Appendix: Lipschitz Continuity in Model-based Reinforcement Learning

We first restate the core lemmas, theorems, and claims presented in our paper below:

**Lemma 1.** A generalized transition function \( \hat{T}_g \) induced by a Lipschitz model class \( F_g \) is Lipschitz with a constant:

\[
K^A_{W,W}(\hat{T}_g) := \sup_a \sup_{\mu_1, \mu_2} \frac{W(\hat{T}_g(\cdot | \mu_1, a), \hat{T}_g(\cdot | \mu_2, a))}{W(\mu_1, \mu_2)} \leq K_F
\]

**Lemma 2.** (Composition Lemma) Define three metric spaces \( (M_1, d_1) \), \( (M_2, d_2) \), and \( (M_3, d_3) \). Define Lipschitz functions \( f : M_2 \mapsto M_3 \) and \( g : M_1 \mapsto M_2 \) with constants \( K_{d_2,d_3}(f) \) and \( K_{d_1,d_2}(g) \). Then, \( h : f \circ g : M_1 \mapsto M_3 \) is Lipschitz with constant \( K_{d_1,d_3}(h) \leq K_{d_2,d_3}(f)K_{d_1,d_2}(g) \).

**Theorem 1.** Define a \( \Delta \)-accurate \( \hat{T}_g \) with the Lipschitz constant \( K_F \) and an MDP with a Lipschitz transition function \( T_g \) with constant \( K_T \). Let \( \bar{K} = \min\{K_F, K_T\} \). Then \( \forall n \geq 1 \):

\[
\delta(n) := W(\hat{T}_g^n(\cdot | \mu), T_g^n(\cdot | \mu)) \leq \Delta \sum_{i=0}^{n-1} (\overline{K})^i .
\]

**Theorem 2.** Assume a Lipschitz model class \( F_g \) with a \( \Delta \)-accurate \( \hat{T} \) with \( \bar{K} = \min\{K_F, K_T\} \). Further, assume a Lipschitz reward function with constant \( K_R = K_{d_s,R}(R) \). Then \( \forall s \in S \) and \( \bar{K} \in [0, \frac{1}{\gamma}) \)

\[
|V_T(s) - V_\hat{T}(s)| \leq \frac{\gamma K_R \Delta}{(1 - \gamma)(1 - \gamma \bar{K})}.
\]

**Lemma 3.** Given a Lipschitz function \( f : S \mapsto \mathbb{R} \) with constant \( K_{d_S,d_W}(f) \):

\[
K^A_{d_S,d_W}( \int \hat{T}(s'|s,a)f(s')ds' ) \leq K_{d_S,d_W}(f)K^A_{d_S,W}(\hat{T}) .
\]

**Lemma 4.** The following operators (Asadi & Littman, 2017) are Lipschitz with constants:

1. \( K_{\|\cdot\|,d_R}(\max(x)) = K_{\|\cdot\|,d_R}(\text{mean}(x)) = K_{\|\cdot\|,d_R}(\epsilon\text{-greedy}(x)) = 1 \)
2. \( K_{\|\cdot\|,d_R}(\text{min}_{\beta}(x)) := \frac{\log \sum_i x_i e^{\delta x_i}}{\beta} = 1 \)
3. \( K_{\|\cdot\|,d_R}(\text{boltz}_{\beta}(x)) := \frac{\sum_{i=1}^n x_i e^{\beta x_i}}{\sum_{i=1}^n e^{\beta x_i}} \leq \sqrt{|A|} + \beta V_{\max}|A| \)

**Theorem 3.** For any non-expansion backup operator \( f \) outlined in Lemma 4, GVI computes a value function with a Lipschitz constant bounded by \( \frac{K^A_{d_S,d_R}(R)}{1 - \gamma K_{d_S,W}(T)} \) if \( \gamma K^A_{d_S,W}(T) < 1 \).

We now provide proofs of various results mentioned in the paper:

**Claim 1.** In a finite MDP, transition probabilities can be expressed using a finite set of deterministic functions and a distribution over the functions.

