Nonparametric Bellman Mappings for Reinforcement Learning: Application to Robust Adaptive Filtering

Yuki Akiyama, Minh Vu, and Konstantinos Slavakis*

***Abstract*—This paper designs novel nonparametric Bellman mappings in reproducing kernel Hilbert spaces (RKHSs) for reinforcement learning (RL). The proposed mappings benefit from the rich approximating properties of RKHSs, adopt no assumptions on the statistics of the data owing to their nonparametric nature, require no knowledge on transition probabilities of Markov decision processes, and may operate without any training data. Moreover, they allow for sampling on-the-fly via the design of trajectory samples, re-use past test data via experience replay, effect dimensionality reduction by random Fourier features, and enable computationally lightweight operations to fit into efficient online or time-adaptive learning. The paper offers also a variational framework to design the free parameters of the proposed Bellman mappings, and shows that appropriate choices of those parameters yield several popular Bellman-mapping designs. As an application, the proposed mappings are employed to offer a novel solution to the problem of countering outliers in adaptive filtering. More specifically, with no prior information on the statistics of the outliers and no training data, a policy-iteration algorithm is introduced to select online, per time instance, the "optimal" coefficient p in the least-mean-p-power-error method. Numerical tests on synthetic data showcase, in most of the cases, the superior performance of the proposed solution over several RL and non-RL schemes.**

***Index Terms*—Bellman mappings, reinforcement learning, nonparametric, adaptive filtering, outliers.**

I. Introduction

A. Motivation: Adaptive filters against outliers

The least-squares (LS) error/loss plays a pivotal role in signal processing, e.g., adaptive filtering (AdaFilt) [1], and machine learning [2, 3]. Notwithstanding, the LS loss is notoriously sensitive to the presence of outliers [4], where outliers are defined as contaminating data that do not adhere to a nominal data-generation model, and are often viewed as random variables (RVs) with non-Gaussian heavy tailed distributions, *e.g.*, α-stable ones [5, 6]. To counter outliers in AdaFilt, non-LS losses, such as the p-norm ($2 > p \in \mathbb{R}_{++}$) [7–14] and correntropy [15, 16] have been studied (henceforth, $\mathbb{R}_{++}$ will denote the set of all positive real numbers).

Consider the classical linear data-generation model in AdaFilt: $y_n = \boldsymbol{\theta}_*^\intercal \mathbf{x}_n + o_n$, where $n \in \mathbb{N}$ denotes discrete time ($\mathbb{N}$ is the set of all non-negative integers), $\boldsymbol{\theta}_* \in \mathbb{R}^L$ is the $L \times 1$ vector/system with real-valued entries that needs to be estimated (estimandum), o_n is the real-valued RV which models outliers/noise, $(\mathbf{x}_n, y_n)$ stands for the input-output pair of available data, where $\mathbf{x}_n \in \mathbb{R}^L$ and $y_n \in \mathbb{R}$, and $\intercal$ denotes vector/matrix transposition. The online-learning setting is considered, that is, data $(\mathbf{x}_n, y_n)_{n\in\mathbb{N}}$ appear to the user/agent in a streaming fashion, a pair $(\mathbf{x}_n, y_n)$ per time index n, while no training data are available. All operations in the following discussion are performed online, so that the time index n coincides with the iteration index of the proposed reinforcement-learning (RL) algorithm.

Motivated by the importance and longevity of p-norm algorithms in robust statistics [4, 7], this study focuses on the least-mean-p-power-error (LMP) method [8] because of its simplicity: LMP is an application of the classical stochastic-gradient-descent (SGD) method to the p-power/norm error/loss $|y_n - \mathbf{x}_n^\intercal \boldsymbol{\theta}|^p$, $p \in [1, 2]$. In other words, for an arbitrarily fixed $\boldsymbol{\theta}_0 \in \mathbb{R}^L$, LMP generates sequence $(\boldsymbol{\theta}_n)_{n\in\mathbb{N}}$ to estimate $\boldsymbol{\theta}_*$:

$$\boldsymbol{\theta}_{n+1} := \boldsymbol{\theta}_n + \rho p \, \mathrm{sgn}(e_n) \, |e_n|^{p-1} \, \mathbf{x}_n \,, \tag{1}$$

where $e_n := y_n - \mathbf{x}_n^\intercal \boldsymbol{\theta}_n = \mathbf{x}_n^\intercal (\boldsymbol{\theta}_* - \boldsymbol{\theta}_n) + o_n$ is the classical a-priori error [1, (10.11)], ρ is the learning rate (step size), and $\mathrm{sgn}(\cdot)\colon \mathbb{R} \to \{\pm 1\}$ provides the sign of a real number. The p-power loss is a convex function of $\boldsymbol{\theta}$ for $p \in [1, 2]$. If $p = 1$ and 2, then (1) boils down to the classical sign-LMS and LMS, respectively [1]. The p-power loss remains convex even if $p > 2$, but such values of p may amplify large-variance outliers o_n via $|e_n|^{p-1}$ and inflict instabilities on (1).

Intuition suggests that the choice of p in (1) should be based on the probability density function (PDF) of the RV o_n. Indeed, if o_n obeys a Gaussian PDF, then $p = 2$ yields the 2-power loss which agrees with the maximum-likelihood criterion. Nevertheless, having prior knowledge on the statistics of the outliers is usually infeasible in practice, as in cases where no training data are available, and in dynamic environments where the statistics of the outliers may be time varying.

Combinations of LMP filters, with different p-power losses [9] as well as forgetting factors [12], have been proposed to surmount the problem of pinpointing the "best" p, but still, the problem remains and translates to that of pinpointing the "best" combination, which again depends on the underlying

*Y. Akiyama, M. Vu, and K. Slavakis are with Institute of Science Tokyo, Department of Information and Communications Engineering, 4259-G2-4 Nagatsuta-Cho, Midori-Ku, Yokohama, Kanagawa, 226-8502 Japan. Emails: {akiyama.y.ce1f, vu.d.a5c3}@m.isct.ac.jp, slavakis@ict.eng.isct.ac.jp.

Cite as: Y. Akiyama, M. Vu and K. Slavakis, "Nonparametric Bellman mappings for reinforcement learning: Application to robust adaptive filtering," IEEE Transactions on Signal Processing, vol. 72, pp. 5644–5658, 2024, DOI: 10.1109/TSP.2024.3505266.

This file contains corrections to missing qualifiers of Q from (7), (10) and (14) of the published article. The proof in Appendix A of the supplementary file has been also corrected to reflect the newly added qualifiers.

File created on January 15, 2025.

outlier PDF. A *data-driven* solution to the problem of *dynamically* selecting p, per time instance n, from streaming data with *no prior knowledge* on the statistics of o_n and *no training data* seems to be missing from the AdaFilt literature.

