

A Proof of Theorem 3.2

This section provides a detailed proof of the existence and convergence of GBSM.

A.1 The Existence of d^{1-2}

To prove the existence of d^{1-2} , we introduce the Knaster-Tarski fixed-point theorem. Let (\mathcal{X}, \preceq) denote a partial order, which means certain pairs of elements within the set \mathcal{X} are comparable under the homogeneous relation \preceq [21]. If this partial order has least upper bounds and greatest lower bounds for its arbitrary subsets, it is called a complete lattice. The Knaster-Tarski fixed-point theorem asserts that for a continuous function on a complete lattice, the iterative application of this function to the least element of the lattice converges to a fixed point \bar{x} , which satisfies $\bar{x} = f(\bar{x})$. Formally, the theorem is stated as follows.

Lemma A.1 (Knaster-Tarski fixed-point theorem [21]). *If the partial order (\mathcal{X}, \preceq) is a complete lattice and $f: \mathcal{X} \rightarrow \mathcal{X}$ is a continuous function. Then, f has a least fixed point, given by*

$$\text{fix}(f) = \sqcup_{n \in \mathbb{N}} f^{(n)}(x_0), \quad (24)$$

where x_0 is the least element of \mathcal{X} , \sqcup denotes the least upper bound, $f^{(n)}(x_0) = f(f^{(n-1)}(x_0))$, and $f^{(1)}(x_0) = f(x_0)$. Here, the continuity of f is defined such that for any increasing sequence $\{x_n\}$ in \mathcal{X} , it satisfies

$$f(\sqcup_{n \in \mathbb{N}} \{x_n\}) = \sqcup_{n \in \mathbb{N}} \{f(x_n)\}. \quad (25)$$

Let \mathcal{D} denote the set of all cost functions, which are defined as maps that satisfy $\mathcal{S}_1 \times \mathcal{S}_2 \rightarrow [0, \frac{\bar{R}}{1-\gamma}]$. Equip \mathcal{D} with the usual pointwise ordering: Consider two cost function, say d and $d' \in \mathcal{D}$, denote $d \leq d'$ if and only if $d(s, s') \leq d'(s, s')$ for any $s \in \mathcal{S}_1$ and $s' \in \mathcal{S}_2$. Then \mathcal{D} forms a complete lattice with the least element d_0^{1-2} , i.e., the constant zero function. Given s and s' , we regard the recursive definition in (5) as a function of d and accordingly define $F: \mathcal{D} \rightarrow \mathcal{D}$ by

$$F(d | s, s') = \max_a \left\{ |R_1(s, a) - R_2(s', a)| + \gamma W_1(\mathbb{P}_1(\cdot | s, a), \mathbb{P}_2(\cdot | s', a); d) \right\}. \quad (26)$$

Utilizing the Knaster-Tarski fixed-point theorem, the existence of d^{1-2} is achieved if the continuity of F holds on \mathcal{D} .

We first prove the continuity of the second term in F . Define $F_{W_1}: \mathcal{D} \rightarrow \mathcal{D}$ by

$$F_{W_1}(d | s, s') = W_1(\mathbb{P}_1(\cdot | s, a), \mathbb{P}_2(\cdot | s', a); d). \quad (27)$$

Lemma A.2. F_{W_1} is continuous on \mathcal{D} .

Proof. We follow the definition of continuity defined in Lemma A.1. Let $s_i \in \mathcal{S}_1$ and $s_j \in \mathcal{S}_2$. Regard $F_{W_1}(s_i, s_j; d)$ as a function of d . Without loss of generality, we denote probability distributions $\{\mathbb{P}_1(\cdot | s_i, a), \mathbb{P}_2(\cdot | s_j, a)\}$ as $\{P, Q\}$ for brevity, and let $\rho \leq \rho'$, $\{\rho, \rho'\} \in \mathcal{D}$. Considering the optimal solution $\{\mu, \nu\}$ for $W_1(P, Q; \rho)$ in the dual LP in (4), we have

$$\mu_i - \nu_j \leq \rho(s_i, s_j) \leq \rho'(s_i, s_j), \quad \forall i, j, \quad (28)$$

which is derived from the pointwise ordering in \mathcal{D} . Here, for the other $W_1(P, Q; \rho')$, $\{\mu, \nu\}$ is a feasible, though not necessarily optimal, solution to the dual LP in (4). Thus, we have

$$\begin{aligned} W_1(P, Q; \rho) &= \sum_{i=1}^{|\mathcal{S}_1|} \mu_i P(s_i) - \sum_{j=1}^{|\mathcal{S}_2|} \nu_j Q(s_j) \\ &\leq W_1(P, Q; \rho'), \quad \forall \rho \leq \rho'. \end{aligned} \quad (29)$$

By such a monotonicity, we have $W_1(P, Q; \rho) \leq W_1(P, Q; \sqcup_{n \in \mathbb{N}} \{\rho_n\})$, $\forall \rho \in \{\rho_n\}$ for any increasing sequence $\{\rho_n\}$ on \mathcal{D} . This further implies that $\sqcup_{n \in \mathbb{N}} \{W_1(P, Q; \rho_n)\} \leq W_1(P, Q; \sqcup_{n \in \mathbb{N}} \{\rho_n\})$.

We use the primal LP for the other side. Let λ^n denote the optimal solution in (3) for $W_1(P, Q; \rho_n)$, which also satisfies the conditions for $W_1(P, Q; \sqcup_{n \in \mathbb{N}} \{\rho_n\})$. Define $\epsilon_{i,j}^n = \sqcup_{n \in \mathbb{N}} \{\rho_n\}(s_i, s_j) -$

$\rho_n(s_i, s_j)$, then $\epsilon_{i,j}^n \geq 0$ and $\lim_{n \rightarrow \infty} \epsilon_{i,j}^n = 0$ due to the monotonicity of the increasing sequence of $\{\rho_n\}$. Then, we have

$$\begin{aligned}
W_1(P, Q; \sqcup_{n \in \mathbb{N}} \{\rho_n\}) &\stackrel{(a)}{\leq} \sum_{i=1}^{|\mathcal{S}_1|} \sum_{j=1}^{|\mathcal{S}_2|} \lambda_{i,j}^n \cdot \sqcup_{n \in \mathbb{N}} \{\rho_n\}(s_i, s_j) \\
&= \sum_{i=1}^{|\mathcal{S}_1|} \sum_{j=1}^{|\mathcal{S}_2|} \lambda_{i,j}^n \rho_n(s_i, s_j) + \sum_{i=1}^{|\mathcal{S}_1|} \sum_{j=1}^{|\mathcal{S}_2|} \lambda_{i,j}^n \epsilon_{i,j}^n \\
&= W_1(P, Q; \rho_n) + \sum_{i,j=1}^{|\mathcal{S}|} \lambda_{i,j}^n \epsilon_{i,j}^n \\
&\leq \sqcup_{n \in \mathbb{N}} \{W_1(P, Q; \rho_n)\} + \max_{i,j} \{\epsilon_{i,j}^n\}.
\end{aligned} \tag{30}$$

Here, step (a) follows from the fact that λ^n is the optimal solution for $W_1(P, Q; \rho_n)$ rather than $W_1(P, Q; \sqcup_{n \in \mathbb{N}} \{\rho_n\})$. Taking $n \rightarrow \infty$, we have $\sqcup_{n \in \mathbb{N}} \{W_1(P, Q; \rho_n)\} \geq W_1(P, Q; \sqcup_{n \in \mathbb{N}} \{\rho_n\})$. Following from the above two inequalities from both directions, it is readily to get $\sqcup_{n \in \mathbb{N}} \{W_1(P, Q; \rho_n)\} = W_1(P, Q; \sqcup_{n \in \mathbb{N}} \{\rho_n\})$. Thus, for any i and j ,

$$\begin{aligned}
F_{W_1}(\sqcup_{n \in \mathbb{N}} \{\rho_n\} | s_i, s_j) &= W_1(\mathbb{P}_1(\cdot | s_i, a), \mathbb{P}_2(\cdot | s_j, a); \sqcup_{n \in \mathbb{N}} \{\rho_n\}) \\
&= \sqcup_{n \in \mathbb{N}} \{W_1(\mathbb{P}_1(\cdot | s_i, a), \mathbb{P}_2(\cdot | s_j, a); \rho_n)\} \\
&= \sqcup_{n \in \mathbb{N}} \{F_{W_1}(\rho_n | s_i, s_j)\}.
\end{aligned} \tag{31}$$

Now that the continuity of F_{W_1} in (27) on \mathcal{D} is established. \square

Armed with Lemma A.2, we are ready to establish the continuity of F as follows.