**Proof.** Let \( Pr(s, a, s') \) denote the probability of a transition from \( s \) to \( s' \) when executing the action \( a \). Define an ordering over states \( s_1, ..., s_n \) with an additional unreachable state \( s_0 \). Now define the cumulative probability distribution:

\[
C(s, a, s_i) := \sum_{j=0}^{i} Pr(s, a, s_j) .
\]

Further define \( L \) as the set of distinct entries in \( C \):

\[
L := \left\{ C(s, a, s_i) \mid s \in S, i \in [0,n] \right\} .
\]
Note that, since the MDP is assumed to be finite, then $|L|$ is finite. We sort the values of $L$ and denote, by $c_i$, $i$th smallest value of the set. Note that $c_0 = 0$ and $c_{|L|} = 1$. We now build deterministic set of functions $f_1, \ldots, f_{|L|}$ as follows: $\forall i = 1 \to |L|$ and $\forall j = 1$ to $n$, define $f_i(s) = s_j$ if and only if:

$$C(s, a, s_{j-1}) < c_i \leq C(s, a, s_j).$$

We also define the probability distribution $g$ over $f$ as follows:

$$g(f_i|a) := c_i - c_{i-1}.$$

Given the functions $f_1, \ldots, f_{|L|}$ and the distribution $g$, we can now compute the probability of a transition to $s_j$ from $s$ after executing action $a$:

$$\sum_i \mathbb{1}(f_i(s) = s_j) g(f_i|a)$$

$$= \sum_i \mathbb{1}(C(s, a, s_{j-1}) < c_i \leq C(s, a, s_j)) (c_i - c_{i-1})$$

$$= C(s, a, s_j) - C(s, a, s_{j-1})$$

$$= Pr(s, a, s_j),$$

where $\mathbb{1}$ is a binary function that outputs one if and only if its condition holds. We reconstructed the transition probabilities using distribution $g$ and deterministic functions $f_1, \ldots, f_{|L|}$.

**Claim 2.** Given a deterministic and linear transition model, and a linear reward signal, the bounds provided in Theorems 1 and 2 are both tight.

Assume a linear transition function $T$ defined as:

$$T(s) = Ks$$

Assume our learned transition function $\hat{T}$:

$$\hat{T}(s) := Ks + \Delta$$

Note that:

$$\max_s |T(s) - \hat{T}(s)| = \Delta$$

and that:

$$\min\{K_T, K_{\hat{T}}\} = K$$

First observe that the bound in Theorem 2 is tight for $n = 2$:

$$\forall s \quad |T^2(s) - \hat{T}^2(s)| = |K^2s - K^2s + \Delta(1 + K)| = \Delta \sum_{i=0}^{1} K^i$$

and more generally and after $n$ compositions of the models, denoted by $T^n$ and $\hat{T}^n$, the following equality holds:

$$\forall s \quad |T^n(s) - \hat{T}^n(s)| = \Delta \sum_{i=0}^{n-1} K^i$$

Let further assume that the reward is linear:

$$R(s) = K_R s$$

Consider the state $s = 0$. Note that clearly $v(0) = 0$. We now compute the value predicted using $\hat{T}$, denoted by $\hat{v}(0)$:

$$\hat{v}(0) = R(0) + \gamma R(0 + \Delta \sum_{i=0}^{0} K^i) + \gamma^2 R(0 + \Delta \sum_{i=0}^{1} K^i) + \gamma^3 R(0 + \Delta \sum_{i=0}^{2} K^i) + \ldots$$

$$= 0 + \gamma K_R \Delta \sum_{i=0}^{0} K^i + \gamma^2 K_R \Delta \sum_{i=0}^{1} K^i + \gamma^3 K_R \Delta \sum_{i=0}^{2} K^i + \ldots$$
We can also prove this using the Kantarovich-Rubinstein duality theorem:

\[ W \sum_{n=0}^{\infty} \gamma^n \sum_{i=0}^{n-1} K^i = \frac{\gamma K_R \Delta}{(1 - \gamma)(1 - \gamma K)}, \]

and so:

\[ |\nu(0) - \hat{\nu}(0)| = \frac{\gamma K_R \Delta}{(1 - \gamma)(1 - \gamma K)} \]

Note that this exactly matches the bound derived in our Theorem 2.

