On-line Reinforcement Learning Using Incremental Kernel-Based Stochastic Factorization

SUPPLEMENTARY MATERIAL

André M. S. Barreto
School of Computer Science
McGill University
Montreal, Canada
amsb@cs.mcgill.ca

Doina Precup
School of Computer Science
McGill University
Montreal, Canada
dprecup@cs.mcgill.ca

Joelle Pineau
School of Computer Science
McGill University
Montreal, Canada
jpineau@cs.mcgill.ca

Abstract

This is the supplementary material for the paper entitled "On-line Reinforcement Learning Using Incremental Kernel-Based Stochastic Factorization" [2]. It contains the details of our theoretical developments that could not be included in the paper due to space constraints. This material should be read in conjunction with the main paper.

1 Preliminaries

- Similarly to Ormoneit and Sen [3], we define a "mother kernel" $\phi(x) : \mathbb{R}^+ \mapsto \mathbb{R}^+$ satisfying
  (i) $\phi(x)$ is continuous in $\mathbb{R}^+$,
  (ii) $\int_0^\infty \phi(x)dx \leq L_\phi < \infty$,
  (iii) $\phi(x) \geq \phi(y)$ if $x < y$,
  (iv) $\exists\, A_\phi, \lambda_\phi > 0, \exists\, B_\phi \geq 0$ such that $A_\phi \exp(-x) \leq \phi(x) \leq \lambda_\phi A_\phi \exp(-x)$ if $x \geq B_\phi$.

  Remarks:
  - Assumption (i) is implied by Ormoneit and Sen's [3] assumption that ϕ is Lipschitz continuous. Ormoneit and Sen also assume that $\int_0^1 \phi(z)dz = 1$ (see Appendix A.1 in [3]).
  - Assumption (iv) implies that the kernel function ϕ will eventually decay exponentially and also that $\phi(z) > 0$ for all $z \in \mathbb{R}^+$.
- Let $\mathbb{S} \subset [0,1]^d$ and let $\| \cdot \|$ be a norm in $\mathbb{R}^d$. Then, we define

$$\mathrm{k}_\tau(s, s') = \phi\left(\frac{\| s - s' \|}{\tau}\right),$$

  where $\tau > 0$ is the "width" of the kernel k_τ.
- Let M be a Markov decision process (MDP) with state space $\mathbb{S}$ and let $S^a = \{(s_k^a, r_k^a, \hat{s}_k^a) | k = 1, 2, ..., n_a\}$ be a set of sample transitions associated with action $a \in A$, where $s_k^a, \hat{s}_k^a \in \mathbb{S}$ and $r_k^a \in \mathbb{R}$. We define the normalized kernel function associated with action a as

$$\kappa_\tau^a(s, s_i^a) = \frac{\mathrm{k}_\tau(s, s_i^a)}{\sum_{j=1}^{n_a} \mathrm{k}_\tau(s, s_j^a)}.$$

- Let $\bar{S} \equiv \{\bar{s}_1, \bar{s}_2, ..., \bar{s}_m\}$ be a set of representative states in $\mathbb{S}$. Define:
  - $\hat{s}_*^a \equiv \hat{s}_k^a$ with $k = \mathrm{argmax}_i \min_j \| \hat{s}_i^a - \bar{s}_j \|$,

– $\bar{s}^a_* \equiv \bar{s}_h$ where $h = \text{argmin}_j \parallel \hat{s}^a_* - \bar{s}_j \parallel$,
– $\hat{s}_* \equiv \hat{s}^b_*$ where $b = \text{argmax}_a \parallel \hat{s}^a_* - \bar{s}^a_* \parallel$,
– $\bar{s}_* \equiv \bar{s}^b_*$ where $b = \text{argmax}_a \parallel \hat{s}^a_* - \bar{s}^a_* \parallel$,
– $\mathfrak{d}^* \equiv \parallel \hat{s}_* - \bar{s}_* \parallel$.