Lemma A.3. *F is continuous on \mathcal{D} .*

Proof. Considering an arbitrary increasing sequence $\{\rho_n\}$ on \mathcal{D} , for any i and j , we have

$$\begin{aligned}
&F(\sqcup_{n \in \mathbb{N}} \{\rho_n\} | s_i, s_j) \\
&= \max_a \{|R_1(s_i, a) - R_2(s_j, a)| + \gamma W_1(\mathbb{P}_1(\cdot | s_i, a), \mathbb{P}_2(\cdot | s_j, a); \sqcup_{n \in \mathbb{N}} \{\rho_n\})\} \\
&= \max_a \{|R_1(s_i, a) - R_2(s_j, a)| + \gamma \sqcup_{n \in \mathbb{N}} \{W_1(\mathbb{P}_1(\cdot | s_i, a), \mathbb{P}_2(\cdot | s_j, a); \rho_n)\}\} \\
&= \sqcup_{n \in \mathbb{N}} \left\{ \max_a \{|R_1(s_i, a) - R_2(s_j, a)| + \gamma W_1(\mathbb{P}_1(\cdot | s_i, a), \mathbb{P}_2(\cdot | s_j, a); \rho_n)\} \right\} \\
&= \sqcup_{n \in \mathbb{N}} \{F(\rho_n | s_i, s_j)\}.
\end{aligned} \tag{32}$$

\square

Now that the existence of d^{1-2} is established by using Lemma A.1 and Lemma A.3.

A.2 The Convergence of d_n^{1-2} to d^{1-2}

Due to the continuity of F and using the induction starting from $d_0^{1-2} \leq d_1^{1-2}$, $\{d_n^{1-2}\}$ forms an increasing sequence on \mathcal{D} . Given that $d^{1-2} = \sqcup_{n \in \mathbb{N}} F^{(n)}(d_0^{1-2})$, we have $d^{1-2} \geq d_n^{1-2}$ for any n . Also,

$$\begin{aligned}
d^{1-2}(s_i, s_j) &= \max_a \left\{ |R_1(s_i, a) - R_2(s_j, a)| + \gamma W_1(\mathbb{P}_1(\cdot | s_i, a), \mathbb{P}_2(\cdot | s_j, a); d^{1-2}) \right\} \\
&\leq \bar{R} + \gamma \max_{i,j} \{d^{1-2}(s_i, s_j)\} \\
&\Rightarrow \max_{i,j} \{d^{1-2}(s_i, s_j)\} \leq \bar{R}/(1 - \gamma), \forall (s_i, s_j) \in \mathcal{S}_1 \times \mathcal{S}_2.
\end{aligned} \tag{33}$$

We begin with a simple inequality for the Wasserstein distance before proving the convergence of GBSM. Let λ^n denote the optimal solution for $W_1(P, Q; d_n^{1-2})$, then for any d_n^{1-2}

$$\begin{aligned}
W_1(P, Q; d^{1-2}) &\leq \sum_{i=1}^{|\mathcal{S}_1|} \sum_{j=1}^{|\mathcal{S}_2|} \lambda_{i,j}^n d^{1-2}(s_i, s_j) \\
&= \sum_{i=1}^{|\mathcal{S}_1|} \sum_{j=1}^{|\mathcal{S}_2|} \lambda_{i,j}^n (d^{1-2}(s_i, s_j) - d_n^{1-2}(s_i, s_j) + d_n^{1-2}(s_i, s_j)) \\
&\leq \max_{i,j} \{d^{1-2}(s_i, s_j) - d_n^{1-2}(s_i, s_j)\} + W_1(P, Q; d_n^{1-2}).
\end{aligned} \tag{34}$$

The first inequality follows from the fact that λ^n is the optimal solution for $W_1(P, Q; d_n^{1-2})$ rather than $W_1(P, Q; d^{1-2})$.

Now we employ the mathematical induction. For the base case, we have

$$\begin{aligned}
& d^{1-2}(s, s') - d_1^{1-2}(s, s') \\
&= \max_a \{ |R_1(s, a) - R_2(s', a)| + \gamma W_1(\mathbb{P}_1(\cdot|s, a), \mathbb{P}_2(\cdot|s', a); d^{1-2}) \} \\
&\quad - \max_a \{ |R_1(s, a) - R_2(s', a)| \} \\
&\leq \max_a \{ |R_1(s, a) - R_2(s', a)| \} + \gamma \max_a \{ W_1(\mathbb{P}_1(\cdot|s, a), \mathbb{P}_2(\cdot|s', a); d^{1-2}) \} \\
&\quad - \max_a \{ |R_1(s, a) - R_2(s', a)| \} \\
&= \gamma \max_a \{ W_1(\mathbb{P}_1(\cdot|s, a), \mathbb{P}_2(\cdot|s', a); d^{1-2}) \} \\
&\leq \gamma \max_{s, s'} \{ d^{1-2}(s, s') \} = \gamma \bar{R}/(1 - \gamma), \quad \forall (s, s') \in \mathcal{S}_1 \times \mathcal{S}_2.
\end{aligned} \tag{35}$$

By the induction hypothesis, we assume that for an arbitrary n ,

$$d^{1-2}(s, s') - d_n^{1-2}(s, s') \leq \gamma^n \bar{R}/(1 - \gamma), \quad \forall (s, s') \in \mathcal{S}_1 \times \mathcal{S}_2. \tag{36}$$

Then we have

$$\begin{aligned}
& d^{1-2}(s, s') - d_{n+1}^{1-2}(s, s') \\
&= \max_a \{ |R_1(s, a) - R_2(s', a)| + \gamma W_1(\mathbb{P}_1(\cdot|s, a), \mathbb{P}_2(\cdot|s', a); d^{1-2}) \} \\
&\quad - \max_a \{ |R_1(s, a) - R_2(s', a)| + \gamma W_1(\mathbb{P}_1(\cdot|s, a), \mathbb{P}_2(\cdot|s', a); d_n^{1-2}) \} \\
&\leq \max_a \{ (|R_1(s, a) - R_2(s', a)| + \gamma W_1(\mathbb{P}_1(\cdot|s, a), \mathbb{P}_2(\cdot|s', a); d^{1-2})) \\
&\quad - (|R_1(s, a) - R_2(s', a)| + \gamma W_1(\mathbb{P}_1(\cdot|s, a), \mathbb{P}_2(\cdot|s', a); d_n^{1-2})) \} \\
&= \gamma \max_a \{ W_1(\mathbb{P}_1(\cdot|s, a), \mathbb{P}_2(\cdot|s', a); d^{1-2}) - W_1(\mathbb{P}_1(\cdot|s, a), \mathbb{P}_2(\cdot|s', a); d_n^{1-2}) \} \\
&\stackrel{(a)}{\leq} \gamma \max_a \{ \max_{s, s'} \{ d^{1-2}(s, s') - d_n^{1-2}(s, s') \} + W_1(\mathbb{P}_1(\cdot|s, a), \mathbb{P}_2(\cdot|s', a); d_n^{1-2}) \\
&\quad - W_1(\mathbb{P}_1(\cdot|s, a), \mathbb{P}_2(\cdot|s', a); d_n^{1-2}) \} \\
&= \gamma \max_{s, s'} \{ d^{1-2}(s, s') - d_n^{1-2}(s, s') \} \leq \gamma^{n+1} \bar{R}/(1 - \gamma), \quad \forall (s, s') \in \mathcal{S}_1 \times \mathcal{S}_2.
\end{aligned} \tag{37}$$

