A. UCB Bonus in OEB3

Recall that we consider the following regularized least-square problem,

$$w_t \leftarrow \operatorname*{argmin}_{w\in\mathbb{R}^d} \sum_{\tau=0}^{m} \big[r_t(s_t^\tau, a_t^\tau) + \max_{a\in\mathcal{A}} Q_{t+1}(s_{t+1}^\tau, a) - w^\top \phi(s_t^\tau, a_t^\tau)\big]^2 + \lambda\|w\|^2. \tag{5}$$

In the sequel, we consider a Bayesian linear regression perspective of (5) that captures the intuition behind the UCB-bonus in OEB3. Our objective is to approximate the action-value function Q_t via fitting the parameter w, such that

$$w^\top \phi(s_t, a_t) \approx r_t(s_t, a_t) + \max_{a\in\mathcal{A}} Q_{t+1}(s_{t+1}, a),$$

where Q_{t+1} is given. We assume that we are given a Gaussian prior of the initial parameter $w \sim \mathcal{N}(0, \mathbf{I}/\lambda)$. With a slight abuse of notation, we denote by w_t the Bayesian posterior of the parameter w given the set of independent observations $\mathcal{D}_m = \{(s_t^\tau, a_t^\tau, s_{t+1}^\tau)\}_{\tau\in[0,m]}$. We further define the following noise with respect to the least-square problem in (5),

$$\epsilon = r_t(s_t, a_t) + \max_{a\in\mathcal{A}} Q_{t+1}(s_{t+1}, a) - w^\top \phi(s_t, a_t), \tag{6}$$

where (s_t, a_t, s_{t+1}) follows the distribution of trajectory. The following theorem justifies the UCB-bonus in OEB3 under the Bayesian linear regression perspective.

Theorem 2 (Formal Version of Theorem 1)**.** *We assume that ϵ follows the standard Gaussian distribution $\mathcal{N}(0, 1)$ given the state-action pair (s_t, a_t) and the parameter w. Let w follows the Gaussian prior $\mathcal{N}(0, \mathbf{I}/\lambda)$. We define*

$$\Lambda_t = \sum_{\tau=0}^{m} \phi(x_t^\tau, a_t^\tau)\phi(x_t^\tau, a_t^\tau)^\top + \lambda \cdot \mathbf{I}. \tag{7}$$

It then holds for the posterior of w_t given the set of independent observations $\mathcal{D}_m = \{(s_t^\tau, a_t^\tau, s_{t+1}^\tau)\}_{\tau\in[0,m]}$ that

$$\mathrm{Var}\big(\phi(s_t, a_t)^\top w_t\big) = \mathrm{Var}\big(\tilde{Q}_t(s_t, a_t)\big) = \phi(s_t, a_t)^\top \Lambda_t^{-1} \phi(s_t, a_t), \quad \forall (s_t, a_t) \in \mathcal{S}\times\mathcal{A}.$$

Here we denote by $\tilde{Q}_t = w_t^\top \phi$ the estimated action-value function.

Proof. The proof follows the standard analysis of Bayesian linear regression. See, e.g., West (1984) for a detailed analysis. We denote the target of the linear regression in (5) by

$$y_t = r_t(s_t, a_t) + \max_{a\in\mathcal{A}} Q_{t+1}(s_{t+1}, a).$$

By the assumption that ϵ follows the standard Gaussian distribution, we obtain that

$$y_t \,|\, (s_t, a_t), w \sim \mathcal{N}\big(w^\top \phi(s_t, a_t), 1\big). \tag{8}$$