We assume that

(v) $\hat{s}^a_*$ and $\bar{s}^a_*$ are unique for all $a \in A$.

2 Data-independent definitions

Definition 1. *For any* $\alpha \in (0,1]$*, the* α-radius *of* k_τ *with respect to* s *and* s' *is defined as*

$$\rho(\mathrm{k}_\tau, s, s', \alpha) = \max\left\{x \in \mathbb{R}^+ | \phi\left(\frac{x}{\tau}\right) = \alpha \mathrm{k}_\tau(s,s')\right\}.$$

Remarks:

- The existence of $\rho(\mathrm{k}_\tau, s, s', \alpha)$ is guaranteed by properties (i), (ii) and (iii).
- $\rho(\mathrm{k}_\tau, s, s', \alpha) \geq \parallel s - s' \parallel$.

Property 1. *If* $\parallel s - s' \parallel < \parallel s - s'' \parallel$*, then* $\rho(\mathrm{k}_\tau, s, s', \alpha) \leq \rho(\mathrm{k}_\tau, s, s'', \alpha)$.

Proof. Let $r = \rho(\mathrm{k}_\tau, s, s', \alpha)$. Then,

$$\phi\left(\frac{r}{\tau}\right) = \alpha \mathrm{k}_\tau(s,s') = \alpha\phi\left(\frac{\parallel s - s' \parallel}{\tau}\right) \geq \alpha\phi\left(\frac{\parallel s - s'' \parallel}{\tau}\right) = \alpha \mathrm{k}_\tau(s,s'').$$

If $\phi(r/\tau) = \alpha \mathrm{k}_\tau(s,s'')$, then $\rho(\mathrm{k}_\tau, s, s'', \alpha) = r$. If $\phi(r/\tau) > \alpha \mathrm{k}_\tau(s,s'')$, then from (iii) it must be the case that $r < \rho(\mathrm{k}_\tau, s, s'', \alpha)$. □

Property 2. *If* $\alpha < \alpha'$*, then* $\rho(\mathrm{k}_\tau, s, s', \alpha) > \rho(\mathrm{k}_\tau, s, s', \alpha')$.

Proof. Let $r = \rho(\mathrm{k}_\tau, s, s', \alpha')$. Then,

$$\phi\left(\frac{r}{\tau}\right) = \alpha' \mathrm{k}_\tau(s,s') = \alpha'\phi\left(\frac{\parallel s - s' \parallel}{\tau}\right) > \alpha\phi\left(\frac{\parallel s - s' \parallel}{\tau}\right) = \alpha \mathrm{k}_\tau(s,s').$$

From (iii) it must be the case that $r < \rho(\mathrm{k}_\tau, s, s', \alpha)$. □

Property 3. *For any* $\alpha \in (0,1)$ *and any* $\varepsilon > 0$*, there is a* $\delta > 0$ *such that* $\rho(\mathrm{k}_\tau, s, s', \alpha) - \parallel s - s' \parallel < \varepsilon$ *if* $\tau < \delta$.

Proof. Let $z = \parallel s - s' \parallel$. We will show that, for any $\varepsilon > 0$, there is a $\delta > 0$ such that $\phi((z+\varepsilon)/\tau) < \alpha\phi(z/\tau)$ if $\tau < \delta$. We know that

$$\frac{\exp(-(z+\varepsilon)/\tau)}{\exp(-z/\tau)} < \alpha/\lambda_\phi \iff \ln\left(\frac{\exp(-(z+\varepsilon)/\tau)}{\exp(-z/\tau)}\right) < \ln(\alpha/\lambda_\phi) \iff$$
$$\iff -\frac{\varepsilon}{\tau} < \ln(\alpha/\lambda_\phi) \iff \tau < -\frac{\varepsilon}{\ln(\alpha/\lambda_\phi)}$$

