SUPPLEMENTARY MATERIAL FOR STABILITY AND GENERALISATION IN BATCH RL

Anonymous authors
Paper under double-blind review

1 OMITTED PROOFS

1.1 ASSUMPTIONS

Assumption 1. *ϕ maps to a compact set in $\mathbb{R}^d$, and is uniformly bounded, with $||\phi|| \leq \phi_{max}$.*

Assumption 2. *The density ratio is bounded from above:*

$$\sup_z \frac{P_{\mathcal{D}'}(z)}{P_{\mathcal{D}}(z)} \leq \rho_{max}.$$

1.2 OFF-DISTRIBUTION STABILITY AND GENERALISATION

Lemma 1. *If an algorithm $\mathcal{A}$ is ϵ-uniformly off-distribution stable for distributions $\mathcal{D}$ and $\mathcal{D}'$ with shared support, then:*

$$\mathcal{G}(\mathcal{A}, \mathcal{D}, \mathcal{D}') \leq \epsilon.$$

Proof.

$$\begin{aligned}
\mathbb{E}_{\mathcal{A},D\sim\mathcal{D}}[\mathcal{R}_{emp}(\mathcal{A}(D))] =& \mathbb{E}_{\mathcal{A},D}\left[\frac{1}{|D|}\sum_{i=1}^{|D|} l(\mathcal{A}(D), x_i)\right] \\
=& \mathbb{E}_{\mathcal{A},D\sim\mathcal{D},\hat{D}\sim\ \mathcal{D}}\left[\frac{1}{|D|}\sum_{i=1}^{|D|} l(A(D_{1:i-1}\cup\hat{x}_i\cup D_{i+1:|D|}), \hat{x}_i)\right] = V \\
=& V + \mathbb{E}_{\mathcal{A},D,D'\sim\mathcal{D}'}\left[\frac{1}{|D|}\sum_{i=1}^{|D|} l(\mathcal{A}(D), x'_i)\right] - \mathbb{E}_{\mathcal{A},D,D'\sim\mathcal{D}'}\left[\frac{1}{|D|}\sum_{i=1}^{|D|} l(\mathcal{A}(D), x'_i)\right] \\
=& \mathbb{E}_{\mathcal{A},D}\left[\mathbb{E}_{x'} l(\mathcal{A}(D), x')\right] + V - \mathbb{E}_{\mathcal{A},D,D'\sim\mathcal{D}'}\left[\frac{1}{|D|}\sum_{i=1}^{|D|} l(\mathcal{A}(D), x'_i)\right]
\end{aligned}$$

Because we are taking the mean of a sample mean for the last two term s, we can remove the dependency on the second data set as follows. With $D^{x_i/\hat{x}_i} := D_{1:i-1}\cup\hat{x}_i\cup D_{i+1:|D|}$, we have:

$$\begin{aligned}
V =& \mathbb{E}_{\mathcal{A},D}\left[\frac{1}{|D|}\left(\sum_{i=1}^{|D|}\mathbb{E}_{\hat{D}}[l(A(D^{x_i/\hat{x}_i}), \hat{x}_i)]\right)\right] \\
=& \mathbb{E}_{\mathcal{A},D}\left[\frac{1}{|D|}\left(\sum_{i=1}^{|D|}\mathbb{E}_{\hat{x}}[l(\mathcal{A}(D^{x_i/\hat{x}}), \hat{x})]\right)\right] = \mathbb{E}_{\mathcal{A},D,\hat{x}}\left[\frac{1}{|D|}\sum_{i=1}^{|D|} l(\mathcal{A}(D^{x_i/\hat{x}}), \hat{x})\right]
\end{aligned}$$

and

$$\mathbb{E}_{\mathcal{A},D,D'\sim\mathcal{D}'}\left[\frac{1}{|D|}\sum_{i=1}^{|D|} l(\mathcal{A}(D), x'_i)\right] = \mathbb{E}_{\mathcal{A},D,x'\sim\mathcal{D}'}\left[l(\mathcal{A}(D), x')\right]$$