**Lemma 1.** A generalized transition function \( \hat{T}_G \) induced by a Lipschitz model class \( F_G \) is Lipschitz with a constant:

\[
K_{W,W}^{\mathcal{A}}(\hat{T}_G) := \sup_a \sup_{\mu_1, \mu_2} \frac{W(\hat{T}_G(\cdot|\mu_1, a), \hat{T}_G(\cdot|\mu_2, a))}{W(\mu_1, \mu_2)} \leq K_F
\]

**Proof.**

\[
W(\hat{T}(\cdot | \mu_1, a), \hat{T}(\cdot | \mu_2, a)) := \inf_j \int s'_1 \int s'_2 \int s''_1 \int s''_2 f(s_1, s_2)ds_1 ds_2
\]

\[
= \inf_j \int s_1 \int s_2 \int s'_1 \int s''_1 \sum_f \mathbb{1}(f(s_1) = s'_1) f(s_2) ds_1 ds_2
\]

\[
= \inf_j \int s_1 \int s_2 \sum_f j(s_1, s_2) f(s_1, s_2) ds_1 ds_2
\]

\[
\leq K_F \inf_j \int s_1 \int s_2 \sum_f g(f) j(s_1, s_2) ds_1 ds_2
\]

\[
= K_F \sup_f \int s_1 \int s_2 j(s_1, s_2) ds_1 ds_2
\]

\[
= K_F \sup_f g(f) W(\mu_1, \mu_2) = K_F W(\mu_1, \mu_2)
\]

Dividing by \( W(\mu_1, \mu_2) \) and taking sup over \( a, \mu_1, \) and \( \mu_2, \) we conclude:

\[
K_{W,W}^{\mathcal{A}}(\hat{T}) = \sup_{\mu_1, \mu_2} \frac{W(\hat{T}(\cdot | \mu_1, a), \hat{T}(\cdot | \mu_2, a))}{W(\mu_1, \mu_2)} \leq K_F.
\]

We can also prove this using the Kantarovich-Rubinstein duality theorem:

For every \( \mu_1, \mu_2, \) and \( a \in \mathcal{A} \) we have:

\[
W(\hat{T}_G(\cdot | \mu_1, a), \hat{T}_G(\cdot | \mu_2, a)) = \sup_{f: K_{\mathcal{L}} \leq 1} \int s (\hat{T}_G(s | \mu_1, a) - \hat{T}_G(s | \mu_2, a)) f(s) ds
\]

\[
= \sup_{f: K_{\mathcal{L}} \leq 1} \int s_0 (\hat{T}(s | s_0, a) \mu_1(s_0) - \hat{T}(s | s_0, a) \mu_2(s_0)) f(s) ds ds_0
\]

\[
= \sup_{f: K_{\mathcal{L}} \leq 1} \int s_0 \sum_t g(t | a) \mathbb{1}(t(s_0) = s) (\mu_1(s_0) - \mu_2(s_0)) f(s) ds ds_0
\]

\[
= \sup_{f: K_{\mathcal{L}} \leq 1} \sum_t g(t | a) \int s_0 \int s \mathbb{1}(t(s_0) = s) (\mu_1(s_0) - \mu_2(s_0)) f(s) ds ds_0
\]

\[
= \sup_{f: K_{\mathcal{L}} \leq 1} \sum_t g(t | a) \int s_0 (\mu_1(s_0) - \mu_2(s_0)) f(t(s_0)) ds_0
\]
Again we conclude by dividing by $W$ and taking sup over $a, \mu_1,$ and $\mu_2$. \hfill \Box

**Lemma 2.** (Composition Lemma) Define three metric spaces $(M_1, d_1)$, $(M_2, d_2)$, and $(M_3, d_3)$. Define Lipschitz functions $f : M_2 \to M_3$ and $g : M_1 \to M_2$ with constants $K_{d_2, d_3}(f)$ and $K_{d_1, d_2}(g)$. Then, $h : f \circ g : M_1 \to M_3$ is Lipschitz with constant $K_{d_1, d_3}(h) \leq K_{d_2, d_3}(f)K_{d_1, d_2}(g)$.

**Proof.**

$$K_{d_1, d_3}(h) = \sup_{s_1, s_2} \frac{d_3(f(g(s_1)), f(g(s_2)))}{d_1(s_1, s_2)}$$

$$= \sup_{s_1, s_2} \frac{d_2(g(s_1), g(s_2))}{d_1(s_1, s_2)} \frac{d_3(f(g(s_1)), f(g(s_2)))}{d_2(g(s_1), g(s_2))}$$

$$\leq \sup_{s_1, s_2} \frac{d_2(g(s_1), g(s_2))}{d_1(s_1, s_2)} \frac{d_3(f(s_1), f(s_2))}{d_2(s_1, s_2)}$$

$$= K_{d_1, d_2}(g)K_{d_2, d_3}(f).$$

**Lemma 3.** Given a Lipschitz function $f : S \to \mathbb{R}$ with constant $K_{d_S, d_0}(f)$:

$$K_{d_S, d_0}^A \left( \int_S \hat{T}(s'|s, a) f(s') ds' \right) \leq K_{d_S, d_0}(f)K_{d_S, W}^A(\hat{T}).$$