Here, step (a) uses (34). Following from (35)-(37), $d^{1-2}(s, s') - d_n^{1-2}(s, s') \leq \gamma^n \bar{R}/(1 - \gamma)$ holds for all $n \in \mathbb{N}$.

B Proof of Theorem 3.3

This proves the optimal value difference bound between MDPs by induction. For the base case, we have

$$\begin{aligned}
|V_1^{(1)}(s_i) - V_2^{(1)}(s_j)| &= |\max_a R_1(s_i, a) - \max_a R_2(s_j, a)| \\
&\leq \max_a |R_1(s_i, a) - R_2(s_j, a)| \\
&= d_1^{1-2}(s_i, s_j), \quad \forall (s_i, s_j) \in \mathcal{S}_1 \times \mathcal{S}_2.
\end{aligned} \tag{38}$$

By the induction hypothesis, we assume that for an arbitrary n ,

$$\begin{aligned}
V_1^{(n)}(s_i) - V_2^{(n)}(s_j) &\leq |V_1^{(n)}(s_i) - V_2^{(n)}(s_j)| \\
&\leq d_n^{1-2}(s_i, s_j), \quad \forall (s_i, s_j) \in \mathcal{S}_1 \times \mathcal{S}_2.
\end{aligned} \tag{39}$$

Then the induction follows

$$\begin{aligned}
& \left| V_1^{(n+1)}(s_i) - V_2^{(n+1)}(s_j) \right| \\
&= \left| \max_a \left\{ R_1(s_i, a) + \gamma \sum_{k=1}^{|\mathcal{S}_1|} \mathbb{P}_1(s_k | s_i, a) V_1^{(n)}(s_k) \right\} \right. \\
&\quad \left. - \max_a \left\{ R_2(s_j, a) + \gamma \sum_{k=1}^{|\mathcal{S}_2|} \mathbb{P}_2(s_k | s_j, a) V_2^{(n)}(s_k) \right\} \right| \\
&\leq \max_a \left\{ \left| \left(R_1(s_i, a) + \gamma \sum_{k=1}^{|\mathcal{S}_1|} \mathbb{P}_1(s_k | s_i, a) V_1^{(n)}(s_k) \right) \right. \right. \\
&\quad \left. \left. - \left(R_2(s_j, a) + \gamma \sum_{k=1}^{|\mathcal{S}_2|} \mathbb{P}_2(s_k | s_j, a) V_2^{(n)}(s_k) \right) \right| \right\} \\
&\leq \max_a \left\{ \left| R_1(s_i, a) - R_2(s_j, a) \right| + \gamma \left| \sum_{k=1}^{|\mathcal{S}_1|} \mathbb{P}_1(s_k | s_i, a) V_1^{(n)}(s_k) - \sum_{k=1}^{|\mathcal{S}_2|} \mathbb{P}_2(s_k | s_j, a) V_2^{(n)}(s_k) \right| \right\} \\
&\stackrel{(a)}{\leq} \max_a \{ |R_1(s_i, a) - R_2(s_j, a)| + \gamma W_1(\mathbb{P}_1(\cdot | s_i, a), \mathbb{P}_2(\cdot | s_j, a); d_n^{1-2}) \} \\
&= d_{n+1}^{1-2}(s_i, s_j), \quad \forall (s_i, s_j) \in \mathcal{S}_1 \times \mathcal{S}_2. \tag{40}
\end{aligned}$$

Here, steps (a) follows from the fact that $(V_1^{(n)}(s_k))_{k=1}^{|\mathcal{S}_1|}$ and $(V_2^{(n)}(s_k))_{k=1}^{|\mathcal{S}_2|}$ form a feasible, but not necessarily the optimal, solution to the dual LP in (4) for $W_1(\mathbb{P}_1(\cdot | s_i, a), \mathbb{P}_2(\cdot | s_j, a); d_n^{1-2})$.

Now from (38)-(40), we have $|V_1^{(n)}(s) - V_2^{(n)}(s')| \leq d_n^{1-2}(s, s'), \quad \forall (s, s') \in \mathcal{S}_1 \times \mathcal{S}_2, \quad \forall n \in \mathbb{N}$. Taking $n \rightarrow \infty$ yields the desired result.

C Proof of Theorem 4.1

This section provides a detailed proof of the regret bound for policy π transferred from \mathcal{M}_1 to \mathcal{M}_2 . By the triangle inequality, for any state $s_j \in \mathcal{S}_2$ and $s_i = f(s_j) \in \mathcal{S}_1$, we have

$$|V_2^*(s_j) - V_2^\pi(s_j)| \leq |V_2^*(s_j) - V_1^*(s_i)| + |V_1^*(s_i) - V_1^\pi(s_i)| + |V_1^\pi(s_i) - V_2^\pi(s_j)|. \tag{41}$$

Within the right-hand side of this inequality, the first summation term $|V_2^*(s_j) - V_1^*(s_i)|$ is upper bounded by $d^{1-2}(s_i, s_j)$ according to Theorem 3.3, and $|V_1^*(s_i) - V_1^\pi(s_i)|$ is upper bounded by $\max_{s \in \mathcal{S}_1} |V_1^*(s) - V_1^\pi(s)|$. For the last term, we have

$$\begin{aligned}
& |V_1^\pi(s_i) - V_2^\pi(s_j)| \\
&= \left| \sum_{a=1}^{|\mathcal{A}|} \pi(a | s_i) \left(R_1(s_i, a) + \gamma \sum_{k=1}^{|\mathcal{S}_1|} \mathbb{P}_1(s_k | s_i, a) V_1^\pi(s_k) \right) \right. \\
&\quad \left. - \sum_{a=1}^{|\mathcal{A}|} \pi(a | f(s_j)) \left(R_2(s_j, a) + \gamma \sum_{k=1}^{|\mathcal{S}_2|} \mathbb{P}_2(s_k | s_j, a) V_2^\pi(s_k) \right) \right| \\
&\leq \sum_{a=1}^{|\mathcal{A}|} \pi(a | s_i) \left(\left| R_1(s_i, a) - R_2(s_j, a) \right| + \gamma \left| \sum_{k=1}^{|\mathcal{S}_1|} \mathbb{P}_1(s_k | s_i, a) V_1^\pi(s_k) - \sum_{k=1}^{|\mathcal{S}_2|} \mathbb{P}_2(s_k | s_j, a) V_2^\pi(s_k) \right| \right) \\
&\leq \sum_{a=1}^{|\mathcal{A}|} \pi(a | s_i) \left(\left| R_1(s_i, a) - R_2(s_j, a) \right| + \gamma \left| \sum_{k=1}^{|\mathcal{S}_1|} \mathbb{P}_1(s_k | s_i, a) V_1^*(s_k) - \sum_{k=1}^{|\mathcal{S}_2|} \mathbb{P}_2(s_k | s_j, a) V_2^*(s_k) \right| \right) \\
&\quad + \gamma \sum_{a=1}^{|\mathcal{A}|} \pi(a | s_i) \left| \sum_{k=1}^{|\mathcal{S}_1|} \mathbb{P}_1(s_k | s_i, a) (V_1^*(s_k) - V_1^\pi(s_k)) \right| \\
&\quad + \gamma \sum_{a=1}^{|\mathcal{A}|} \pi(a | s_i) \left| \sum_{k=1}^{|\mathcal{S}_2|} \mathbb{P}_2(s_k | s_j, a) (V_2^*(s_k) - V_2^\pi(s_k)) \right|
\end{aligned}$$