B. Contributions

Building on its short preliminary version [17], this manuscript offers a solution to the aforementioned AdaFilt problem by reinforcement learning (RL) [18, 19]. In RL, an agent takes a decision/action based on feedback provided by the surrounding environment on the agent's past actions. RL is a sequential-decision-making framework with the goal of minimizing the long-term loss/price (*a.k.a.* Q-function) to be paid by the agent for its own decisions. Central to RL are *Bellman mappings (B-Maps)* which operate on the Q-functions, have deep roots in dynamic programming [18, 20], and a range of applications which extend from autonomous navigation, robotics, resource planning, sensor networks, biomedical imaging, and can reach even to gaming [18].

Rather than adopting a popular off-the-shelf RL method, this manuscript designs a novel family of B-Maps to solve the AdaFilt problem at hand. The contributions of this work are summarized as follows.

- **(C1)** In contrast with the majority of existing B-Maps, which are defined in Banach spaces (no inner product available), the proposed B-Maps, as well as the Q-functions, are specifically defined in reproducing kernel Hilbert spaces (RKHSs) to take advantage of the rich approximating properties of RKHSs [21, 22] and the flexibility an RKHS inner product brings into the design of loss functions and constraints.
- **(C2)** The proposed B-Maps possess ample degrees of freedom; indeed, Proposition 1 offers a variational framework to identify their free parameters, and shows that by appropriately designing those parameters, several popular B-Maps fall as special cases under the umbrella of the proposed design. Section II-C provides a thorough literature review on the prior art of B-Maps.
- **(C3)** Owing to the kernel functions, the proposed B-Maps are rendered nonparametric, with no need for statistical priors and assumptions on the data, in an effort to reduce as much as possible the bias inflicted on data modeling by the user [23]. The price to be paid for this distribution-free approach is that the dimensions of the Q-function estimates scale with the number of observed data. To surmount this "curse of dimensionality," a dimensionality-reduction strategy based on random Fourier features is offered in Section III-D.
- **(C4)** The proposed B-Maps allow for sampling on-the-fly via the design of trajectory samples in Section III-C, do not require any knowledge on transition probabilities of Markov decision processes, and enable computationally lightweight operations to fit into the online or time-adaptive learning required by the AdaFilt problem at hand.
- **(C5)** For the first time in the literature, this manuscript and its short preliminary version [17] offer an RL-based solution (Algorithm 1) to the problem of countering outliers in AdaFilt.

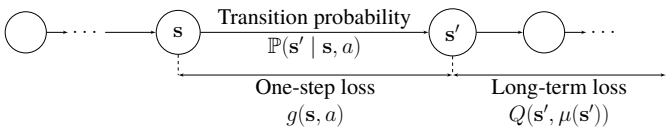


Fig. 1. RL as a sequential-decision-making framework: Identify the agent's policy $\mu(\cdot)$ (a decision- or action-making function) which minimizes the total loss (= one-step loss + long-term loss) to be paid by the agent for its sequence of decisions/actions.

The proposed B-Maps are introduced in Section II-C, their properties are established in Theorems 2 and 4, a solution to the AdaFilt problem of Section I-A is provided in Section III, and a performance analysis of Algorithm 1 is offered in Section IV. The well-known policy-iteration (PI) strategy [18] is adopted in Algorithm 1, because of its well-documented merits (*e.g.*, [24–26]), especially for continuous state spaces. To keep the discussion simple, a quadratic loss on Q-functions is defined via the proposed B-Maps, and the classical SGD rule is applied to update the Q-function estimates (Section III-B). To promote the use of past data, experience replay [27] is employed in Section III-C2, while [17] uses rollout [18]. Being approximates of the classical B-Maps, the proposed B-Maps aim at the theoretical/algorithmic core of RL. Hence, although this manuscript considers AdaFilt, the proposed B-Maps can be applied potentially to any domain where B-Maps are needed; see the beginning paragraph of this section for examples of such application domains.

With regards to (C4), it is worth noting here that the popular deep learning (DeepL), *e.g.*, [28], offers an alternative parametric way of designing rich approximating spaces for Q-functions. However, this is achieved at the price of learning from training data prior to the online mode of operation (test-stage), or even re-training during the test-stage to address the often met scenario of facing test data with different statistics than those of the training data (dynamic environments). Such modes of learning raise computational-complexity issues and discourage the application of DeepL solutions to online modes of operation where a small computational-complexity footprint is desired.

The RL-based AdaFilt solution is built on a continuous state space, because of the nature of $(\mathbf{x}_n, y_n)$. In contrast with [17], where the state space is the high-dimensional $\mathbb{R}^{2L+1}$, this study confines the state space to the low-dimensional $\mathbb{R}^4$. The action space is considered to be discrete; an action is a value of p taken from a finite grid of the interval $[1, 2]$.

Numerical tests on synthetic data in Section V support the theoretical findings and demonstrate that the advocated framework outperforms, in most of the cases, several RL and non-RL schemes. Due to limited space, all proofs and appendices are included in the supplementary file of the manuscript.

II. Novel Nonparametric Bellman Mappings for RL

A. Notation and preliminaries

A *continuous* state space $\mathfrak{S} \subset \mathbb{R}^D$ is considered, with state vector $\mathbf{s} \in \mathfrak{S}$, for some $D \in \mathbb{N}_*$ ($\mathbb{N}_*$ is the set of all positive integers). The action space is denoted by $\mathfrak{A}$, with action $a \in \mathfrak{A}$. For convenience, the state-action tuple is defined as $\mathbf{z} := (\mathbf{s}, a) \in \mathfrak{Z} := \mathfrak{S} \times \mathfrak{A}$. Moreover, let all mappings $\mathcal{M} := \{\mu(\cdot) \mid \mu(\cdot)\colon \mathfrak{S} \to \mathfrak{A} : \mathbf{s} \mapsto \mu(\mathbf{s}),$ and μ is surjective$\}$, and define a deterministic policy $\pi \in \Pi := \mathcal{M}^{\mathbb{N}} := \{(\mu_0, \mu_1, \ldots, \mu_n, \ldots) \mid \mu_n \in \mathcal{M}, n \in \mathbb{N}\}$. Given $\mu \in \mathcal{M}$, the stationary policy $\pi_\mu \in \Pi$ is defined as $\pi_\mu := (\mu, \mu, \ldots, \mu, \ldots)$. By abuse of notation, μ will hereafter denote also the stationary policy π_μ.

RL can be viewed as a sequential-decision framework; *cf.* Figure 1. In short, an agent, currently at state $\mathbf{s} \in \mathfrak{S}$, takes an action/decision $a \in \mathfrak{A}$ and transitions to a new state $\mathbf{s}' \in \mathfrak{S}$ with transition (conditional) probability $\mathbb{P}(\mathbf{s}' \mid \mathbf{s}, a)$ at the price of the *one-step* loss $g(\mathbf{s}, a)$. Quantity $Q(\mathbf{s}', \mu(\mathbf{s}'))$ denotes the *long-term* loss, or, the price to be paid if the agent continues to take actions, from the state $\mathbf{s}'$ and on, according to the stationary policy $\mu(\cdot)$. Typically, $g(\cdot)\colon \mathfrak{Z} \to \mathbb{R}$ and $Q(\cdot)\colon \mathfrak{Z} \to \mathbb{R}$ are considered points of the functional Banach space $\mathcal{B}$ of all (essentially) bounded functions, equipped with the $\mathcal{L}_\infty$-norm [18, 19]. Recall that by definition a Banach space $\mathcal{B}$ is not equipped with an inner product.