Recall that we have the prior distribution $w \sim \mathcal{N}(0, \mathbf{I}/\lambda)$. Our objective is to compute the posterior density $w_t = w \,|\, \mathcal{D}_m$, where $\mathcal{D}_m = \{(s_t^\tau, a_t^\tau, s_{t+1}^\tau)\}_{\tau\in[0,m]}$ is the set of observations. It holds from Bayes rule that

$$\log p(w \,|\, \mathcal{D}_m) = \log p(w) + \log p(\mathcal{D}_m \,|\, w) + Const., \tag{9}$$

where $p(\cdot)$ denote the probability density function of the respective distributions. Plugging (8) and the probability density function of Gaussian distribution into (9) yields

$$\begin{aligned} \log p(w \,|\, \mathcal{D}_m) &= -\|w\|^2/2 - \sum_{\tau=1}^{m} \|w^\top \phi(s_t^\tau, a_t^\tau) - y_t^\tau\|^2/2 + Const. \\ &= -(w-\mu_t)^\top \Lambda_t^{-1}(w-\mu_t)/2 + Const., \end{aligned} \tag{10}$$

where we define

$$\mu_t = \Lambda_t^{-1} \sum_{\tau=1}^{m} \phi(s_t^\tau, a_t^\tau) y_t^\tau, \qquad \Lambda_t = \sum_{\tau=0}^{m} \phi(x_t^\tau, a_t^\tau)\phi(x_t^\tau, a_t^\tau)^\top + \lambda \cdot \mathbf{I}.$$

Thus, by (10), we obtain that $w_t = w \,|\, \mathcal{D}_m \sim \mathcal{N}(\mu_t, \Lambda_t^{-1})$. It then holds for all $(s_t, a_t) \in \mathcal{S} \times \mathcal{A}$ that

$$\mathrm{Var}\big(\phi(s_t, a_t)^\top w_t\big) = \mathrm{Var}\big(\tilde{Q}_t(s_t, a_t)\big) = \phi(s_t, a_t)^\top \Lambda_t^{-1} \phi(s_t, a_t),$$

which concludes the proof of Theorem 2. $\square$

Remark 1 (Extension to Neural Network Parameterization)**.** We remark that our proof can be extended to explain deep neural network parametrization under the overparameterized network regime (Arora et al., 2019). Under such a setting, a two-layer neural network $f(\cdot; W)$ with parameter W and ReLU activation function can be approximated by

$$f(x; W) \approx f(x; W_0) + \phi_{W_0}(x)^\top (W - W_0) = \phi_{W_0}(x)^\top W, \quad \forall x \in \mathcal{X},$$

where the approximation holds if the neural network is sufficiently wide (Arora et al., 2019). Here W_0 is the Gaussian distributed initial parameter and $\phi_{W_0} = ([\phi_{W_0}]_1, \ldots, [\phi_{W_0}]_m)^\top$ is the feature embedding defined as follows,

$$[\phi_{W_0}(x)]_r = \frac{1}{\sqrt{m}} \sigma\big(x^\top [W_0]_r\big), \quad \forall x \in \mathcal{X},\ r \in [m].$$

Hence, if we consider a Bayesian perspective of training neural network, where the parameter W is obtained by solving a Bayesian linear regression with the feature ϕ_{W_0}, then the proof of Theorem 2 can be applied to the setting upon conditioning on the random initialization W_0. Thus, Theorem 2 applies to the neural network parameterization under such an overparameterized neural network regime.

B. Raw Scores of all 49 Atari Games

Table 2. Raw scores for Atari games. Bold scores signify the best score out of all methods.