(note that it must be the case that $\alpha/\lambda_\phi \neq 1$). Thus, by taking $\delta < \min(-\varepsilon/\ln(\alpha/\lambda_\phi), z/B_\phi)$ and resorting to Assumption (iv), we can write:

$$\begin{aligned}
\alpha/\lambda_\phi &> \frac{\exp(-(z+\varepsilon)/\delta)}{\exp(-z/\delta)} = \frac{A_\phi \exp(-(z+\varepsilon)/\delta)}{A_\phi \exp(-z/\delta)} \\
&\geq \frac{A_\phi \exp(-(z+\varepsilon)/\delta)}{\phi(z/\delta)} = \frac{\lambda_\phi A_\phi \exp(-(z+\varepsilon)/\delta)}{\lambda_\phi \phi(z/\delta)} \\
&\geq \frac{\phi((z+\varepsilon)/\delta)}{\lambda_\phi \phi(z/\delta)},
\end{aligned}$$

and therefore $\dfrac{\phi((z+\varepsilon)/\tau)}{\phi(z/\tau)} < \alpha$ if $\tau \leq \delta$. □

Remarks:

- $\rho(\mathrm{k}_\tau, s, s', \alpha) - \| s - s' \| < \varepsilon$ if $\tau < \min(-\varepsilon/\ln(\alpha/\lambda_\phi), \| s - s' \| / B_\phi)$, where λ_ϕ and B_ϕ depend on the particular choice of function ϕ (see Assumption (iv)).

- Given s, s', and s'', with $\| s - s' \| < \| s - s'' \|$, Property 3 states that for any $\alpha \in (0,1)$, there is a $\delta > 0$ such that $\mathrm{k}_\tau(s, s'') < \alpha \mathrm{k}_\tau(s, s')$ if $\tau < \delta$ (to see why this is so, it suffices to make $\varepsilon = \| s - s'' \| - \| s - s' \|$).

3 Data-dependent definitions

Definition 2. *Given $\beta > 0$, the*β-dissimilarity *between s and s' with respect to κ_τ^a is defined as*

$$\psi(\kappa_\tau^a, s, s', \beta) = \begin{cases} \sum_{k=1}^{n_a} |\kappa_\tau^a(s, s_k^a) - \kappa_\tau^a(s', s_k^a)|, \text{ if } \| s - s' \| \leq \beta, \\ 0, \text{ otherwise.} \end{cases}$$

Remark: $\psi(\kappa_\tau^a, s, s', \beta) \in [0, 2]$.

Property 4. *For any $\beta > 0$ and any $\varepsilon > 0$, there is a $\delta > 0$ such that $\psi(\kappa_\tau^a, s, s', \beta) < \varepsilon$ if $\| s - s' \| < \delta$.*

Proof. If $\beta < \| s - s' \|$, then $\psi(\kappa_\tau^a, s, s', \beta) = 0$ and the result follows (see Definition 2). Otherwise:

$$\begin{aligned} \psi(\kappa_\tau^a, s, s', \beta) \equiv \psi_{\tau,s,\beta}^a(s') &= \sum_{k=1}^{n_a} \left| \frac{\mathrm{k}_\tau(s, s_k^a)}{\sum_{l=1}^{n_a} \mathrm{k}_\tau(s, s_l^a)} - \frac{\mathrm{k}_\tau(s', s_k^a)}{\sum_{l=1}^{n_a} \mathrm{k}_\tau(s', s_l^a)} \right| \\ &= \sum_{k=1}^{n_a} \left| \frac{\phi\left(\| s - s_k^a \| / \tau\right)}{\sum_{l=1}^{n_a} \phi\left(\| s - s_l^a \| / \tau\right)} - \frac{\phi\left(\| s' - s_k^a \| / \tau\right)}{\sum_{l=1}^{n_a} \phi\left(\| s' - s_l^a \| / \tau\right)} \right|. \end{aligned}$$

From the definition of ϕ, it is obvious that $\psi_{\tau,s,\beta}^a(s')$ is continuous in s'. The property follows from the fact that $\lim_{s' \to s} \psi_{\tau,s,\beta}^a(s') = 0$. □

Remarks:

- $\psi(\kappa_\tau^a, s, s', \beta)$ does *not* necessarily increase with $\| s - s' \|$.