Departing from standard RL routes which revolve around Banach spaces $\mathcal{B}$, this study considers a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ [21, 22] as the ambient space where g and Q belong to. The RKHS $\mathcal{H}$ is a Hilbert space with inner product $\langle \cdot \mid \cdot \rangle_{\mathcal{H}}$, norm $\|\cdot\|_{\mathcal{H}} := \langle \cdot \mid \cdot \rangle_{\mathcal{H}}^{1/2}$, and a reproducing kernel $\kappa(\cdot, \cdot) : \mathfrak{Z} \times \mathfrak{Z} \to \mathbb{R}$ such that $\kappa(\mathbf{z}, \cdot) \in \mathcal{H}$, $\forall \mathbf{z} \in \mathfrak{Z}$, and the *reproducing property* holds true: $Q(\mathbf{z}) = \langle Q \mid \kappa(\mathbf{z}, \cdot) \rangle_{\mathcal{H}}$, $\forall Q \in \mathcal{H}$, $\forall \mathbf{z} \in \mathfrak{Z}$. Space $\mathcal{H}$ may be infinite dimensional; *e.g.*, the case where $\kappa(\cdot, \cdot)$ is a Gaussian kernel [21, 22]. For compact notations, let the feature mapping $\varphi(\mathbf{z}) := \kappa(\mathbf{z}, \cdot)$ and $Q^\intercal Q' := \langle Q \mid Q' \rangle_{\mathcal{H}}$.

Finally, notation $\mathcal{T}_N := \{(\mathbf{s}_i, a_i, \mathbf{s}_i')\}_{i=1}^N \subset \mathfrak{S} \times \mathfrak{A} \times \mathfrak{S}$, for $N \in \mathbb{N}_*$, will be used hereafter to denote a collection of trajectory samples, with $\mathbf{s}_i'$ being a potential subsequent state of $\mathbf{s}_i$ after the agent takes action a_i. This study does *not* necessitate $\mathbf{s}_{i+1} = \mathbf{s}_i'$, allowing thus for looser assumptions than Markovian ones in stochastic analyses; see, for example, Assumption 3(ii). Moreover, g_i will stand for either $g(\mathbf{s}_i, a_i, \mathbf{s}_i')$ or $g(\mathbf{s}_i, a_i)$, depending on the context of discussion. For a reproducing kernel $\kappa(\cdot, \cdot)$ and its feature mapping $\varphi(\cdot)$, let for convenience $\mathbf{\Phi}_{\mathcal{T}_N} := [\varphi(\mathbf{z}_1), \ldots, \varphi(\mathbf{z}_N)]$, where $\mathbf{z}_i := (\mathbf{s}_i, a_i)$, and define then $\mathbf{K}_{\mathcal{T}_N} := \mathbf{\Phi}_{\mathcal{T}_N}^\intercal \mathbf{\Phi}_{\mathcal{T}_N}$ as the $N \times N$ kernel matrix whose (i, i')th entry is $\kappa(\mathbf{z}_i, \mathbf{z}_{i'}) = \langle \varphi(\mathbf{z}_i) \mid \varphi(\mathbf{z}_{i'}) \rangle_{\mathcal{H}} = \varphi^\intercal(\mathbf{z}_i)\varphi(\mathbf{z}_{i'})$. Moreover, let $\mathbf{\Phi}_\mu' := [\varphi(\mathbf{s}_1', \mu(\mathbf{s}_1')), \ldots, \varphi(\mathbf{s}_N', \mu(\mathbf{s}_N'))]$, and $\mathbf{g} := [g(\mathbf{s}_1, a_1), \ldots, g(\mathbf{s}_N, a_N)]^\intercal$.

B. The classical B-Maps

The classical B-Maps are defined in a Banach space $\mathcal{B}$ and quantify the total loss (= one-step loss + expected long-term loss) to be paid by the agent when taking action a at state $\mathbf{s}$ [29]. More specifically, $T_\mu^\diamond, T^\diamond\colon \mathcal{B} \to \mathcal{B}$, where $\forall Q \in \mathcal{B}$,

$$(T_\mu^\diamond Q)(\mathbf{s}, a) := g(\mathbf{s}, a) + \alpha\, \mathbb{E}_{\mathbf{s}'|(\mathbf{s},a)}\{Q(\mathbf{s}', \mu(\mathbf{s}'))\}\,, \tag{2a}$$

$$(T^\diamond Q)(\mathbf{s}, a) := g(\mathbf{s}, a) + \alpha\, \mathbb{E}_{\mathbf{s}'|(\mathbf{s},a)}\{\inf_{a' \in \mathfrak{A}} Q(\mathbf{s}', a')\}\,, \tag{2b}$$

and where $\mathbb{E}_{\mathbf{s}'|(\mathbf{s},a)}\{\cdot\}$ stands for the conditional expectation [30] over all possible subsequent states $\mathbf{s}'$ of $\mathbf{s}$, conditioned on $(\mathbf{s}, a)$, and α is the discount factor with typical values in $(0, 1)$. Mapping (2a) refers to the case where the agent takes actions according to the stationary policy $\mu(\cdot)$, while (2b) serves as a greedy variation of (2a). Note that (2b) can be recast in the form of (2a) whenever the $\inf$ in (2b) is achievable:

$$\begin{aligned}(T^\diamond Q)(\mathbf{s}, a) &:= g(\mathbf{s}, a) + \alpha\, \mathbb{E}_{\mathbf{s}'|(\mathbf{s},a)}\{Q(\mathbf{s}', \mu_Q(\mathbf{s}'))\} \\ &= (T_{\mu_Q}^\diamond Q)(\mathbf{s}, a)\,,\end{aligned} \tag{2c}$$

where, given Q, the stationary policy $\mu_Q(\cdot)$ is defined as $\mu_Q(\mathbf{s}') := \arg\min_{a' \in \mathfrak{A}} Q(\mathbf{s}', a')$, so that $Q(\mathbf{s}', \mu_Q(\mathbf{s}')) = \min_{a' \in \mathfrak{A}} Q(\mathbf{s}', a')$.