**Proof.**

$$K_{d_S, d_0}^A \left( \int_S \hat{T}(s'|s, a) f(s') ds' \right) = \sup_a \frac{\left| \int_{s'} (\hat{T}(s'|s_1, a) - \hat{T}(s'|s_2, a)) f(s') ds' \right|}{d(s_1, s_2)}$$

$$= \sup_a \frac{\left| \int_{s'} (\hat{T}(s'|s_1, a) - \hat{T}(s'|s_2, a)) f(s') \frac{K_{d_S, d_0}(f)}{K_{d_S, d_0}(f)} ds' \right|}{d(s_1, s_2)}$$

$$= K_{d_S, d_0}(f) \sup_a \frac{\left| \int_{s'} (\hat{T}(s'|s_1, a) - \hat{T}(s'|s_2, a)) f(s') ds' \right|}{d(s_1, s_2)}$$

$$\leq K_{d_S, d_0}(f) \sup_a \frac{\sup_{g : K_{d_S, d_0}(g) \leq 1} \int_{s'} (\hat{T}(s'|s_1, a) - \hat{T}(s'|s_2, a)) g(s') ds'}{d(s_1, s_2)}$$

$$= K_{d_S, d_0}(f) \sup_a \frac{W(\hat{T}(-|s_1, a), \hat{T}(-|s_2, a))}{d(s_1, s_2)}$$

$$= K_{d_S, d_0}(f)K_{d_S, W}^A(\hat{T}).$$
Lemma 4. The following operators (Asadi & Littman, 2017) are Lipschitz with constants:

1. \( K_{||\cdot||_{d_R}}(\text{max}(x)) = K_{||\cdot||_{d_R}}(\text{mean}(x)) = K_{||\cdot||_{d_R}}(\epsilon\text{-greedy}(x)) = 1 \)

2. \( K_{||\cdot||_{d_R}}(\text{mm}_\beta(x) := \frac{\log \sum_i e^{\beta x_i}}{\beta}) = 1 \)

3. \( K_{||\cdot||_{d_R}}(\text{boltz}_\beta(x) := \frac{\sum_i x_i e^{\beta x_i}}{\sum_i e^{\beta x_i}}) \leq \sqrt{|A|} + \beta V_{\max}|A| \)

Proof. 1 was proven by Littman & Szepesvári (1996), and 2 is proven several times (Fox et al., 2016; Asadi & Littman, 2017; Nachum et al., 2017; Neu et al., 2017). We focus on proving 3. Define

\[
\rho(x) = \frac{e^{\beta x_i}}{\sum_{i=1}^n e^{\beta x_i}},
\]

and observe that \( \text{boltz}_\beta(x) = x^\top \rho(x) \). Gao & Pavel (2017) showed that \( \rho \) is Lipschitz:

\[
||\rho(x_1) - \rho(x_2)||_2 \leq \beta ||x_1 - x_2||_2
\]

Using their result, we can further show:

\[
\begin{align*}
|\rho(x_1)^\top x_1 - \rho(x_2)^\top x_2| &\leq |\rho(x_1)^\top x_1 - \rho(x_1)^\top x_2| + |\rho(x_1)^\top x_2 - \rho(x_2)^\top x_2| \\
&\leq ||\rho(x_1)||_2 ||x_1 - x_2||_2 + ||x_2||_2 ||\rho(x_1) - \rho(x_2)||_2 \quad \text{(Cauchy-Schwartz)} \\
&\leq (1 + \beta V_{\max} \sqrt{|A|}) ||x_1 - x_2||_2 \\
&\leq \frac{||x_1||_\infty + \beta V_{\max} ||A||}{|A|} ||x_1 - x_2||_\infty,
\end{align*}
\]

dividing both sides by \( ||x_1 - x_2||_\infty \) leads to 3.

Below, we derive the Lipschitz constant for various functions.