$$\begin{aligned}
&\leq \max_a \left\{ \left| R_1(s_i, a) - R_2(s_j, a) \right| + \gamma \left| \sum_{k=1}^{|S_1|} \mathbb{P}_1(s_k | s_i, a) V_1^*(s_k) - \sum_{k=1}^{|S_2|} (\mathbb{P}(s_k | s_j, a) V_2^*(s_k)) \right| \right\} \\
&\quad + \gamma \max_{s \in \mathcal{S}_1} |V_1^*(s) - V_1^\pi(s)| + \gamma \max_{s \in \mathcal{S}_2} |V_2^*(s) - V_2^\pi(s)| \\
&\stackrel{(a)}{\leq} \max_a \{ |R_1(s_i, a) - R_2(s_j, a)| + \gamma W_1(\mathbb{P}_1(\cdot | s_i, a), \mathbb{P}_2(\cdot | s_j, a); d^{1-2}) \} \\
&\quad + \gamma \max_{s \in \mathcal{S}_1} |V_1^*(s) - V_1^\pi(s)| + \gamma \max_{s \in \mathcal{S}_2} |V_2^*(s) - V_2^\pi(s)| \\
&= d^{1-2}(s_i, s_j) + \gamma \max_{s \in \mathcal{S}_1} |V_1^*(s) - V_1^\pi(s)| + \gamma \max_{s \in \mathcal{S}_2} |V_2^*(s) - V_2^\pi(s)|. \tag{42}
\end{aligned}$$

Here, step (a) stems from the fact that, according to Theorem 3.3, $(V_1^*(s_k))_{k=1}^{|S_1|}$ and $(V_2^*(s_k))_{k=1}^{|S_2|}$ form a feasible, but not necessarily the optimal, solution to the dual LP in (4) for $W_1(\mathbb{P}_1(\cdot | s_i, a), \mathbb{P}_2(\cdot | s_j, a); d^{1-2})$. Combining the above inequalities on all three summation terms in (41) and taking the maximum of both sides, we have

$$\begin{aligned}
\max_{s' \in \mathcal{S}_2} |V_2^*(s') - V_2^\pi(s')| &\leq \underbrace{\max_{s' \in \mathcal{S}_2} d^{1-2}(f(s'), s')}_{\text{1st term}} + \underbrace{\max_{s \in \mathcal{S}_1} |V_1^*(s) - V_1^\pi(s)|}_{\text{2nd term}} \\
&\quad + \underbrace{\max_{s' \in \mathcal{S}_2} d^{1-2}(f(s'), s') + \gamma \max_{s \in \mathcal{S}_1} |V_1^*(s) - V_1^\pi(s)| + \gamma \max_{s' \in \mathcal{S}_2} |V_2^*(s') - V_2^\pi(s')|}_{\text{3rd term}} \\
&\leq 2 \max_{s \in \mathcal{S}_2} d^{1-2}(f(s), s) + (1 + \gamma) \max_{s \in \mathcal{S}_1} |V_1^*(s) - V_1^\pi(s)| + \gamma \max_{s' \in \mathcal{S}_2} |V_2^*(s') - V_2^\pi(s')|. \tag{43}
\end{aligned}$$

Rearranging the inequality yields the desired result.

D Proofs of Lax GBSM Properties

Definition D.1 (Lax generalized bisimulation metric). Given two MDPs $\mathcal{M}_1 = \langle \mathcal{S}_1, \mathcal{A}_1, \mathbb{P}_1, R_1, \gamma \rangle$ and $\mathcal{M}_2 = \langle \mathcal{S}_2, \mathcal{A}_2, \mathbb{P}_2, R_2, \gamma \rangle$, we first define an intermediate metric as

$$\delta(d)((s, a), (s', a')) = |R_1(s, a) - R_2(s', a')| + \gamma W_1(\mathbb{P}_1(\cdot | s, a), \mathbb{P}_2(\cdot | s', a'); d), \tag{44}$$

and define the lax function as

$$F_{\text{lax}}(d | s, s') = H(X_s, X_{s'}; \delta(d)), \tag{45}$$

where $X_s = \{(s, a) | a \in \mathcal{A}_1\}$, $X_{s'} = \{(s', a') | a' \in \mathcal{A}_2\}$, and H is the Hausdorff metric defined by $H(X, Y; d) = \max \{ \sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y) \}$. Iterating from $d_{\text{lax}, 0}^{1-2}(s, s') = 0$ and $d_{\text{lax}, n+1}^{1-2} = F_{\text{lax}}(d_{\text{lax}, n}^{1-2} | s, s')$, $d_{\text{lax}, n}^{1-2}$ converges to a similar fixed point d_{lax}^{1-2} with $n \rightarrow \infty$.

The proof of the existence and convergence of Lax GBSM is quite similar to the one for GBSM in Appendix A. We omit it and mainly prove its core property, i.e., the optimal value difference bound between MDPs, and its tightness compared to GBSM in this section.

Theorem D.2 (Lax GBSM optimal value difference bound). Let V_1^* and V_2^* denote the optimal value functions in \mathcal{M}_1 and \mathcal{M}_2 , respectively. Then lax GBSM provides an upper bound for the difference between the optimal values for any state pair $(s, s') \in \mathcal{S}_1 \times \mathcal{S}_2$:

$$|V_1^*(s) - V_2^*(s')| \leq d_{\text{lax}}^{1-2}(s, s'). \tag{46}$$

Proof. For the base case, we have

$$|V_1^{(0)}(s_i) - V_2^{(0)}(s_j)| = d_{\text{lax}, 0}^{1-2}(s_i, s_j) = 0, \forall (s_i, s_j) \in \mathcal{S}_1 \times \mathcal{S}_2. \tag{47}$$

By the induction hypothesis, we assume that for an arbitrary n ,

$$|V_1^{(n)}(s_i) - V_2^{(n)}(s_j)| \leq d_{\text{lax}, n}^{1-2}(s_i, s_j), \forall (s_i, s_j) \in \mathcal{S}_1 \times \mathcal{S}_2. \tag{48}$$