Given a mapping $T\colon \mathcal{B} \to \mathcal{B}$, its fixed-point set is defined as $\operatorname{Fix} T := \{Q \in \mathcal{B} \mid TQ = Q\}$. It is well known that $\operatorname{Fix} T_\mu^\diamond$ and $\operatorname{Fix} T^\diamond$ play a central role in identifying the policy which minimizes the total loss [18]. Typically, the discount factor $\alpha \in (0, 1)$ to render $T_\mu^\diamond, T^\diamond$ contractions [18, 31], so that $\operatorname{Fix} T_\mu^\diamond$ and $\operatorname{Fix} T^\diamond$ become singletons [31].

C. The proposed B-Maps and prior art

In most real-world cases, (complete) knowledge of the conditional PDF of $\mathbf{s}'$ on $(\mathbf{s}, a)$ is unavailable to the agent, so that the computation of the conditional expectation $\mathbb{E}_{\mathbf{s}'|(\mathbf{s},a)}\{\cdot\}$ in (2) is rendered infeasible. A classical way to surmount this obstacle is to approximate $\mathbb{E}_{\mathbf{s}'|(\mathbf{s},a)}\{\cdot\}$ by sample averaging. This classical route, popular in RL, is also followed here to introduce the proposed B-Maps in (3).

Hereafter, losses g, Q are assumed to belong to an RKHS $\mathcal{H}$. For some $N_{\text{av}} \in \mathbb{N}_*$, consider the state-space vectors $\mathcal{S}_{\text{av}} := \{\mathbf{s}_i^{\text{av}}\}_{i=1}^{N_{\text{av}}} \subset \mathfrak{S}$, chosen by the user to enable sampling of trajectory samples on-the-fly to approximate the conditional expectation in (2). Define also, for convenience in notation and for a $\mu \in \mathcal{M}$, $\mathbf{\Phi}_\mu^{\text{av}} := [\varphi(\mathbf{s}_1^{\text{av}}, \mu(\mathbf{s}_1^{\text{av}})), \ldots, \varphi(\mathbf{s}_{N_{\text{av}}}^{\text{av}}, \mu(\mathbf{s}_{N_{\text{av}}}^{\text{av}}))]$. Consider also the user-defined $\{\psi_i\}_{i=1}^{N_{\text{av}}} \subset \mathcal{H}$, with $\mathbf{\Psi} := [\psi_1, \ldots, \psi_{N_{\text{av}}}]$. Then, the proposed B-Maps $T_\mu, T\colon \mathcal{H} \to \mathcal{H}$ are defined as follows: $\forall Q \in \mathcal{H}$, $\forall \mu \in \mathcal{M}$,

$$\begin{aligned}T_\mu(Q) &:= g + \alpha \sum\nolimits_{i=1}^{N_{\text{av}}} Q(\mathbf{s}_i^{\text{av}}, \mu(\mathbf{s}_i^{\text{av}})) \cdot \psi_i \\ &= g + \alpha \mathbf{\Psi} \mathbf{\Phi}_\mu^{\text{av}\intercal} Q\,,\end{aligned} \tag{3a}$$

$$\begin{aligned}T(Q) &:= g + \alpha \sum\nolimits_{i=1}^{N_{\text{av}}} \inf\nolimits_{a_i \in \mathfrak{A}} Q(\mathbf{s}_i^{\text{av}}, a_i) \cdot \psi_i \\ &= g + \alpha \mathbf{\Psi}\ \mathbf{inf}_{\mu \in \mathcal{M}}\, \mathbf{\Phi}_\mu^{\text{av}\intercal} Q\,,\end{aligned} \tag{3b}$$

where $\mathbf{inf}_{\mu \in \mathcal{M}}\, \mathbf{\Phi}_\mu^{\text{av}\intercal} Q$ is defined as the $N_{\text{av}} \times 1$ vector whose ith entry is $\inf_{\mu \in \mathcal{M}} Q(\mathbf{s}_i^{\text{av}}, \mu(\mathbf{s}_i^{\text{av}})) = \inf_{a_i \in \mathfrak{A}} Q(\mathbf{s}_i^{\text{av}}, a_i)$, and where the last equality is established because of the surjectivity of μ in the definition of $\mathcal{M}$. Let also the $N_{\text{av}} \times N_{\text{av}}$ kernel matrices $\mathbf{K}_\Psi := \mathbf{\Psi}^\intercal \mathbf{\Psi}$ and $\mathbf{K}_\mu^{\text{av}} := \mathbf{\Phi}_\mu^{\text{av}\intercal} \mathbf{\Phi}_\mu^{\text{av}}$. Vectors $\mathcal{S}_{\text{av}}$ may be also used to incorporate training data in (3).

The proposed B-Maps (3) serve as approximates of the classical B-Maps, where conditional expectations need not be computed, and the expected long-term loss in (2) is approximated by a sampling average in (3), that is, by a linear combination of $\{\psi_i\}_i$ with coefficients which are defined via samples of the long-term loss. The asymptotic consistency of a special class of (3) with the classical B-Maps, as the number of samples goes to infinity, is established formally in Theorem 4.

The proposed B-Maps are instances of the general strategy of "RL with function approximation" [18]. Both the one-step g and the long-term loss Q are considered to be elements of an RKHS $\mathcal{H}$ to capitalize on the rich geometry introduced by the reproducing inner product $\langle\cdot \mid \cdot\rangle_{\mathcal{H}}$ [21, 22]. For example, the samples of the long-term loss in (3) can be viewed as inner-product evaluations via the reproducing property $Q(\mathbf{s},a) = \langle Q \mid \varphi(\mathbf{s},a)\rangle_{\mathcal{H}} = \varphi^{\intercal}(\mathbf{s},a)Q$. This elementary observation establishes the latter formulations in (3a) and (3b), and will be also used later in (27) to derive simple recursions for Algorithm 1.

It is worth stressing here that the price to be paid for working in an RKHS, and not a more general Banach space, which is a typical case in prior-art designs (*cf.* Section II-C), is that an RKHS may fail to model and capture non-smooth and possibly non-continuous Q-functions, which can be instead modeled by the more general $\mathcal{L}_\infty$-Banach space of all essentially bounded functions. Nevertheless, a Banach space does not offer the convenience of an inner product.

To underline the ample degrees of freedom offered by the user-defined $\{\psi_i\}_i$, the following proposition demonstrates that by appropriately tuning the $\{\psi_i\}_i$ through a variational framework, the proposed B-Maps (3) yield as special cases the popular designs (7), (10), (15), as well as [32, (7)] and [33, (3)].