- Given $\varepsilon > 0$, δ is data-dependent.

4 Main Results

Lemma 1. *For any $\alpha \in (0,1]$ and any $t \geq m - 1$, let $\delta_a = \rho(\mathrm{k}_{\bar{\tau}}, \hat{s}_*^a, \bar{s}_*^a, \alpha/t)$, let $\psi_\delta^a = \max_{i,j} \psi(\kappa_\tau^a, \hat{s}_i^a, \bar{s}_j, \delta_a)$ and let $\psi_{\max}^a = \max_{i,j} \psi(\kappa_\tau^a, \hat{s}_i^a, \bar{s}_j, \infty)$. Then,*

$$\|\mathbf{P}^a - \mathbf{DK}^a\|_\infty \leq \frac{1}{1+\alpha} \psi_\delta^a + \frac{\alpha}{1+\alpha} \psi_{\max}^a. \quad (1)$$

Proof. Let $\tilde{\mathbf{P}}^a = \mathbf{D}\mathbf{K}^a$. Recalling that $\kappa_\tau^a(s, s_j^a) = \frac{\mathrm{k}_\tau(s, s_j^a)}{\sum_{k=1}^{n_a} \mathrm{k}_\tau(s, s_k^a)}$, we can write:

$$
\begin{aligned}
\|\mathbf{p}_i^a - \tilde{\mathbf{p}}_i^a\|_1 &= \sum_{j=1}^{n_a} |p_{ij}^a - \sum_{k=1}^{m} \dot{d}_{ik}^a \dot{k}_{kj}^a| \\
&= \sum_{j=1}^{n_a} \left| \frac{\mathrm{k}_\tau(\hat{s}_i^a, s_j^a)}{\sum_{l=1}^{n} \mathrm{k}_\tau(\hat{s}_i^a, s_l^a)} - \sum_{k=1}^{m} \left(\frac{\mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_k)}{\sum_{l=1}^{m} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l)} \frac{\mathrm{k}_\tau(\bar{s}_k, s_j^a)}{\sum_{l=1}^{n} \mathrm{k}_\tau(\bar{s}_k, s_l^a)} \right) \right| \\
&= \sum_{j=1}^{n_a} \left| \frac{\mathrm{k}_\tau(\hat{s}_i^a, s_j^a)}{\sum_{l=1}^{n} \mathrm{k}_\tau(\hat{s}_i^a, s_l^a)} - \frac{1}{\sum_{l=1}^{m} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l)} \sum_{k=1}^{m} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_k) \frac{\mathrm{k}_\tau(\bar{s}_k, s_j^a)}{\sum_{l=1}^{n} \mathrm{k}_\tau(\bar{s}_k, s_l^a)} \right| \\
&= \sum_{j=1}^{n_a} \left| \frac{1}{\sum_{l=1}^{m} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l)} \sum_{k=1}^{m} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_k) \frac{\mathrm{k}_\tau(\hat{s}_i^a, s_j^a)}{\sum_{l=1}^{n} \mathrm{k}_\tau(\hat{s}_i^a, s_l^a)} - \frac{1}{\sum_{l=1}^{m} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l)} \sum_{k=1}^{m} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_k) \frac{\mathrm{k}_\tau(\bar{s}_k, s_j^a)}{\sum_{l=1}^{n} \mathrm{k}_\tau(\bar{s}_k, s_l^a)} \right| \\
&\le \sum_{j=1}^{n_a} \frac{1}{\sum_{l=1}^{m} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l)} \sum_{k=1}^{m} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_k) \left| \frac{\mathrm{k}_\tau(\hat{s}_i^a, s_j^a)}{\sum_{l=1}^{n} \mathrm{k}_\tau(\hat{s}_i^a, s_l^a)} - \frac{\mathrm{k}_\tau(\bar{s}_k, s_j^a)}{\sum_{l=1}^{n} \mathrm{k}_\tau(\bar{s}_k, s_l^a)} \right| \\
&= \frac{1}{\sum_{l=1}^{m} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l)} \sum_{k=1}^{m} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_k) \sum_{j=1}^{n_a} \left| \kappa_\tau^a(\hat{s}_i^a, s_j^a) - \kappa_\tau^a(\bar{s}_k, s_j^a) \right|.
\end{aligned}
$$