	Random	Human	BEBU	BEBU-UCB	BEBU-IDS	OEB3
Alien	227.8	6,875.0	**1,118.0**	811.1	857.9	916.9
Amidar	5.8	1676.0	81.7	**166.4**	148.1	94.0
Assault	222.4	1,496.0	1,377.0	**3,574.5**	2,441.8	2,996.2
Asterix	210.0	8,503.0	2,315.0	2,709.3	2,433.9	**2,719.0**
Asteroids	719.1	13,157.0	962.8	**1,025.0**	868.8	959.9
Atlantis	12,850.0	29,028.0	3,020,500.0	3,191,600.0	3,144,440.0	**3,146,300.0**
Bank Heist	14.2	734.4	331.8	277.0	361.6	**378.6**
Battle Zone	2,360.0	37,800.0	5,446.4	**16,348.8**	10,520.0	13,454.5
BeamRider	363.9	5,775.0	2,930.0	3,208.3	3,391.0	**3,736.7**
Bowling	23.1	154.8	29.9	30.7	**40.2**	30.0
Boxing	0.1	4.3	72.4	68.3	69.8	**75.1**
Breakout	1.7	31.8	**473.2**	382.3	412.7	423.1
Centipede	2,090.9	11,963.0	2,547.2	2,377.9	**3,328.4**	2,661.8
Chopper Command	811.0	9,882.0	930.6	1,013.4	1,100.0	**1,100.3**
Crazy Climber	10,780.5	35,411.0	49,735.7	39,187.5	42,242.9	**53,346.7**
Demon Attack	152.1	3,401.0	6,506.3	6,840.4	**7,080.0**	6,794.6
Double Dunk	-18.6	-15.5	-18.9	**-16.5**	-17.0	-18.2
Enduro	0.0	309.6	504.1	697.8	513.6	**719.0**
Fishing Derby	-91.7	5.5	-56.7	-83.8	**-53.3**	-60.1
Freeway	0.0	29.6	21.5	21.6	21.3	**32.1**
Frostbite	65.2	4,335.0	393.4	470.4	466.2	**1,277.3**
Gopher	257.6	2,321.0	4,842.6	**7,211.8**	7,171.5	6,359.5
Gravitar	173.0	2,672.0	256.1	321.0	283.3	**393.6**
H.E.R.O	1,027.0	25,763.0	2,951.4	2,905.0	3,059.4	**3,302.5**
Ice Hockey	-11.2	0.9	-5.4	-6.5	-4.6	**-4.2**
Jamesbond	29.0	406.7	**650.0**	360.3	302.1	434.3
Kangaroo	52.0	3,035.0	3624.2	2,711.1	**4,448.0**	2,387.0
Krull	1,598.0	2,395.0	15,716.7	11,499.0	10,818.0	**45,388.8**
Kung-Fu Master	258.5	22,736.0	56.0	20,738.9	**26,909.7**	16,272.2
Montezuma's Revenge	0.0	4,376.0	0.0	0.0	0.0	**0.0**
Ms. Pacman	307.3	15,693.0	1,723.8	1,706.8	1,615.5	**1,794.9**
Name This Game	2,292.3	4,076.0	8,275.3	6,573.9	**8,925.0**	8,576.8
Pong	-20.7	9.3	18.1	18.5	17.2	**18.7**
Private Eye	24.9	69,571.0	1,185.8	1,925.2	1,897.1	1,174.1
Q*Bert	163.9	13,455.0	3,588.4	3,783.2	3,696.0	**4,275.0**
River Raid	1,338.5	13,513.0	3,127.5	**3,617.7**	3,169.1	2,926.5
Road Runner	11.5	7,845.0	11,483.0	20,990.7	17,281.4	**21,831.4**
Robotank	2.2	11.9	10.3	13.3	10.7	**13.5**
Seaquest	68.4	20,182.0	447.0	**492.3**	332.4	332.1
Space Invaders	148.0	1,652.0	814.4	782.2	794.7	**904.9**
Star Gunner	664.0	10,250.0	1,467.2	1,201.5	1,158.9	**1,290.2**
Tennis	-23.8	-8.9	-1.0	-2.0	-1.0	**-1.0**
Time Pilot	3,568.0	5,925.0	2,622.1	3,321.2	1,950.6	**3,404.5**
Tutankham	11.4	167.6	167.0	151.0	80.5	**297.0**
Up and Down	533.4	9,082.0	**5,954.8**	4,530.2	4,619.7	5,100.8
Venture	0.0	1,188.0	42.9	3.4	**150.0**	16.1
Video Pinball	16,256.9	17,298.0	26,829.6	48,959.1	58,398.3	**80,607.0**
Wizard of Wor	563.5	4,757.0	810.8	**1,316.7**	578.2	480.7
Zaxxon	32.5	9,173.0	1,587.5	2,104.8	1,594.2	**2,842.0**