Proposition 1. (Variational framework for B-Maps) Consider the user-defined loss function $\mathcal{L}\colon \mathbb{R}^N \times \mathbb{R}^{N\times N_{\text{av}}} \to \mathbb{R}\colon (\boldsymbol{\gamma},\boldsymbol{\Upsilon}) \mapsto \mathcal{L}(\boldsymbol{\gamma},\boldsymbol{\Upsilon})$ and the regularizing function $\mathcal{R}\colon \mathbb{R}^N \times \mathbb{R}^{N\times N_{\text{av}}} \to \mathbb{R}\colon (\boldsymbol{\gamma},\boldsymbol{\Upsilon}) \mapsto \mathcal{R}(\boldsymbol{\gamma},\boldsymbol{\Upsilon})$, and let $(\boldsymbol{\gamma}_\star,\boldsymbol{\Upsilon}_\star)$ stand for the minimizers of the following variational problem:

$$(\boldsymbol{\gamma}_\star,\boldsymbol{\Upsilon}_\star) \in \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^N,\boldsymbol{\Upsilon}\in\mathbb{R}^{N\times N_{\text{av}}}} \mathcal{L}(\boldsymbol{\gamma},\boldsymbol{\Upsilon}) + \sigma\mathcal{R}(\boldsymbol{\gamma},\boldsymbol{\Upsilon})\,, \quad (4)$$

where $\sigma \in \mathbb{R}_+$, and where (4) is assumed to possess a solution.

(i) Consider a stationary policy μ. Let $\mu_\star \in \mathcal{M}$ be a stationary policy s.t. $\mu_\star(\mathbf{s}_i) \coloneqq a_i$ and $\mu_\star(\mathbf{s}'_i) \coloneqq \mu(\mathbf{s}'_i)$, $\forall i \in \{1,\ldots,N\}$. Define also $\mathbf{s}^{\text{av}}_i \coloneqq \mathbf{s}_i$, $\forall i \in \{1,\ldots,N\}$, and $\mathbf{s}^{\text{av}}_i \coloneqq \mathbf{s}'_{i-N}$, $\forall i \in \{N+1,\ldots,2N\}$, so that $\boldsymbol{\Phi}^{\text{av}}_{\mu_\star} = [\boldsymbol{\Phi}_{\mathcal{T}_N}, \boldsymbol{\Phi}'_\mu]$ and $N_{\text{av}} = 2N$. Define also

$$\mathcal{L}(\boldsymbol{\gamma},\boldsymbol{\Upsilon}) \coloneqq \|\mathbf{K}_{\mathcal{T}_N}(\boldsymbol{\gamma} + \alpha\boldsymbol{\Upsilon}\boldsymbol{\Phi}^{\text{av}\intercal}_{\mu_\star}Q) - \mathbf{g} - \alpha\boldsymbol{\Phi}'^{\intercal}_{\mu}Q\|^2_{\mathbb{R}^N}\,, \quad (5a)$$

$$\mathcal{R}(\boldsymbol{\gamma},\boldsymbol{\Upsilon}) \coloneqq (\boldsymbol{\gamma} + \alpha\boldsymbol{\Upsilon}\boldsymbol{\Phi}^{\text{av}\intercal}_{\mu_\star}Q - \mathbf{K}^{\dagger}_{\mathcal{T}_N}\boldsymbol{\Phi}^{\intercal}_{\mathcal{T}_N}Q)^{\intercal}\mathbf{K}_{\mathcal{T}_N} \cdot (\boldsymbol{\gamma} + \alpha\boldsymbol{\Upsilon}\boldsymbol{\Phi}^{\text{av}\intercal}_{\mu_\star}Q - \mathbf{K}^{\dagger}_{\mathcal{T}_N}\boldsymbol{\Phi}^{\intercal}_{\mathcal{T}_N}Q)\,, \quad (5b)$$

where $\mathbf{K}^{\dagger}_{\mathcal{T}_N}$ is the Moore-Penrose pseudoinverse of $\mathbf{K}_{\mathcal{T}_N}$, to verify that the solution to (4) is:

$$\boldsymbol{\gamma}_\star \coloneqq (\mathbf{K}_{\mathcal{T}_N} + \sigma\mathbf{I}_N)^{-1}\mathbf{g}\,, \quad (6a)$$

$$\boldsymbol{\Upsilon}_\star \coloneqq (\mathbf{K}_{\mathcal{T}_N} + \sigma\mathbf{I}_N)^{-1}[(\sigma/\alpha)\mathbf{K}^{\dagger}_{\mathcal{T}_N}, \mathbf{I}_N]\,. \quad (6b)$$

Then, the proposed $T_{\mu_\star}(Q)$ in (3a), with $g \coloneqq \boldsymbol{\Phi}_{\mathcal{T}_N}\boldsymbol{\gamma}_\star$ and $\boldsymbol{\Psi} \coloneqq \boldsymbol{\Phi}_{\mathcal{T}_N}\boldsymbol{\Upsilon}_\star$, yields the least-squares-policy-evaluation (LSPE) B-Map: for $Q = \boldsymbol{\Phi}_{\mathcal{T}_N}\mathbf{w}$ and any arbitrarily chosen $\mathbf{w} \in \mathbb{R}^N$,

$$\begin{aligned} &T_{\text{LSPE},\mu}(Q) \\ &\coloneqq \arg\min_{Q'\in\mathcal{H}} \sum\nolimits_{i=1}^{N} \big[Q'(\mathbf{z}_i) - g_i - \alpha Q(\mathbf{s}'_i,\mu(\mathbf{s}'_i))\big]^2 \\ &\qquad + \sigma\|Q'-Q\|^2_{\mathcal{H}} \qquad (7) \\ &= \boldsymbol{\Phi}_{\mathcal{T}_N}\boldsymbol{\gamma}_\star + \alpha\boldsymbol{\Phi}_{\mathcal{T}_N}\boldsymbol{\Upsilon}_\star\boldsymbol{\Phi}^{\text{av}\intercal}_{\mu_\star}Q \\ &= \boldsymbol{\Phi}_{\mathcal{T}_N}(\mathbf{K}_{\mathcal{T}_N} + \sigma\mathbf{I}_N)^{-1}\mathbf{g} \\ &\quad + \alpha\boldsymbol{\Phi}_{\mathcal{T}_N}(\mathbf{K}_{\mathcal{T}_N} + \sigma\mathbf{I}_N)^{-1}[(\sigma/\alpha)\mathbf{K}^{\dagger}_{\mathcal{T}_N}, \mathbf{I}_N] \\ &\qquad \cdot \begin{bmatrix} \boldsymbol{\Phi}^{\intercal}_{\mathcal{T}_N}Q \\ \boldsymbol{\Phi}'^{\intercal}_{\mu}Q \end{bmatrix}, \end{aligned}$$

which was originally introduced for Euclidean spaces via (7) in [34, 35], and extended for RKHSs in [36].