Let $H = \{k| \parallel \hat{s}_i^a - \bar{s}_k \parallel \le \delta_a\}$ and let $\bar{H} = \{1, 2, ..., m\} - H$. Then,

$$
\begin{aligned}
\|\mathbf{p}_i^a - \tilde{\mathbf{p}}_i^a\|_1 &\le \sum_{k \in H} \frac{\mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_k)}{\sum_{l=1}^{m} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l)} \sum_{j=1}^{n_a} \left| \kappa_\tau^a(\hat{s}_i^a, s_j^a) - \kappa_\tau^a(\bar{s}_k, s_j^a) \right| + \sum_{k \in \bar{H}} \frac{\mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_k)}{\sum_{l=1}^{m} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l)} \sum_{j=1}^{n_a} \left| \kappa_\tau^a(\hat{s}_i^a, s_j^a) - \kappa_\tau^a(\bar{s}_k, s_j^a) \right| \\
&\le \frac{\sum_{k \in H} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_k)}{\sum_{l=1}^{m} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l)} \max_{h \in H} \psi(\kappa_\tau^a, \hat{s}_i^a, \bar{s}_h, \infty) + \frac{\sum_{k \in \bar{H}} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_k)}{\sum_{l=1}^{m} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l)} \max_{h \in \bar{H}} \psi(\kappa_\tau^a, \hat{s}_i^a, \bar{s}_h, \infty) \\
&\le \frac{\sum_{k \in H} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_k)}{\sum_{l=1}^{m} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l)} \psi_\delta^a + \frac{\sum_{k \in \bar{H}} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_k)}{\sum_{l=1}^{m} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l)} \max_{h} \psi(\kappa_\tau^a, \hat{s}_i^a, \bar{s}_h, \infty) \\
&\le \frac{\sum_{k \in H} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_k)}{\sum_{l=1}^{m} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l)} \psi_\delta^a + \frac{\sum_{k \in \bar{H}} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_k)}{\sum_{l=1}^{m} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l)} \psi_{\max}^a \\
&= \frac{\sum_{k \in H} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_k)}{\sum_{l \in H} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l) + \sum_{l \in \bar{H}} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l)} \psi_\delta^a + \frac{\sum_{k \in \bar{H}} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_k)}{\sum_{l \in H} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l) + \sum_{l \in \bar{H}} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l)} \psi_{\max}^a.
\end{aligned} \tag{2}
$$

From the definition of δ_a we can write

$$
\mathrm{k}_{\bar{\tau}}(\hat{s}_*^a, \bar{s}_k) \le \frac{\alpha}{t} \mathrm{k}_{\bar{\tau}}(\hat{s}_*^a, \bar{s}_*^a) \text{ if } \parallel \hat{s}_*^a - \bar{s}_k \parallel \ge \delta_a. \tag{3}
$$