Assuming equal support for $\mathcal{D}$ and $\mathcal{D}'$, this allows us to combine our expectations using a density ratio:

$$V - \mathbb{E}_{\mathcal{A},D,D'\sim\mathcal{D}'}\left[\frac{1}{|D|}\sum_{i=1}^{|D|} l(\mathcal{A}(D), x_i')\right] = \mathbb{E}_{\mathcal{A},D,\hat{x}\sim\mathcal{D}}\left[\frac{1}{|D|}\sum_{i=1}^{|D|} l(\mathcal{A}(D^{x_i/\hat{x}}), \hat{x})\right] - \mathbb{E}_{\mathcal{A},D,\hat{x}\sim\mathcal{D}}\left[\frac{P_{\mathcal{D}'}(\hat{x})}{P_{\mathcal{D}}(\hat{x})} l(\mathcal{A}(D), \hat{x})\right]$$

$$= \mathbb{E}_{\mathcal{A},D,\hat{x}\sim\mathcal{D}}\left[\frac{1}{|D|}\sum_{i=1}^{|D|} l(\mathcal{A}(D^{x_i/\hat{x}}), \hat{x}) - \frac{P_{\mathcal{D}'}(\hat{x})}{P_{\mathcal{D}}(\hat{x})} l(\mathcal{A}(D), \hat{x})\right]$$

Since the datasets differ at most at one point, we bound with suprema over datasets differing at most by one point, as well as the test point, giving us the desired results:

$$\left|\mathbb{E}_{\mathcal{A},D,\hat{x}\sim\mathcal{D}}\left[\frac{1}{|D|}\sum_{i=1}^{|D|} l(\mathcal{A}(D^{x_i/\hat{x}}), \hat{x}) - \frac{P_{\mathcal{D}'}(\hat{x})}{P_{\mathcal{D}}(\hat{x})} l(\mathcal{A}(D), \hat{x})\right]\right| \leq$$

$$\sup_{D,x,x',z}\left|\mathbb{E}_{\mathcal{A}}\left[l(\mathcal{A}(D\cup x), z) - \frac{P_{\mathcal{D}'}(\hat{z})}{P_{\mathcal{D}}(\hat{z})} l(\mathcal{A}(D\cup x'), z)\right]\right|$$

□

1.3 Expansivity of Update Rule

Lemma 2. *Under Assumption 1, and the additional projection step, the update rule $G(\hat{w})_X$ is η-expansive, with:*

$$\eta \leq 1 + \alpha\,(1+\gamma)\,\phi_{max}^2. \tag{1}$$

Proof. Starting from the definition of our update rule we have:

$$\begin{aligned}
\|G(\hat{u})_{X_{0:k}} - G(\hat{v})_{X_{0:k}}\| = \Bigg\| &\hat{u} - \alpha\begin{bmatrix}\phi(s,a)\\ \mathbf{0}\end{bmatrix}\left(\phi(s,a)^\top u - r - \gamma\sum_{a'}\pi(a'|s')\phi(s',a')^\top\bar{u}\right) \\
&-\hat{v} + \alpha\begin{bmatrix}\phi(s,a)\\ \mathbf{0}\end{bmatrix}\left(\phi(s,a)^\top v - r - \gamma\sum_{a'}\pi(a'|s')\phi(s',a')^\top\bar{v}\right)\Bigg\| \\
= \Bigg\| &\hat{u} - \hat{v} + \alpha\left(\hat{\phi}(s,a)\hat{\phi}(s,a)^\top\begin{bmatrix}v-u\\ \mathbf{0}\end{bmatrix} + \gamma\hat{\phi}(s,a)\left(\sum_{a'}\pi(a'|s')\hat{\phi}(s',a')^\top\right)\begin{bmatrix}\bar{u}-\bar{v}\\ \mathbf{0}\end{bmatrix}\right)\Bigg\| \\
\leq &\|\hat{u}-\hat{v}\| + \alpha\left(\left\|\phi(s,a)\phi(s,a)^\top\right\|\|v-u\| + \gamma\left\|\phi(s,a)\left(\sum_{a'}\pi(a'|s')\phi(s',a')^\top\right)\right\|\|\bar{u}-\bar{v}\|\right) \\
\leq &\|\hat{u}-\hat{v}\| + \alpha\left(\phi_{max}^2\|\hat{v}-\hat{u}\| + \gamma\phi_{max}^2\|\hat{u}-\hat{v}\|\right)
\end{aligned}$$