Without loss of generality, assume that $V_1^{(n+1)}(s_i) \geq V_2^{(n+1)}(s_j)$, and the induction follows

$$\begin{aligned}
& \left| V_1^{(n+1)}(s_i) - V_2^{(n+1)}(s_j) \right| \\
&= \left| \max_a \left\{ R_1(s_i, a) + \gamma \sum_{k=1}^{|\mathcal{S}_1|} \mathbb{P}_1(s_k | s_i, a) V_1^{(n)}(s_k) \right\} \right. \\
&\quad \left. - \max_{a'} \left\{ R_2(s_j, a') + \gamma \sum_{k=1}^{|\mathcal{S}_2|} \mathbb{P}_2(s_k | s_j, a') V_2^{(n)}(s_k) \right\} \right| \\
&= \left| \left\{ R_1(s_i, a_{\max}) + \gamma \sum_{k=1}^{|\mathcal{S}_1|} \mathbb{P}_1(s_k | s_i, a_{\max}) V_1^{(n)}(s_k) \right\} \right. \\
&\quad \left. - \left\{ R_2(s_j, a'_{\max}) + \gamma \sum_{k=1}^{|\mathcal{S}_2|} \mathbb{P}_2(s_k | s_j, a'_{\max}) V_2^{(n)}(s_k) \right\} \right| \\
&= \min_{a'} \left| \left\{ R_1(s_i, a_{\max}) + \gamma \sum_{k=1}^{|\mathcal{S}_1|} \mathbb{P}_1(s_k | s_i, a_{\max}) V_1^{(n)}(s_k) \right\} \right. \\
&\quad \left. - \left\{ R_2(s_j, a') + \gamma \sum_{k=1}^{|\mathcal{S}_2|} \mathbb{P}_2(s_k | s_j, a') V_2^{(n)}(s_k) \right\} \right| \\
&\leq \max_a \min_{a'} \left| \left\{ R_1(s_i, a) + \gamma \sum_{k=1}^{|\mathcal{S}_1|} \mathbb{P}_1(s_k | s_i, a) V_1^{(n)}(s_k) \right\} \right. \\
&\quad \left. - \left\{ R_2(s_j, a') + \gamma \sum_{k=1}^{|\mathcal{S}_2|} \mathbb{P}_2(s_k | s_j, a') V_2^{(n)}(s_k) \right\} \right| \\
&\leq \max_a \min_{a'} \left\{ \left| R_1(s_i, a) - R_2(s_j, a') \right| + \gamma \left| \sum_{k=1}^{|\mathcal{S}_1|} \mathbb{P}_1(s_k | s_i, a) V_1^{(n)}(s_k) - \sum_{k=1}^{|\mathcal{S}_2|} \mathbb{P}_2(s_k | s_j, a') V_2^{(n)}(s_k) \right| \right\} \\
&\leq \max_a \min_{a'} \{ |R_1(s_i, a) - R_2(s_j, a')| + \gamma W_1(\mathbb{P}_1(\cdot | s_i, a), \mathbb{P}_2(\cdot | s_j, a'); d_n^{1-2}) \} \\
&\leq d_{\text{lax}, n+1}^{1-2}(s_i, s_j), \quad \forall (s_i, s_j) \in \mathcal{S}_1 \times \mathcal{S}_2. \tag{49}
\end{aligned}$$

Taking $n \rightarrow \infty$ yields the desired result. \square

Theorem D.3 (Inequality between GBSM and lax GBSM). When \mathcal{M}_1 and \mathcal{M}_2 share the same \mathcal{A} ,

$$d_{\text{lax}}^{1-2}(s, s') \leq d^{1-2}(s, s'). \tag{50}$$

Proof. For the base case, we have

$$d_{\text{lax}, 0}^{1-2}(s, s') = d_0^{1-2}(s, s') = 0, \quad \forall (s, s') \in \mathcal{S}_1 \times \mathcal{S}_2. \tag{51}$$

By the induction hypothesis, we assume that for an arbitrary n ,

$$d_{\text{lax}, n}^{1-2}(s, s') \leq d_n^{1-2}(s, s'), \quad \forall (s, s') \in \mathcal{S}_1 \times \mathcal{S}_2. \tag{52}$$

By the continuity of F_{lax} , we have that $F_{\text{lax}}(d_{\text{lax}, n}^{1-2} | s, s') \leq F_{\text{lax}}(d_n^{1-2} | s, s')$, which implies

$$\begin{aligned}
d_{\text{lax}, n+1}^{1-2}(s, s') &\leq \max \left\{ \max_a \min_{a'} \delta(d_n^{1-2})((s, a), (s', a')), \max_{a'} \min_a \delta(d_n^{1-2})((s, a), (s', a')) \right\} \\
&\leq \max_a \delta(d_n^{1-2})((s, a), (s', a)) = d_{n+1}^{1-2}(s, s'), \quad \forall (s, s') \in \mathcal{S}_1 \times \mathcal{S}_2. \tag{53}
\end{aligned}$$

Taking $n \rightarrow \infty$ yields the desired result. \square

Using the above two theorems and the definition of Wasserstein distance, we derive similar metric properties as

$$d_{\text{lax}}^{1-2}(s, s') = d_{\text{lax}}^{2-1}(s', s), \quad \forall (s, s') \in \mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{S}_3, \quad (\text{Theorem 3.4}), \tag{54}$$

$$d_{\text{lax}}^{1-2}(s, s') \leq d_{\text{lax}}^{1-3}(s, s'') + d_{\text{lax}}^{3-2}(s'', s'), \quad \forall (s, s', s'') \in \mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{S}_3, \quad (\text{Theorem 3.5}), \tag{55}$$

$$\text{and } d_{\text{lax}}^{1-2} \leq d^{1-2} \leq d_{\text{TV}}^{1-2} / (1 - \gamma) \text{ when } \mathcal{S}_1 = \mathcal{S}_2 \text{ and } \mathcal{A}_1 = \mathcal{A}_2. \quad (\text{Theorem 3.6}) \tag{56}$$

Since these fundamental metric properties hold, the bounds for state aggregation (Theorem 4.5) and estimation (Theorem 4.6) also follow directly. In terms of policy transfer (Theorem 4.1), besides the state mapping $f : \mathcal{S}_2 \rightarrow \mathcal{S}_1$, we need to define an additional action mapping $g : \mathcal{A}_1 \rightarrow \mathcal{A}_2$ for policy transfer between different action spaces. Then, we have