Now, let $w = \mathrm{argmin}_k \parallel \hat{s}_i^a - \bar{s}_k \parallel$. We know that $\parallel \hat{s}_i^a - \bar{s}_w \parallel \le \parallel \hat{s}_*^a - \bar{s}_*^a \parallel$, and thus, from Assumption (iii), it follows that $\mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_w) \ge \mathrm{k}_{\bar{\tau}}(\hat{s}_*^a, \bar{s}_*^a)$. This fact together with (3) imply that

$$
\mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_k) \le \frac{\alpha}{t} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_w) \text{ if } \bar{s}_w \in \bar{H},
$$

which allows us to write

$$
\sum_{l \in \bar{H}} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l) \le \frac{\alpha |\bar{H}|}{t} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_w). \tag{4}
$$

Plugging (4) back into (2), we can write:

$$
\begin{aligned}
\|\mathbf{p}_i^a - \tilde{\mathbf{p}}_i^a\|_1 &\le \frac{\sum_{k \in H} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_k)}{\sum_{l \in H} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l) + \frac{\alpha|\bar{H}|}{t} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_w)} \psi_\delta^a + \frac{\frac{\alpha|\bar{H}|}{t} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_w)}{\sum_{l \in H} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l) + \frac{\alpha|\bar{H}|}{t} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_w)} \psi_{\max}^a & (5) \\
&\le \frac{\sum_{k \in H} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_k)}{\sum_{l \in H} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l) + \frac{\alpha t}{t} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_w)} \psi_\delta^a + \frac{\frac{\alpha t}{t} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_w)}{\sum_{l \in H} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l) + \frac{\alpha t}{t} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_w)} \psi_{\max}^a & (6) \\
&= \frac{\sum_{k \in H} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_k)}{\sum_{l \in H} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l) + \alpha \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_w)} \psi_\delta^a + \frac{\alpha \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_w)}{\sum_{l \in H} \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_l) + \alpha \mathrm{k}_{\bar{\tau}}(\hat{s}_i^a, \bar{s}_w)} \psi_{\max}^a, & (7)
\end{aligned}
$$

where in (5) and (6) we used the fact that the coefficients multiplying ψ_δ^a and $\psi_{\max}^a$ define a convex combination, and we are increasing the weight of the latter. Noticing that

$$\frac{\alpha \mathrm{k}_{\bar\tau}(\hat s_i^a, \bar s_w)}{\sum_{l\in H} \mathrm{k}_{\bar\tau}(\hat s_i^a, \bar s_l) + \alpha \mathrm{k}_{\bar\tau}(\hat s_i^a, \bar s_w)} \le \frac{\alpha \mathrm{k}_{\bar\tau}(\hat s_i^a, \bar s_w)}{\mathrm{k}_{\bar\tau}(\hat s_i^a, \bar s_w) + \alpha \mathrm{k}_{\bar\tau}(\hat s_i^a, \bar s_w)} = \frac{\alpha}{1+\alpha},$$

and applying the same reasoning to the coefficients of ψ_δ^a and $\psi_{\max}^a$ in (7), we can finally write

$$\|\mathbf{p}_i^a - \tilde{\mathbf{p}}_i^a\|_1 \le \frac{1}{1+\alpha}\psi_\delta^a + \frac{\alpha}{1+\alpha}\psi_{\max}^a.$$

□

Remarks:

- $\psi_\delta^a \to \psi_{\max}^a$ as $\alpha \to 0$.
- There is an $\alpha^* \in (0,1]$ that minimizes the right-hand side of (1).

Let $\tilde{\mathbf{v}} = \Gamma \mathbf{D}\bar{\mathbf{Q}}^*$, where Γ is the 'max' operator applied row wise, that is, $\tilde v_i = \max_a (\mathbf{D}\bar{\mathbf{Q}}^*)_{ia}$. Recalling that $\mathfrak{d}^*$ is the maximum distance from a sampled state $\hat s_i^a$ to the closest representative state and that $\bar\tau$ is the width of kernel $\mathrm{k}_{\bar\tau}$, we present the following result:

Proposition 1. *For any $\varepsilon > 0$, there are $\delta_1, \delta_2 > 0$ such that $\|\hat{\mathbf{v}}^* - \tilde{\mathbf{v}}\|_\infty < \varepsilon$ if $\mathfrak{d}^* < \delta_1$ and $\bar\tau < \delta_2$.*