The first inequality follows from application of the triangle inequality to separate terms and the submultiplicativity of the induced matrix norm. The second comes from the fact that the policy weighted features forms a convex combination and our assumption on the feature map, as well as the fact that adding additional dimensions with nonzero elements can only increase distances. Dividing through by $\|\hat{u}-\hat{v}\|$ gives us the desired result. □

1.4 Linear Partially Fitted Expected SARSA is σ-bounded

Lemma 3. *Under Assumption 1, for finite integers k and p, and the additional projection step, the update rule $G(w)_{X_k}$ is σ-bounded, with:*

$$\sigma \leq \alpha\phi_{max}\left((1+\gamma)\phi_{max}w_{max} + r_{max}\right)\left(\alpha\phi_{max}^2 + 1\right)^k. \tag{2}$$

Proof. Since our update rule doesn't modify the target network parameters, these cancel out in the difference, and are omitted for clarity. Starting from our update rule:

$$||w_k - G(w_k)_{X_{0:k}}|| = \left\|w - w + \alpha\phi(S_k,A_k)\left(\phi(S_k,A_k)^\top w_k - R_k - \gamma\sum_{A'}\pi(S_k',A')\phi(S_k',A')\bar{w}\right)\right\|,$$

by the triangle inequality, submultiplicativity of the induced norm, and the fact that the sum over A' is a convex combination:

$$||w_k - G(w_k)_{X_{0:k}}|| \leq \alpha\phi_{max}\left(\phi_{max}||w_k|| + r_{max} + \gamma\phi_{max}w_{max}\right),$$

unrolling w_k, and applying the triangle inequality:

$$\begin{aligned}
||w_k - G(w_k)_{X_{0:k}}|| &\leq \alpha\phi_{max}\left(\phi_{max}(\|w_{k-1}\| + \alpha\phi_{max}\left(\phi_{max}||w_{k-1}|| + r_{max} + \gamma\phi_{max}w_{max}\right) + r_{max} + \gamma\phi_{max}w_{max}\right), \\
&\leq \alpha\phi_{max}\Bigg(\phi_{max}((1+\alpha\phi_{max}^2)\|w_{k-1}\| + \alpha\phi_{\max}r_{max} + \gamma\alpha\phi_{max}^2 w_{max}) \\
&\quad + r_{max} + \gamma\phi_{max}w_{max}\Bigg), \\
&\leq \alpha\phi_{max}\left(\phi_{max}\left((1+\alpha\phi_{max}^2)^k w_{max} + \sum_{i=0}^{k-1}(1+\alpha\phi_{max}^2)^i(\alpha\phi_{max}r + \gamma\alpha\phi_{max}^2 w_{max})\right)\right. \\
&\quad \left. + r_{max} + \gamma\phi_{max}w_{max}\right), \\
&\leq \alpha\phi_{max}\left((1+\gamma)\phi_{max}w_{max} + r_{max}\right)\left(\alpha\phi_{max}^2 + 1\right)^k.
\end{aligned}$$