$$\begin{aligned}
& |V_1^\pi(s_i) - V_2^\pi(s_j)| \\
&= \left| \sum_{a=1}^{|\mathcal{A}_1|} \pi(a|s_i) (R_1(s_i, a) + \gamma \sum_{k=1}^{|\mathcal{S}_1|} \mathbb{P}_1(s_k|s_i, a) V_1^\pi(s_k)) \right. \\
&\quad \left. - \sum_{a=1}^{|\mathcal{A}_1|} \pi(a|f(s_j)) (R_2(s_j, g(a)) + \gamma \sum_{k=1}^{|\mathcal{S}_2|} \mathbb{P}_2(s_k|s_j, g(a)) V_2^\pi(s_k)) \right| \\
&\leq \sum_{a=1}^{|\mathcal{A}_1|} \left(\pi(a|s_i) (|R_1(s_i, a) - R_2(s_j, g(a))| \right. \\
&\quad \left. + \gamma \left| \sum_{k=1}^{|\mathcal{S}_1|} \mathbb{P}_1(s_k|s_i, a) V_1^\pi(s_k) - \sum_{k=1}^{|\mathcal{S}_2|} \mathbb{P}_2(s_k|s_j, g(a)) V_2^\pi(s_k) \right| \right) \\
&\leq \sum_{a=1}^{|\mathcal{A}_1|} \left(\pi(a|s_i) (|R_1(s_i, a) - R_2(s_j, g(a))| \right. \\
&\quad \left. + \gamma \left| \sum_{k=1}^{|\mathcal{S}_1|} \mathbb{P}_1(s_k|s_i, a) V_1^*(s_k) - \sum_{k=1}^{|\mathcal{S}_2|} \mathbb{P}_2(s_k|s_j, g(a)) V_2^*(s_k) \right| \right) \\
&\quad + \gamma \sum_{a=1}^{|\mathcal{A}_1|} \pi(a|s_i) \left| \sum_{k=1}^{|\mathcal{S}_1|} \mathbb{P}_1(s_k|s_i, a) (V_1^*(s_k) - V_1^\pi(s_k)) \right| \\
&\quad + \gamma \sum_{a=1}^{|\mathcal{A}_1|} \pi(a|s_i) \left| \sum_{k=1}^{|\mathcal{S}_2|} \mathbb{P}_2(s_k|s_j, g(a)) (V_2^*(s_k) - V_2^\pi(s_k)) \right| \\
&\leq \max_{a \in \mathcal{A}_1} \left\{ |R_1(s_i, a) - R_2(s_j, g(a))| \right. \\
&\quad \left. + \gamma \left| \sum_{k=1}^{|\mathcal{S}_1|} \mathbb{P}_1(s_k|s_i, a) V_1^*(s_k) - \sum_{k=1}^{|\mathcal{S}_2|} \mathbb{P}_2(s_k|s_j, g(a)) V_2^*(s_k) \right| \right\} \\
&\quad + \gamma \max_{s \in \mathcal{S}_1} |V_1^*(s) - V_1^\pi(s)| + \gamma \max_{s \in \mathcal{S}_2} |V_2^*(s) - V_2^\pi(s)| \\
&\stackrel{(a)}{\leq} \max_{a \in \mathcal{A}_1} \min_{a' \in \mathcal{A}_2} \{ |R_1(s_i, a) - R_2(s_j, a')| + \gamma W_1(\mathbb{P}_1(\cdot|s_i, a), \mathbb{P}_2(\cdot|s_j, a'); d_{\text{lax}}^{1-2}) \} \\
&\quad + \gamma \max_{s \in \mathcal{S}_1} |V_1^*(s) - V_1^\pi(s)| + \gamma \max_{s \in \mathcal{S}_2} |V_2^*(s) - V_2^\pi(s)| \\
&\leq H(\delta(d_{\text{lax}}^{1-2}))(X_{s_i}, X'_{s_j}) + \gamma \max_{s \in \mathcal{S}_1} |V_1^*(s) - V_1^\pi(s)| + \gamma \max_{s \in \mathcal{S}_2} |V_2^*(s) - V_2^\pi(s)| \\
&= d_{\text{lax}}^{1-2}(s_i, s_j) + \gamma \max_{s \in \mathcal{S}_1} |V_1^*(s) - V_1^\pi(s)| + \gamma \max_{s \in \mathcal{S}_2} |V_2^*(s) - V_2^\pi(s)|
\end{aligned}$$

Due to the introduction of max-min term via Hausdorff metric, step (a) requires that action mapping satisfies $g(a) = \arg \min_{a'} \delta((f(s'), a), (s', a'); d_{\text{lax}}^{1-2})$ for each s' and a .

E Proofs of on-policy GBSM Properties

Theorem E.1 (On-policy GBSM optimal value difference bound). *Let V_1^π and V_2^π denote the value functions with policy π in \mathcal{M}_1 and \mathcal{M}_2 , respectively. Then lax GBSM provides an upper bound for the difference between the value functions for any state pair $(s, s') \in \mathcal{S}_1 \times \mathcal{S}_2$:*

$$|V_1^\pi(s) - V_2^\pi(s')| \leq d_{\pi}^{1-2}(s, s'). \quad (57)$$

Proof. For the base case, we have

$$|V_1^{\pi, (0)}(s_i) - V_2^{\pi, (0)}(s_j)| = d_{\pi, 0}^{1-2}(s_i, s_j) = 0, \forall (s_i, s_j) \in \mathcal{S}_1 \times \mathcal{S}_2. \quad (58)$$

By the induction hypothesis, we assume that for an arbitrary n ,

$$|V_1^{\pi, (n)}(s_i) - V_2^{\pi, (n)}(s_j)| \leq d_{\pi, n}^{1-2}(s_i, s_j), \forall (s_i, s_j) \in \mathcal{S}_1 \times \mathcal{S}_2. \quad (59)$$

The induction follows

$$\begin{aligned}
& |V_1^{\pi, (n+1)}(s_i) - V_2^{\pi, (n+1)}(s_j)| \\
&= \left| \{R_1^\pi(s_i) + \gamma \sum_{k=1}^{|\mathcal{S}_1|} \mathbb{P}_1^\pi(s_k|s_i) V_1^{\pi, (n)}(s_k)\} - \{R_2^\pi(s_j) + \gamma \sum_{k=1}^{|\mathcal{S}_2|} \mathbb{P}_2^\pi(s_k|s_j) V_2^{\pi, (n)}(s_k)\} \right| \\
&\leq \left\{ |R_1^\pi(s_i) - R_2^\pi(s_j)| + \gamma \left| \sum_{k=1}^{|\mathcal{S}_1|} \mathbb{P}_1^\pi(s_k|s_i) V_2^{\pi, (n)}(s_k) - \sum_{k=1}^{|\mathcal{S}_2|} \mathbb{P}_2^\pi(s_k|s_j) V_2^{\pi, (n)}(s_k) \right| \right\} \\
&\leq \left\{ |R_1^\pi(s_i) - R_2^\pi(s_j)| + \gamma W_1(\mathbb{P}_1^\pi(\cdot|s_i), \mathbb{P}_2^\pi(\cdot|s_j); d_n^{1-2}) \right\} \\
&= d_{n+1}^{1-2}(s_i, s_j), \quad \forall (s_i, s_j) \in \mathcal{S}_1 \times \mathcal{S}_2.
\end{aligned} \tag{60}$$

Taking $n \rightarrow \infty$ yields the desired result. \square

Theorem E.2 (On-policy GBSM distance bound on identical state spaces). *When \mathcal{M}_1 and \mathcal{M}_2 share the same \mathcal{S} ,*

$$\max_s d_\pi^{1-2}(s, s) \leq \frac{1}{1-\gamma} \max_{s,a} \left\{ |R_1^\pi(s) - R_2^\pi(s)| + \frac{\gamma \bar{R}}{1-\gamma} \text{TV}(\mathbb{P}_1^\pi(\cdot|s), \mathbb{P}_2^\pi(\cdot|s)) \right\}, \tag{61}$$

where TV represents the total variation distance defined by $\text{TV}(P, Q) = \frac{1}{2} \sum_{s \in \mathcal{S}} |P(s) - Q(s)|$.