Proof. We have previously showed that

$$\|\hat{\mathbf{v}}^* - \tilde{\mathbf{v}}\|_\infty \le \frac{1}{1-\gamma}\max_a \|\hat{\mathbf{r}}^a - \mathbf{D}\bar{\mathbf{r}}^a\|_\infty + \frac{1}{(1-\gamma)^2}\left(\bar C \max_i (1 - \max_j d_{ij}) + \frac{\hat C \gamma}{2}\max_a \left\|\hat{\mathbf{P}}^a - \mathbf{D}\mathbf{K}^a\right\|_\infty\right), \tag{8}$$

where $\|\cdot\|_\infty$ is the infinity norm, $\hat{\mathbf{v}}^* \in \mathbb{R}^n$ is the optimal value function of KBRL's MDP, $\hat C = \max_{a,i} \hat r_i^a - \min_{a,i} \hat r_i^a$, $\bar C = \max_{a,i} \bar r_i^a - \min_{a,i} \bar r_i^a$, and $\mathbf{K}^a$ is matrix $\mathbf{K}$ with all elements equal to zero except for those corresponding to matrix $\dot{\mathbf{K}}^a$ (see [1, 2] for details). Let $\check{\mathbf{r}} \equiv [(\mathbf{r}^1)^\intercal, (\mathbf{r}^2)^\intercal, ..., (\mathbf{r}^{|A|})^\intercal]^\intercal \in \mathbb{R}^n$, where $\mathbf{r}^a \in \mathbb{R}^{n_a}$ is the vector composed of sample rewards r_i^a. Then,

$$\|\hat{\mathbf{r}}^a - \mathbf{D}\bar{\mathbf{r}}^a\|_\infty = \left\|\hat{\mathbf{P}}^a\check{\mathbf{r}} - \mathbf{D}\dot{\mathbf{K}}^a\mathbf{r}^a\right\|_\infty = \left\|\hat{\mathbf{P}}^a\check{\mathbf{r}} - \mathbf{D}\mathbf{K}^a\check{\mathbf{r}}\right\|_\infty = \left\|(\hat{\mathbf{P}}^a - \mathbf{D}\mathbf{K}^a)\check{\mathbf{r}}\right\|_\infty \le \left\|\hat{\mathbf{P}}^a - \mathbf{D}\mathbf{K}^a\right\|_\infty \|\check{\mathbf{r}}\|_\infty, \tag{9}$$

where the equality $\hat{\mathbf{r}}^a = \hat{\mathbf{P}}^a\check{\mathbf{r}}$ is a consequence of the fact that KBRL's reward function $R^a(s,s')$ is independent of the start state s (see (1) in the main paper [2]). Thus, plugging (9) back into (8), it is clear that there is a $\eta > 0$ such that $\|\hat{\mathbf{v}}^* - \tilde{\mathbf{v}}\|_\infty < \varepsilon$ if $\max_a \left\|\hat{\mathbf{P}}^a - \mathbf{D}\mathbf{K}^a\right\|_\infty < \eta$ and $\max_i (1 - \max_j d_{ij}) < \eta$. We start by showing that if $\mathfrak{d}^*$ and $\bar\tau$ are small enough, then $\max_a \left\|\hat{\mathbf{P}}^a - \mathbf{D}\mathbf{K}^a\right\|_\infty < \eta$. From Lemma 1 we know that, for any set of $m \le n$ representative states, and for any $\alpha \in (0,1]$, the following must hold:

$$\max_a \|\mathbf{P}^a - \mathbf{D}\mathbf{K}^a\|_\infty \le \frac{1}{1+\alpha}\psi_\rho + \frac{\alpha}{1+\alpha}\psi_{\mathrm{MAX}},$$