The second last line comes from unrolling all the way to w_0, and by the projection step $||w_0|| \leq w_{\max}$. The final line comes from the partial sum of a geometric series:

$$\sum_{i=0}^{k-1}(1+a)^i = \frac{(1+a)^k - 1}{a},$$

and simplifying. □

1.5 Linear Partially Fitted Expected SARSA is Uniformly Stable

Before proving the main result, first we provide proof of the following lemma:

Lemma 4. *Let w_t and w'_t be the output from running Algorithm 1 on our original dataset, and the perturbed dataset respectively, with algorithm randomness held constant across both runs. Let ζ_t represent the parameter gap in supremum norm at time t: $|w_t - w'_t|_\infty$. Then, under Assumption 1 with supremum norm, and using supremum norm for the projection step, for every $t_0 \in \{1, ..., |D|)\}, z \in \Omega$, we have:*

$$\mathbb{E}_{\mathcal{A}}|MSBE(w_T; z) - MSBE(w'_T; z)| \leq \frac{t_0}{|D|}((1+\gamma)d\phi_{max}w_{max} + r_{max})^2 + M\sqrt{d}\mathbb{E}[\zeta_T|\zeta_{t_0} = 0]. \tag{3}$$

where:

$$M = \left(d^{\frac{3}{2}}(1+\gamma)^2\phi_{max}^2 w_{max} + \sqrt{d}2(1+\gamma)\phi_{max}r_{max}\right)$$

Proof. We consider the parameter gap at epoch t_0, ζ_{t_0}. There are two possibilities: either the algorithm has not encountered the perturbed data point by t_0, in which case we can ensure that the gap is zero at t_0, or the data point has been encountered, for which we assume the worst-case. Let l here represent the MSBE, then:

$$\begin{aligned}
\mathbb{E}_{\mathcal{A}}|l(w_T; z) - l(w'_T; z)| &\leq P(\zeta_{t_0} \neq 0)\left((1+\gamma)d\phi_{max}w_{max} + r_{max}\right)^2 \\
&\quad + P(\zeta_{t_0} = 0)\mathbb{E}_{\mathcal{A}}|l(w_T; z) - l(w'_T; z)|\zeta_{t_0} = 0|.
\end{aligned}$$

The probability that the perturbed data point is selected in a batch is $K/|D|$, and this is done for t_0/K batches. Thus, the probability that the data point is selected at or before step t_0 is upper bounded, by union bound as: $\frac{t_0}{|D|}$. We use this as an upper bound on $P(\zeta_{t_0} \neq 0)$. Then:

$$\mathbb{E}_{\mathcal{A}}|l(w_T; z) - l(w'_T; z)| \leq \frac{t_0}{|D|}\left((1+\gamma)d\phi_{max}w_{max} + r_{max}\right)^2 + M\sqrt{d}\mathbb{E}_{\mathcal{A}}[\zeta_T|\zeta_{t_0} = 0],$$

which we get from dropping the complementary probability term, M comes from the definition of MSBE, and the $\sqrt{d}$ comes from bounding the 2-norm (induced by the inner products in the MSBE) with the ∞-norm. □

We now move on to proving the main result:

Theorem 1. *Under Assumptions 1 (under supremum norm) and 2 and the projection step applied under supremum norm, for monotonically decreasing $\alpha_t \leq \frac{c}{t}$, after $T = NK$ updates, Algorithm 1 is ϵ-uniformly off-distributionally stable, with:*

$$\epsilon \leq (1 - \rho_{max})((1 + \gamma)d\phi_{max}w_{max} + r_{max})^2 + \rho_{max}E, \tag{4}$$

where:

$$E = \left((1 + \gamma)d\phi_{max}w_{max} + r_{max}\right)^{\frac{2\beta c}{\beta c+1}} \left(\frac{1 + \frac{1}{\beta c}}{|D| - 1}\right) \left(2M\sqrt{d}Lc\right)^{\frac{1}{\beta c+1}} T^{\frac{\beta c}{\beta c+1}}$$

and:

$$M = \left(d^{\frac{3}{2}}(1 + \gamma)^2\phi_{max}^2 w_{max} + \sqrt{d}2(1 + \gamma)\phi_{max}r_{max}\right)$$

and:

$$L = \phi_{max}\left((1 + \gamma)\phi_{max}w_{max} + r_{max}\right)\left(\alpha\phi_{max}^2 + 1\right)^K,$$

and:

$$\beta = (1 + \gamma)\,\phi_{max}^2.$$

Proof. We let l be the MSBE, and w_T, w'_T outputs from $\mathcal{A}(D \cup x)$ and $\mathcal{A}(D \cup x')$ for arbitrary D, x, x' after $T = K\tau$ time steps, for natural K, τ. Let z be an arbitrary data point. From the definition of uniform off-distribution stability, we have:

$$\begin{aligned}\mathbb{E}_{\mathcal{A}}\left[l(w_T, z) - \frac{P_{\mathcal{D}'}(z)}{P_{\mathcal{D}}(z)}l(w'_T, z)\right] &\leq \mathbb{E}_{\mathcal{A}}\left[|(1 - \rho_{max})l(w_T, z) + \rho_{max}\big(l(w_T, z) - l(w'_T, z)\big)|\right], \\ &\leq |(1 - \rho_{max})\big((1 + \gamma)d\phi_{max}w_{max} + r_{max}\big)^2| + \rho_{max}\mathbb{E}_{\mathcal{A}}\left[\big|l(w_T, z) - l(w'_T, z)\big|\right].\end{aligned}$$

The first inequality comes from adding and subtracting a cross term, as well as Assumption 2. Since the features are bounded by Assumption 1, the rewards by MDP definition, and the parameters by the projection step, we can bound the loss trivially, which, alongside the triangle inequality and linearity of expectation, gives the second line. This leaves us to bound the loss gap as in previous algorithmic stability work. We use the same hitting time argument as that of Hardt et al. (2016). From Lemma 4, we have that:

$$\mathbb{E}_{\mathcal{A}}\left[|l(w_T, z) - l(w'_T, z)|\right] \leq \frac{t_0}{|D| - 1}\big((1+\gamma)d\phi_{max}w_{max}+r_{max}\big)^2 + M\sqrt{d}\mathbb{E}_{\mathcal{A}}[\zeta_T|\zeta_{t_0} = 0]. \tag{5}$$

Which leaves us needing to bound $\mathbb{E}_{\mathcal{A}}[\zeta_T|\zeta_{t_0} = 0]$, after which we can minimise over t_0. Our proof follows identically to that of Hardt et al. (2016) here, where we bound $\mathbb{E}_{\mathcal{A}}[\zeta_T|\zeta_{t_0} = 0]$ recursively. At each step, the algorithm selects a data point, which, with probability $\frac{1}{|D|}$ is the perturbed point. If this is the case, the update rules used by the algorithms are different, and the parameters may step in opposite directions by the maximum step size, L. On the other hand, with probability $1 - \frac{1}{|D|}$, the same data point is selected, and the same update rule is applied, leading parameters to diverge by at most the expansivity of the update. Since the parameter fixing update is non-expansive in infinity norm, we can ignore these updates. From Hardt et al. (2016) pp. 13-14, we then have:

$$\mathbb{E}_{\mathcal{A}}[\zeta_T|\zeta_{t_0} = 0] \leq \frac{2L}{\beta(|D| - 1)}\left(\frac{T}{t_0}\right)^{\beta c}$$

Which we plug into (5), and minimise over t_0, which leads us to:

$$t_0 = \left(\frac{2cM\sqrt{d}LT^{\beta c}}{\big((1 + \gamma)\phi_{max}w_{max} + r_{max}\big)^2}\right)^{\frac{1}{\beta c}}$$

which, when plugged into (5) alongside the previous and simplified, leads to the result. □

REFERENCES

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.