**ReLU non-linearity** We show that \( \text{ReLU} : \mathbb{R}^n \rightarrow \mathbb{R}^n \) has Lipschitz constant 1 for \( p \).

\[
K_{||\cdot||_p,||\cdot||_p}(\text{ReLU}) = \sup_{x_1, x_2} \frac{||\text{ReLU}(x_1) - \text{ReLU}(x_2)||_p}{||x_1 - x_2||_p}
\]

\[
= \sup_{x_1, x_2} \frac{\left(\sum |\text{ReLU}(x_1)_i - \text{ReLU}(x_2)_i|^p\right)^{\frac{1}{p}}}{||x_1 - x_2||_p}
\]

(We can show that \( |\text{ReLU}(x_1)_i - \text{ReLU}(x_2)_i| \leq |x_1,i - x_2,i| \) and so):

\[
\leq \sup_{x_1, x_2} \frac{\left(\sum |x_1,i - x_2,i|^p\right)^{\frac{1}{p}}}{||x_1 - x_2||_p}
\]

\[
= \sup_{x_1, x_2} \frac{||x_1 - x_2||_p}{||x_1 - x_2||_p} = 1
\]

**Matrix multiplication** Let \( W \in \mathbb{R}^{n \times m} \). We derive the Lipschitz continuity for the function \( x \circ W(x) = Wx \).

For \( p = \infty \) we have:

\[
K_{||\cdot||_\infty,||\cdot||_\infty}(\times W(x))
\]
We sampled each function 30 times, where the input was chosen uniformly randomly from

\[ \mathbb{R}^4 \]

We used the following 5 functions to generate the dataset:

- Vector addition

\[ j \text{ refers to } j = 2 \]

\[ \leq \sup_j \|W_j\| \|x_1 - x_2\|_\infty \]

\[ = \sup_j \|W_j\|_1 \]

where \( W_j \) refers to \( j \)th row of the weight matrix \( W \). Similarly, for \( p = 1 \) we have:

\[
K_x \|1\|, \|1\| \left( \times W(x_1) \right) \\
= \sup_{x_1, x_2} \frac{\|xW(x_1) - xW(x_2)\|_1}{\|x_1 - x_2\|_1} = \sup_{x_1, x_2} \frac{\|Wx_1 - Wx_2\|_1}{\|x_1 - x_2\|_1} = \sup_{x_1, x_2} \frac{\|W(x_1 - x_2)\|_1}{\|x_1 - x_2\|_1}
\]

\[
= \sup_{x_1, x_2} \sum_j |W_j(x_1 - x_2)|
\]

\[
\leq \sup_{x_1, x_2} \frac{\sum_j \|W_j\|_1 \|x_1 - x_2\|_1}{\|x_1 - x_2\|_1} = \sum_j \|W_j\|_1
\]

and finally for \( p = 2 \):

\[
K_{x} \|2\|, \|2\| \left( \times W(x_1) \right) \\
= \sup_{x_1, x_2} \frac{\|xW(x_1) - xW(x_2)\|_2}{\|x_1 - x_2\|_2} = \sup_{x_1, x_2} \frac{\|Wx_1 - Wx_2\|_2}{\|x_1 - x_2\|_2} = \sup_{x_1, x_2} \frac{\|W(x_1 - x_2)\|_2}{\|x_1 - x_2\|_2}
\]

\[
= \sup_{x_1, x_2} \sqrt{\sum_j |W_j(x_1 - x_2)|^2} \]

\[
\leq \sup_{x_1, x_2} \frac{\sqrt{\sum_j \|W_j\|_2^2 \|x_1 - x_2\|_2^2}}{\|x_1 - x_2\|_2} = \sqrt{\sum_j \|W_j\|_2^2}
\]

**Vector addition** We show that \( +b : \mathbb{R}^n \rightarrow \mathbb{R}^n \) has Lipschitz constant 1 for \( p = 0, 1, \infty \) for all \( b \in \mathbb{R}^n \).

\[
K_{\|p\|, \|p\|} (\text{ReLU}) = \sup_{x_1, x_2} \frac{\|+b(x_1) - +b(x_2)\|_p}{\|x_1 - x_2\|_p} = \sup_{x_1, x_2} \frac{\|\{(x_1 + b) - (x_2 + b)\|_p}{\|x_1 - x_2\|_p} = \|x_1 - x_2\|_p = 1
\]

**Supervised-learning domain** We used the following 5 functions to generate the dataset:

\[
f_0(x) = \tanh(x) + 3 \\
f_1(x) = x \times x \\
f_2(x) = \sin(x) - 5 \\
f_3(x) = \sin(x) - 3 \\
f_4(x) = \sin(x) \times \sin(x)
\]

We sampled each function 30 times, where the input was chosen uniformly randomly from \([-2, 2]\) each time.
References