Proof. Using inequality (12) in Theorem 3.6, we have

$$\begin{aligned}
\max_s d_\pi^{1-2}(s, s) &\leq |R_1^\pi(s) - R_2^\pi(s)| + \gamma \text{TV}(\mathbb{P}_1^\pi(\cdot|s), \mathbb{P}_2^\pi(\cdot|s)) \max_{s,s'} d_\pi^{1-2}(s, s') \\
&\quad + \gamma (1 - \text{TV}(\mathbb{P}_1^\pi(\cdot|s), \mathbb{P}_2^\pi(\cdot|s))) \max_s d_\pi^{1-2}(s, s) \\
&\leq |R_1^\pi(s) - R_2^\pi(s)| + \frac{\gamma \bar{R}}{1-\gamma} \text{TV}(\mathbb{P}_1^\pi(\cdot|s), \mathbb{P}_2^\pi(\cdot|s)) + \gamma \max_s d_\pi^{1-2}(s, s).
\end{aligned}$$

Rearranging the inequality yields the desired result. \square

Theorem E.3 (VFA error bound with non-optimal policy).

$$\max_s |V_1^\pi(s) - V_{[1]}^\pi(s)| \leq \max_s d_\pi^{1-[1]}(s, s) \leq \max_s \tilde{d}_\pi(s, [s]) / (1-\gamma) \tag{62}$$

Proof. The first inequality follows directly from Theorem E.1, while the second is established using a derivation analogous to the proof of Theorem 4.4. \square

F Additional Numerical Results

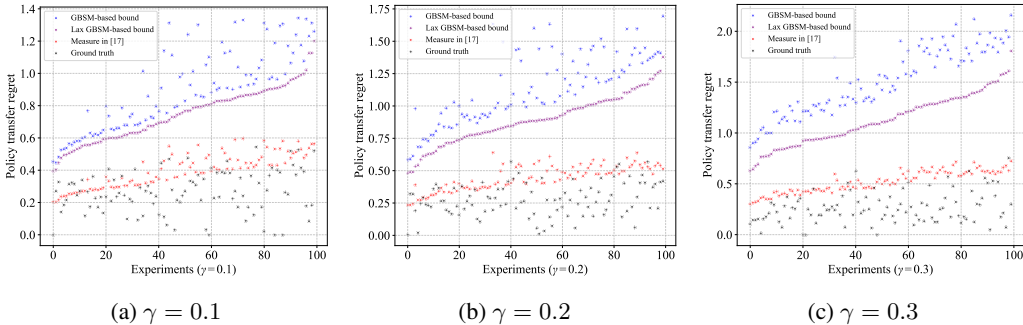


Figure 2: Experiments on random Garnet MDPs (policy transfer, $\gamma = 0.1$ to 0.3).

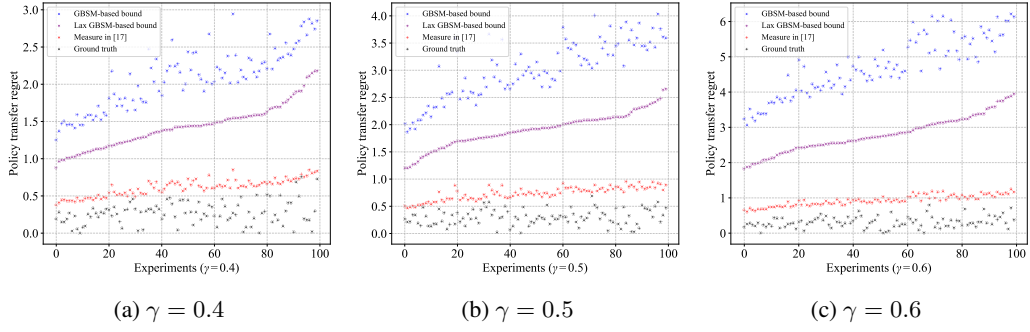


Figure 3: Experiments on random Garnet MDPs (policy transfer, $\gamma = 0.4$ to 0.6).

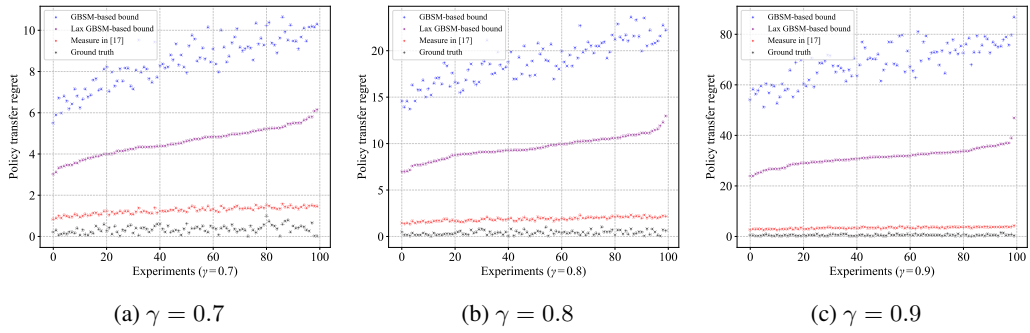


Figure 4: Experiments on random Garnet MDPs (policy transfer, $\gamma = 0.7$ to 0.9).

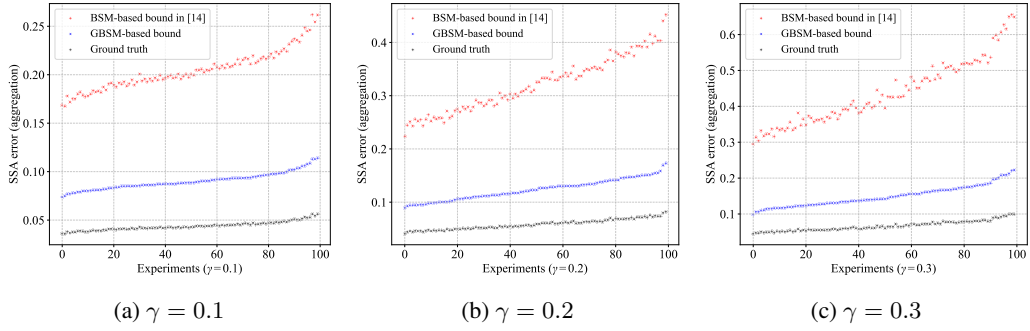


Figure 5: Experiments on random Garnet MDPs (SSA with aggregation, $\gamma = 0.1$ to 0.3).

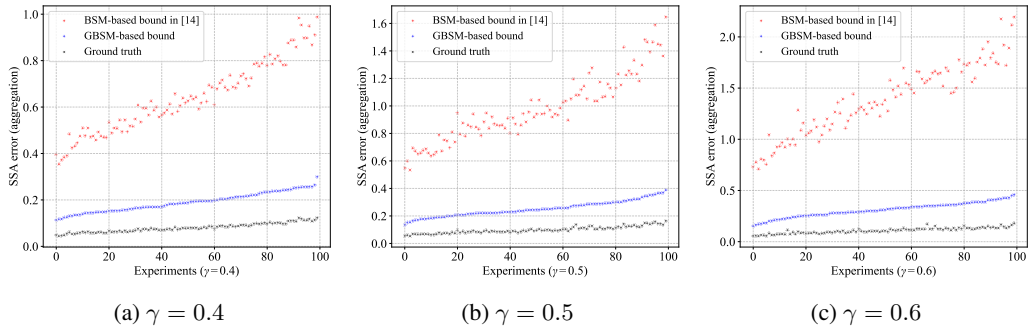


Figure 6: Experiments on random Garnet MDPs (SSA with aggregation, $\gamma = 0.4$ to 0.6).

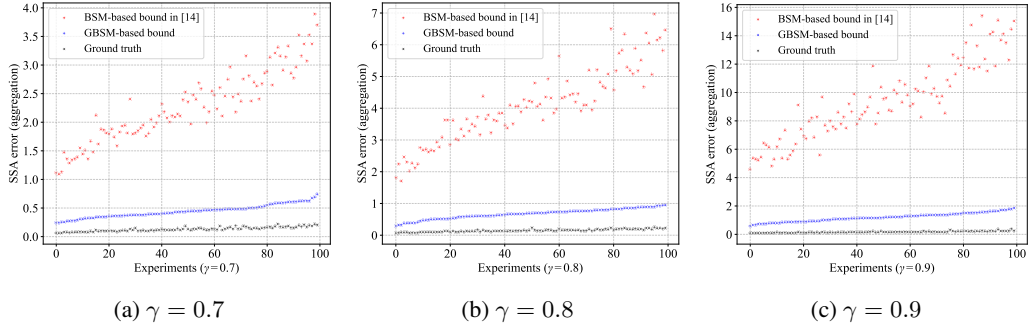


Figure 7: Experiments on random Garnet MDPs (SSA aggregation, $\gamma = 0.7$ to 0.9).