(ii) Consider a stationary policy μ, and define $\mu_\star$ and $\boldsymbol{\Phi}^{\text{av}}_{\mu_\star}$ as in Proposition 1(i). Define also the temporal-difference (TD) feature vectors $\boldsymbol{\Phi}_{\text{TD}} \coloneqq \boldsymbol{\Phi}_{\mathcal{T}_N} - \alpha\boldsymbol{\Phi}'_\mu$ and $\mathbf{K}_{\text{TD}} \coloneqq \boldsymbol{\Phi}^{\intercal}_{\text{TD}}\boldsymbol{\Phi}_{\text{TD}}$. Moreover, let

$$\mathcal{L}(\boldsymbol{\gamma},\boldsymbol{\Upsilon}) \coloneqq \|\mathbf{K}_{\text{TD}}(\boldsymbol{\gamma} + \alpha\boldsymbol{\Upsilon}\boldsymbol{\Phi}^{\text{av}\intercal}_{\mu_\star}Q) - \mathbf{g}\|^2_{\mathbb{R}^N}\,, \quad (8a)$$

$$\mathcal{R}(\boldsymbol{\gamma},\boldsymbol{\Upsilon}) \coloneqq (\boldsymbol{\gamma} + \alpha\boldsymbol{\Upsilon}\boldsymbol{\Phi}^{\text{av}\intercal}_{\mu_\star}Q - \mathbf{K}^{\dagger}_{\text{TD}}\boldsymbol{\Phi}^{\intercal}_{\text{TD}}Q)^{\intercal}\mathbf{K}_{\text{TD}} \cdot (\boldsymbol{\gamma} + \alpha\boldsymbol{\Upsilon}\boldsymbol{\Phi}^{\text{av}\intercal}_{\mu_\star}Q - \mathbf{K}^{\dagger}_{\text{TD}}\boldsymbol{\Phi}^{\intercal}_{\text{TD}}Q)\,. \quad (8b)$$

to verify that

$$\boldsymbol{\gamma}_\star \coloneqq (\mathbf{K}_{\text{TD}} + \sigma\mathbf{I}_N)^{-1}\mathbf{g}\,, \quad (9a)$$

$$\boldsymbol{\Upsilon}_\star \coloneqq (\mathbf{K}_{\text{TD}} + \sigma\mathbf{I}_N)^{-1}\mathbf{K}^{\dagger}_{\text{TD}}[(\sigma/\alpha)\mathbf{I}_N, -\sigma\mathbf{I}_N]\,, \quad (9b)$$

solve (4).

Then, the proposed $T_{\mu_\star}(Q)$ in (3a), with $g \coloneqq \boldsymbol{\Phi}_{\text{TD}}\boldsymbol{\gamma}_\star$ and $\boldsymbol{\Psi} \coloneqq \boldsymbol{\Phi}_{\text{TD}}\boldsymbol{\Upsilon}_\star$, yields the Bellman-residual (BR) B-Map: for $Q = \boldsymbol{\Phi}_{\text{TD}}\mathbf{w}$ and any arbitrarily chosen $\mathbf{w} \in \mathbb{R}^N$,

$$\begin{aligned} &T_{\text{BR},\mu}(Q) \\ &\coloneqq \arg\min_{Q'\in\mathcal{H}} \sum\nolimits_{i=1}^{N} \big[Q'(\mathbf{z}_i) - g_i - \alpha Q'(\mathbf{s}'_i,\mu(\mathbf{s}'_i))\big]^2 \\ &\qquad + \sigma\|Q'-Q\|^2_{\mathcal{H}} \qquad (10) \\ &= \boldsymbol{\Phi}_{\text{TD}}\boldsymbol{\gamma}_\star + \alpha\boldsymbol{\Phi}_{\text{TD}}\boldsymbol{\Upsilon}_\star\boldsymbol{\Phi}^{\text{av}\intercal}_{\mu_\star}Q \\ &= \boldsymbol{\Phi}_{\text{TD}}(\mathbf{K}_{\text{TD}} + \sigma\mathbf{I}_N)^{-1}\mathbf{g} \\ &\quad + \alpha\boldsymbol{\Phi}_{\text{TD}}(\mathbf{K}_{\text{TD}} + \sigma\mathbf{I}_N)^{-1}\mathbf{K}^{\dagger}_{\text{TD}}[(\sigma/\alpha)\mathbf{I}_N, -\sigma\mathbf{I}_N] \\ &\qquad \cdot \begin{bmatrix} \boldsymbol{\Phi}^{\intercal}_{\mathcal{T}_N}Q \\ \boldsymbol{\Phi}'^{\intercal}_{\mu}Q \end{bmatrix}, \end{aligned}$$

which was originally presented via (10) in [37, 38].

(iii) Given a stationary policy $\mu \in \mathcal{M}$, let $\mu_\star \coloneqq \mu$, $\mathbf{s}^{\text{av}}_i \coloneqq \mathbf{s}'_i$, $\forall i \in \{1,\ldots,N\}$, and $\boldsymbol{\Phi}^{\text{av}}_{\mu_\star} \coloneqq \boldsymbol{\Phi}'_\mu$. Define also

$$\mathcal{L}(\boldsymbol{\gamma},\boldsymbol{\Upsilon}) \coloneqq \|\mathbf{K}_{\mathcal{T}_N}(\boldsymbol{\gamma} + \alpha\boldsymbol{\Upsilon}\boldsymbol{\Phi}^{\text{av}\intercal}_{\mu_\star}Q) - \mathbf{g} - \alpha\boldsymbol{\Phi}'^{\intercal}_{\mu}Q\|^2_{\mathbb{R}^N}\,, \quad (11a)$$

$$\mathcal{R}(\boldsymbol{\gamma},\boldsymbol{\Upsilon}) \coloneqq (\boldsymbol{\gamma} + \alpha\boldsymbol{\Upsilon}\boldsymbol{\Phi}^{\text{av}\intercal}_{\mu_\star}Q)^{\intercal}\mathbf{K}_{\mathcal{T}_N}(\boldsymbol{\gamma} + \alpha\boldsymbol{\Upsilon}\boldsymbol{\Phi}^{\text{av}\intercal}_{\mu_\star}Q)\,, \quad (11b)$$

to verify that

$$\boldsymbol{\gamma}_\star \coloneqq (\mathbf{K}_{\mathcal{T}_N} + \sigma\mathbf{I}_N)^{-1}\mathbf{g}\,, \quad (12a)$$

$$\boldsymbol{\Upsilon}_\star \coloneqq (\mathbf{K}_{\mathcal{T}_N} + \sigma\mathbf{I}_N)^{-1}\,, \quad (12b)$$

solve (4).