where $\psi_{\mathrm{MAX}} = \max_{a,i,s} \psi(\mathrm{k}_\tau, \hat s_i^a, s, \infty)$ and $\psi_\rho = \max_a \psi_\rho^a = \max_{a,i,j} \psi(\kappa_\tau^a, \hat s_i^a, \bar s_j, \rho^a)$, with $\rho^a = \rho(\mathrm{k}_{\bar\tau}, \hat s_*^a, \bar s_*^a, \alpha/(n-1))$. Note that ψ_{MAX} is independent of the representative states. Define α such that $\alpha/(1+\alpha)\psi_{\mathrm{MAX}} < \eta$. We have to show that, if we define the representative states in such a way that $\mathfrak{d}^*$ is small enough, and set $\bar\tau$ accordingly, then we can make $\psi_\rho < (1-\alpha)\eta - \alpha\psi_{\mathrm{MAX}} \equiv \eta'$. From Property 4 we know that there is a $\delta_1 > 0$ such that $\psi_\rho < \eta'$ if $\rho^a < \delta_1$ for all $a \in A$. From Property 1 we know that $\rho^a \le \rho(\mathrm{k}_{\bar\tau}, \hat s_*, \bar s_*, \alpha/(n-1))$ for all $a \in A$. From Property 3 we know that, for any $\varepsilon' > 0$, there is a $\delta' > 0$ such that $\rho(\mathrm{k}_{\bar\tau}, \hat s_*, \bar s_*, \alpha/(n-1)) < \mathfrak{d}^* + \varepsilon'$ if $\bar\tau < \delta'$. Therefore, if $\mathfrak{d}^* < \delta_1$, we can take any $\varepsilon' < \delta_1 - \mathfrak{d}^*$ to have an upper bound δ' for $\bar\tau$. It remains to show that there is a $\delta > 0$ such that $\min_i \max_j d_{ij} > 1 - \eta$ if $\bar\tau < \delta$. Recalling that $d_{ij}^a = \mathrm{k}_{\bar\tau}(\hat s_i^a, \bar s_j)/\sum_{k=1}^m \mathrm{k}_{\bar\tau}(\hat s_i^a, \bar s_k)$, let $w = \mathrm{argmax}_j \mathrm{k}_{\bar\tau}(\hat s_i^a, \bar s_j)$, and let $y_i^a = \mathrm{k}_{\bar\tau}(\hat s_i^a, \bar s_w)$ and $\check y_i^a = \max_{j\ne w} \mathrm{k}_{\bar\tau}(\hat s_i^a, \bar s_j)$. Then, for any i,

$$\max_j d_{ij}^a = \frac{y_i^a}{\left(y_i^a + \sum_{j\ne w} \mathrm{k}_{\bar\tau}(\hat s_i^a, \bar s_j)\right)} \ge \frac{y_i^a}{(y_i^a + (m-1)\check y_i^a)}.$$

From Assumption (v) and Property 3 we know that there is a $\delta_i^a > 0$ such that $y_i^a > (m-1)(1-\eta)\check y_i^a/\eta$ if $\bar\tau < \delta_i^a$. Thus, by making $\delta = \min_{a,i} \delta_i^a$, we can guarantee that $\min_i \max_j d_{ij} > 1 - \eta$. Finally, if we take $\delta_2 = \min(\delta, \delta')$, the result follows. □

Remark: If we define a "net" over $\mathbb{S}$ using the representative states, then we know that $\mathfrak{d}^*$ is smaller than the resolution of the net.

References

[1] A. M. S. Barreto, D. Precup, and J. Pineau. Reinforcement learning using kernel-based stochastic factorization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 720–728, 2011.

[2] A. M. S. Barreto, D. Precup, and J. Pineau. On-line reinforcement learning using incremental kernel-based stochastic factorization. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[3] D. Ormoneit and S. Sen. Kernel-based reinforcement learning. *Machine Learning*, 49 (2–3): 161–178, 2002.