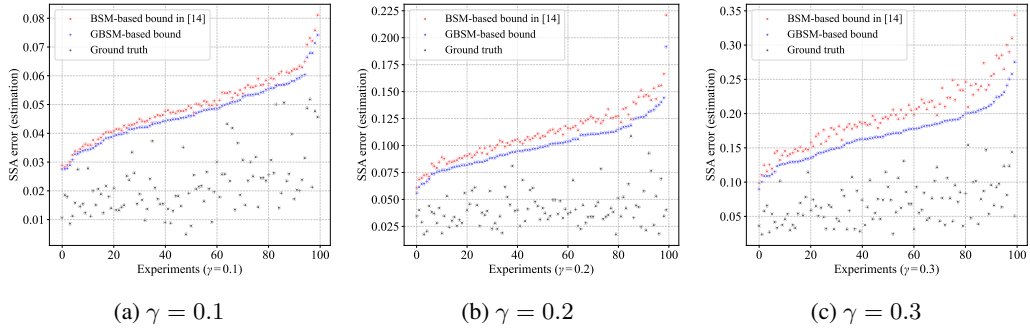


Figure 8: Experiments on random Garnet MDPs (SSA with estimation, $\gamma = 0.1$ to 0.3).

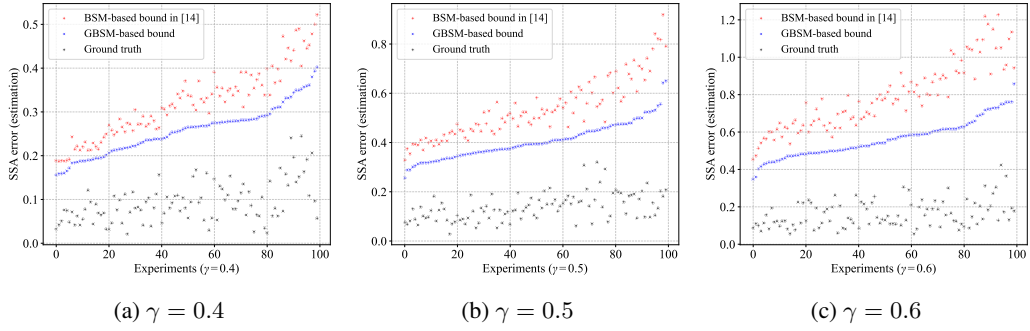


Figure 9: Experiments on random Garnet MDPs (SSA with aggregation, $\gamma = 0.4$ to 0.6).

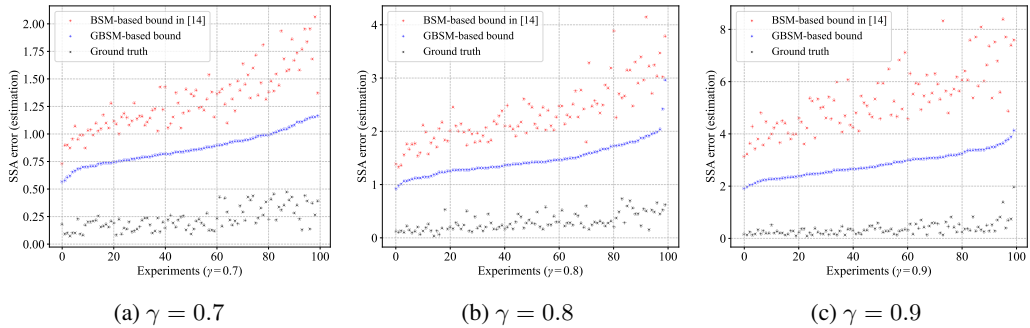


Figure 10: Experiments on random Garnet MDPs (SSA with aggregation, $\gamma = 0.7$ to 0.9).

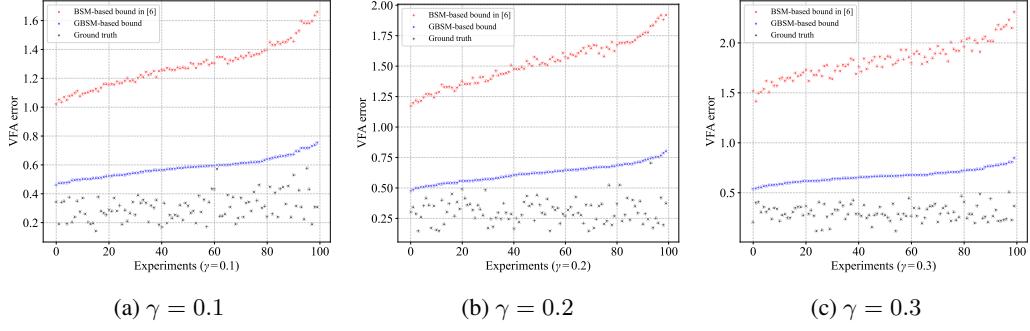


Figure 11: Experiments on random Garnet MDPs (VFA, $\gamma = 0.1$ to 0.3).

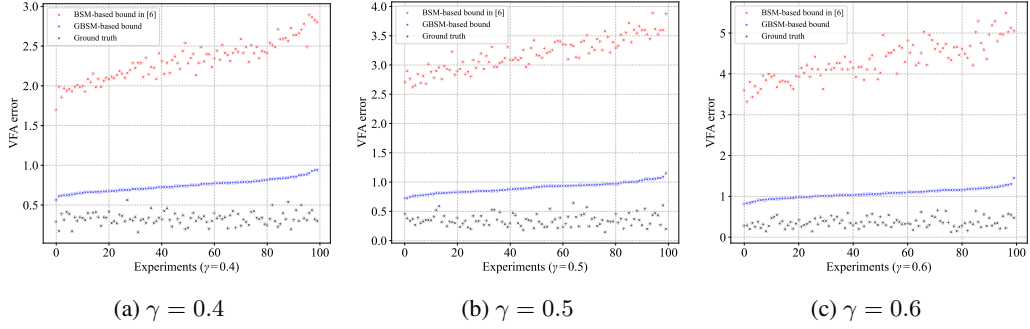


Figure 12: Experiments on random Garnet MDPs (VFA, $\gamma = 0.4$ to 0.6).

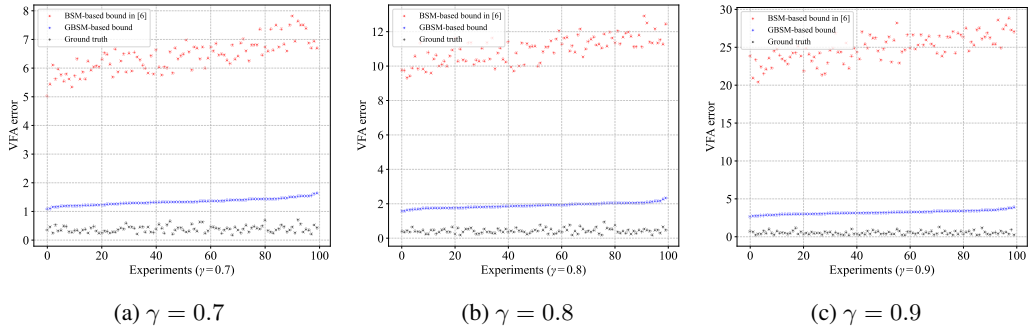


Figure 13: Experiments on random Garnet MDPs (VFA, $\gamma = 0.7$ to 0.9).