i.i.d. from  $\mathcal{N}(y|\mu(x), \sigma^2(x))$ .  $\mathbb{E}_z$  and  $\mathbb{V}_z$  refer to the expectation and variance over the variable  $z$ , e.g.  $x$  or  $y$ . We have estimated the function  $f_{\mathcal{D}}(x) \approx \mathbb{E}_y[y|x]$  with some type of *unbiased regression* on the data set  $\mathcal{D} := \{x_t, y_t\}_{t=1}^n \subset \mathcal{X} \times \mathbb{R}$ , which implies  $\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(x)|x] = \mu(x)$ .

- (a) [1 point] Prove that  $\mathbb{E}_z[(g(z) - \alpha)^2] = \mathbb{V}_z[g(z)] + (\mathbb{E}_z[g(z)] - \alpha)^2, \forall g : \mathcal{Z} \rightarrow \mathbb{R}, \forall \alpha \in \mathbb{R}$ .
- (b) [1 point] Prove that the *expected generalization error* for a mean-squared loss of  $f_{\mathcal{D}}$  is

$$\mathbb{E}_{\mathcal{D}} \left[ \underbrace{\mathbb{E}_x [\mathbb{E}_y [(y - f_{\mathcal{D}}(x))^2 | x]]}_{\text{generalization error of } f_{\mathcal{D}}} \right] = \underbrace{\mathbb{E}_x [\sigma^2(x)]}_{\text{aleatoric}} + \underbrace{\mathbb{E}_x [\mathbb{V}_{\mathcal{D}} [f_{\mathcal{D}}(x) | x]]}_{\text{epistemic uncertainty}}.$$

Note that the generalization error splits into an irreducible aleatoric uncertainty and an epistemic uncertainty over what we have inferred from the data. The latter reduces to zero if and when the variance between predictions for different  $\mathcal{D}$  vanishes (which it should in the limit of infinite data).

- (c) [1 point] Prove that the epistemic uncertainty of the *empirical mean*  $f_{\mathcal{D}} = \frac{1}{n} \sum_{t=1}^n y_t$  for the labels of one constant input sample  $x_t = x, \forall t$ , is  $\mathbb{V}_{\mathcal{D}}[f_{\mathcal{D}}(x)|x] = \frac{\sigma^2(x)}{n}$ .

#### Solution

- (a) Remember the definition  $\mathbb{V}_z[g(z)] = \mathbb{E}_z[(g(z) - \mathbb{E}_z[g(z)])^2]$ :

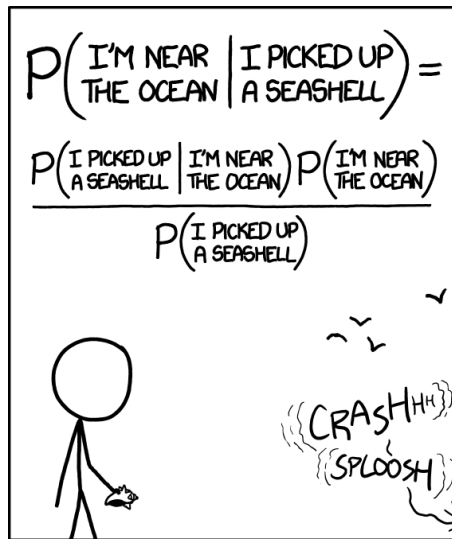
$$\begin{aligned}
\mathbb{E}_z[(g(z) - \alpha)^2] &= \mathbb{E}_z \left[ (g(z) - \mathbb{E}_z[g(z)] + \mathbb{E}_z[g(z)] - \alpha)^2 \right] \\
&= \mathbb{E}_z \left[ (g(z) - \mathbb{E}_z[g(z)])^2 + 2(g(z) - \mathbb{E}_z[g(z)])(\mathbb{E}_z[g(z)] - \alpha) + (\mathbb{E}_z[g(z)] - \alpha)^2 \right] \\
&= \underbrace{\mathbb{E}_z[(g(z) - \mathbb{E}_z[g(z)])^2]}_{\mathbb{V}_z[g(z)]} + \underbrace{2(\mathbb{E}_z[g(z)] - \mathbb{E}_z[g(z)])(\mathbb{E}_z[g(z)] - \alpha)}_0 + (\mathbb{E}_z[g(z)] - \alpha)^2 \\
&= \mathbb{V}_z[g(z)] + (\mathbb{E}_z[g(z)] - \alpha)^2.
\end{aligned}$$

- (b) In this proof we need to use the previous result, swap expectations  $\mathbb{E}_x[\mathbb{E}_z[g(x, z)]] = \mathbb{E}_z[\mathbb{E}_x[g(x, z)]]$  and use the unbiased property  $\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(x)|x] = \mu(x)$ :

$$\begin{aligned}
 \mathbb{E}_{\mathcal{D}} \left[ \mathbb{E}_x \left[ \mathbb{E}_y \left[ (y - f_{\mathcal{D}}(x))^2 | x \right] \right] \right] &\stackrel{(a)}{=} \mathbb{E}_{\mathcal{D}} \left[ \mathbb{E}_x \left[ \overbrace{\mathbb{V}_y[y|x]}^{\sigma^2(x)} + \left( \overbrace{\mathbb{E}_y[y|x]}^{\mu(x)} - f_{\mathcal{D}}(x) \right)^2 \right] \right] \\
 &\stackrel{(a)}{=} \mathbb{E}_x[\sigma^2(x)] + \mathbb{E}_x \left[ \mathbb{V}_{\mathcal{D}}[f_{\mathcal{D}}(x)|x] + \underbrace{\left( \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(x)|x] - \mu(x) \right)^2}_{\mu(x)} \right] \\
 &= \mathbb{E}_x[\sigma^2(x)] + \mathbb{E}_x[\mathbb{V}_{\mathcal{D}}[f_{\mathcal{D}}(x)|x]] .
 \end{aligned}$$

- (c) This is actually a variation of an old exam question. The trick is that  $\mathcal{D}$  is drawn i.i.d., and thus  $\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(x)] = \frac{1}{n} \sum_{t=1}^n \mathbb{E}_{y_t}[y_t|x] = \mu(x)$ ,  $\mathbb{E}_{y_t}[(y_t - \mu(x))^2] = \sigma^2(x)$  and  $\mathbb{E}_{y_i, y_j}[(y_i - \mu(x))(y_j - \mu(x)) - \mu(x)] = 0$  if  $i \neq j$ :

$$\begin{aligned}
 \mathbb{V}_{\mathcal{D}}[f_{\mathcal{D}}(x)] &= \mathbb{E}_{\mathcal{D}} \left[ \left( \frac{1}{n} \sum_{t=1}^n y_t - \mu(x) \right)^2 \right] = \mathbb{E}_{\mathcal{D}} \left[ \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n (y_i - \mu(x))(y_j - \mu(x)) \right] \\
 &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \underbrace{\mathbb{E}_{y_i, y_j} [(y_i - \mu(x))(y_j - \mu(x))]}_{\delta(i=j) \sigma^2(x)} = \frac{\sigma^2(x)}{n} .
 \end{aligned}$$



STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND DON'T HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.

<https://xkcd.com/1236>

**Total 17 points.**