Then, the proposed $T_{\mu_\star}(Q)$ in (3a), with $g \coloneqq \mathbf{\Phi}_{\mathcal{T}_N}\boldsymbol{\gamma}_\star$ and $\mathbf{\Psi} \coloneqq \mathbf{\Phi}_{\mathcal{T}_N}\mathbf{\Upsilon}_\star$, takes the following form:

$$\begin{aligned} T_{\mu_\star}(Q) &= \mathbf{\Phi}_{\mathcal{T}_N}\boldsymbol{\gamma}_\star + \alpha\mathbf{\Phi}_{\mathcal{T}_N}\mathbf{\Upsilon}_\star\mathbf{\Phi}_\mu^{\mathrm{av}\intercal}Q \\ &= \mathbf{\Phi}_{\mathcal{T}_N}(\mathbf{K}_{\mathcal{T}_N}+\sigma\mathbf{I}_N)^{-1}\mathbf{g} \\ &\quad + \alpha\mathbf{\Phi}_{\mathcal{T}_N}(\mathbf{K}_{\mathcal{T}_N}+\sigma\mathbf{I}_N)^{-1}\mathbf{\Phi}_\mu^{\mathrm{av}\intercal}Q\,. \end{aligned} \tag{13}$$

Proof: See Appendix A. ∎

LSPE in Proposition 1(i) shows strong connections with the classical temporal difference (TD) [19, 35, 39–41], where the TD recursion (Q-learning) is an SGD step on the loss in (7) for $N=1$. The popular LS temporal difference (LSTD) [26, 42–45] computes a fixed point $Q_{\mathrm{LSTD},\mu} \in \mathrm{Fix}\, T_{\mathrm{LSPE},\mu} \cap \mathrm{span}\,\{\varphi(\mathbf{z}_i)\}_{i=1}^N$, where

$$\begin{aligned} &\mathrm{Fix}\, T_{\mathrm{LSPE},\mu} \cap \mathrm{span}\,\{\varphi(\mathbf{z}_i)\}_{i=1}^N \\ &= \big\{\, Q \in \mathrm{span}\,\{\varphi(\mathbf{z}_i)\}_{i=1}^N \,\big|\, T_{\mathrm{LSPE},\mu}Q = Q \,\big\} \\ &= \{\, Q \in \mathrm{span}\,\{\varphi(\mathbf{z}_i)\}_{i=1}^N \mid (\mathbf{\Phi}_{\mathcal{T}_N}\mathbf{\Phi}_{\mathcal{T}_N}^\intercal - \alpha\mathbf{\Phi}_{\mathcal{T}_N}\mathbf{\Phi}_\mu'^\intercal)Q \\ &\qquad\qquad = \mathbf{\Phi}_{\mathcal{T}_N}\mathbf{g}\,\}\,; \end{aligned} \tag{14}$$

see Appendix A for a proof of (14). That fixed point becomes unique whenever $\mathbf{K}_{\mathcal{T}_N} - \alpha\mathbf{\Phi}_\mu'^\intercal\mathbf{\Phi}_{\mathcal{T}_N}$ is invertible:

$$Q_{\mathrm{LSTD},\mu} = \mathbf{\Phi}_{\mathcal{T}_N}(\mathbf{K}_{\mathcal{T}_N} - \alpha\mathbf{\Phi}_\mu'^\intercal\mathbf{\Phi}_{\mathcal{T}_N})^{-1}\mathbf{g}\,, \tag{15}$$

where (15) follows easily from (14) after using $(\mathbf{\Phi}_{\mathcal{T}_N}\mathbf{\Phi}_{\mathcal{T}_N}^\intercal - \alpha\mathbf{\Phi}_{\mathcal{T}_N}\mathbf{\Phi}_\mu'^\intercal)^{-1}\mathbf{\Phi}_{\mathcal{T}_N} = \mathbf{\Phi}_{\mathcal{T}_N}(\mathbf{K}_{\mathcal{T}_N} - \alpha\mathbf{\Phi}_\mu'^\intercal\mathbf{\Phi}_{\mathcal{T}_N})^{-1}$. Interestingly, it has been demonstrated that LSPE/LSTD perform better in general than TD in numerical tests [44].

Mapping $T_{\mathrm{BR},\mu}$ in Proposition 1(ii) can be viewed as the popular proximal mapping defined by $\mathrm{Prox}_{f/(2\sigma)}(Q) \coloneqq \arg\min_{Q'\in\mathcal{H}} f(Q') + 2\sigma\cdot(1/2)\|Q-Q'\|_{\mathcal{H}}^2$ [31], with $f(Q') \coloneqq \sum_{i=1}^N [Q'(\mathbf{z}_i) - g_i - \alpha Q'(\mathbf{s}_i', \mu(\mathbf{s}_i'))]^2$. Extensions to cases where the loss in (10) is further regularized by additional convex functions, such as the ℓ_1-norm loss for example to impose structure onto the desired solutions, can be found in [46–48].

Operator $\mathbf{\Phi}_{\mathcal{T}_N}(\mathbf{K}_{\mathcal{T}_N}+\sigma\mathbf{I}_N)^{-1}\mathbf{\Phi}_\mu^{\mathrm{av}\intercal}$ of (13) appears also in [32, (7)] and [33, (3)], where RKHSs are used as approximating spaces for conditional expectations via distribution embeddings in the following sense: under certain conditions, the existence of an $h_{(\mathbf{s},a)}^\mu \in \mathcal{H}$ such that $\langle Q \mid h_{(\mathbf{s},a)}^\mu\rangle_{\mathcal{H}} = \mathbb{E}_{\mathbf{s}'|(\mathbf{s},a)}\{Q(\mathbf{s}', \mu(\mathbf{s}'))\}$, $\forall (\mathbf{s},a) \in \mathfrak{Z}$, is guaranteed [33, 49, 50]. By tailoring the arguments of [33] to the current context, a B-Map can be defined as $\mathcal{H} \ni Q \mapsto T_{\mathrm{emb},\mu}(Q)(\mathbf{s},a) \coloneqq g(\mathbf{s},a) + \alpha\langle Q \mid \hat{h}_{(\mathbf{s},a)}^\mu\rangle_{\mathcal{H}}$, where

$$\hat{h}_{(\mathbf{s},a)}^\mu \coloneqq \sum\nolimits_{i=1}^N \frac{c_i(\mathbf{s},a)}{\sum_{j=1}^N |c_j(\mathbf{s},a)|}\varphi(\mathbf{s}_i', \mu(\mathbf{s}_i')) \in \mathcal{H}$$

serves as an estimate for the unknown $h_{(\mathbf{s},a)}^\mu$, with $\mathbf{c}(\mathbf{s},a) \coloneqq [c_1(\mathbf{s},a), c_2(\mathbf{s},a), \ldots, c_N(\mathbf{s},a)]^\intercal \coloneqq (\mathbf{K}_{\mathcal{T}_N} + \sigma\mathbf{I}_N)^{-1}\mathbf{\Phi}_{\mathcal{T}_N}^\intercal\varphi(\mathbf{s},a)$. Due to the existence of the denominator in the definition of $\hat{h}_{(\mathbf{s},a)}^\mu$, $T_{\mathrm{emb},\mu}(Q)$ is not guaranteed in general to belong to $\mathcal{H}$, even if $Q \in \mathcal{H}$. As such, $T_{\mathrm{emb},\mu}(Q)$ is treated as an element of a Banach space (of all essentially bounded functions) in [33, (6), (7)]. Moreover, notice that $\mathbf{c}(\mathbf{s},a)$ needs to be computed at each point $(\mathbf{s},a) \in \mathfrak{Z}$, which poses computational obstacles in cases where $\mathfrak{Z}$ is either continuous or of massive cardinality. Further, the previous idea to approximate conditional expectations by inner products do not seem to work smoothly in the case of (2b), because the hypothetical existence of an $h_{(\mathbf{s},a)}$ that satisfies $\langle Q \mid h_{(\mathbf{s},a)}\rangle_{\mathcal{H}} = \mathbb{E}_{\mathbf{s}'|(\mathbf{s},a)}\{\inf_{a'\in\mathfrak{A}} Q(\mathbf{s}', a')\}$, and the linearity of the inner product, would suggest that the previous conditional expectation is a linear function of Q; however, this is not true in general due to the existence of the $\inf$ operator.

Motivated by the Nadaraya-Watson kernel estimate [23], and for a non-negative and not necessarily reproducing kernel function $\chi(\cdot,\cdot)\colon \mathfrak{S}\times\mathfrak{S} \to \mathbb{R}$, another major prior-art path is kernel-based (KB)RL [24, 25], built on the following B-Maps: $\forall (\mathbf{s},a) \in \mathfrak{Z}$,

$$T_{\mathrm{KBRL},\mu}(Q)(\mathbf{s},a) \coloneqq \sum_{(\mathbf{s}_i,a_i,\mathbf{s}_i')\in\mathcal{T}_N^a} \chi(\mathbf{s},\mathbf{s}_i)\big(g_i + \alpha Q(\mathbf{s}_i', \mu(\mathbf{s}_i'))\big)\,, \tag{16a}$$

$$T_{\mathrm{KBRL}}(Q)(\mathbf{s},a) \coloneqq \sum_{(\mathbf{s}_i,a_i,\mathbf{s}_i')\in\mathcal{T}_N^a} \chi(\mathbf{s},\mathbf{s}_i)\big(g_i + \alpha \inf_{a'\in\mathfrak{A}} Q(\mathbf{s}_i', a')\big)\,, \tag{16b}$$

where $\mathcal{T}_N^a \coloneqq \{(\mathbf{s}_i, a_i, \mathbf{s}_i') \in \mathcal{T}_N \,|\, a_i = a\}$, $\mathcal{T}_N$ is typically considered to comprise "historical" (training) trajectory data, $g_i \coloneqq g(\mathbf{s}_i, a_i, \mathbf{s}_i')$, and χ needs to satisfy $\sum_{(\mathbf{s}_i,a_i,\mathbf{s}_i')\in\mathcal{T}_N^a}\chi(\mathbf{s},\mathbf{s}_i) = 1$. Following [24], a simple way to enforce the previous constraint on χ for every $\mathbf{s}\in\mathfrak{S}$ is via another non-negative "mother" kernel function $\zeta(\cdot,\cdot)$:

$$\chi(\mathbf{s},\mathbf{s}_i) \coloneqq \frac{\zeta(\mathbf{s},\mathbf{s}_i)}{\sum_{(\mathbf{s}_j,a_j,\mathbf{s}_j')\in\mathcal{T}_N^a}\zeta(\mathbf{s},\mathbf{s}_j)}\,. \tag{17}$$

Even if ζ is a reproducing kernel of an RKHS $\mathcal{H}$ and $Q \in \mathcal{H}$, due to the denominator of (17), there is no guarantee, in general, that χ, $T_{\mathrm{KBRL},\mu}(Q)$, and $T_{\mathrm{KBRL}}(Q)$ belong to $\mathcal{H}$ [24]. For such a reason, the discussion in [24, 25] stays in a Banach space (of all essentially bounded functions), with no use of an RKHS inner product. KBRL mappings (16) have been also adopted in [51–54].

More variations of (3) can be generated from (4) by tuning the loss functions $\mathcal{L}, \mathcal{R}$ appropriately. For example, robust B-Maps against outliers in sampling can be designed by letting the ℓ_1-norm take the place of the quadratic one in (5a) and (8a). Task (4) for general (non)smooth convex $\mathcal{L}$ and $\mathcal{R}$ can be handled efficiently by [55]. Such designs are deferred to future publications.

D. Properties of the proposed B-Maps

Several properties of the proposed mappings (3) are now in order. First, it is demonstrated that the proposed B-Maps (3) are continuous.

Theorem 2. (Lipschitz continuity) Mappings (3) are Lipschitz continuous: $\forall Q_1, Q_2 \in \mathcal{H}$,

$$\|T_\mu(Q_1) - T_\mu(Q_2)\|_{\mathcal{H}} \le \beta\|Q_1 - Q_2\|_{\mathcal{H}}\,, \tag{18a}$$

$$\|T(Q_1) - T(Q_2)\|_{\mathcal{H}} \le \beta\|Q_1 - Q_2\|_{\mathcal{H}}\,, \tag{18b}$$

where

$$\beta \coloneqq \alpha\left(\|\mathbf{K}_\Psi\|_2 \sup_{\mu'\in\mathcal{M}}\|\mathbf{K}_{\mu'}^{\mathrm{av}}\|_2\right)^{1/2}\,, \tag{19}$$

and $\|\cdot\|_2$ stands for the spectral norm of a matrix. Hence, if $\beta = 1$, mappings (3) are nonexpansive [31], whereas, if $\beta < 1$, they are contractions [56] in $(\mathcal{H}, \langle\cdot \mid \cdot\rangle_{\mathcal{H}})$.

Proof: See Appendix B. ∎

To state Theorem 4, a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is necessary, with sample space Ω, σ-algebra $\mathcal{F}$ of events, and probability measure $\mathbb{P}$ [30]. A statement $(\ldots)$ will be said to hold true almost surely (a.s.), if $(\ldots)$ holds true on an event $\mathcal{E} \in \mathcal{F}$ with $\mathbb{P}(\mathcal{E}) = 1$. Moreover, by a slight abuse of terminology, a bounded linear and self-adjoint mapping $\mathcal{A}\colon \mathcal{H} \to \mathcal{H}$ will be called positive definite, if its minimum spectral value $\sigma_{\min}(\mathcal{A}) > 0$, where $\sigma_{\min}(\cdot)$ is defined by (47a).

Assumptions 3.

- **(i)** RKHS $\mathcal{H}$ is separable, for a stationary policy $\mu(\cdot) \in \mathcal{M}$ operators $\Sigma_{zz}, \Sigma^{\mu}_{s'z}, \Sigma^{\mu}_{s'|z}$, defined by (46), are bounded linear, Σ_{zz} of (46a) is positive definite, and $\Sigma_{zz}^{-3/2}\Sigma^{\mu}_{s'z}$ of (46c) is Hilbert-Schmidt (*cf.* Theorem 12 in the supplementary file).
- **(ii)** Let $\mathbf{s}, a, \mathbf{s}'$ be random variables (RVs) on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and assume that trajectory points $\{(\mathbf{s}_i, a_i, \mathbf{s}'_i)\}_{i=1}^N$ are also RVs, but independent and identically distributed (IID) copies of $(\mathbf{s}, a, \mathbf{s}')$.
- **(iii)** Motivated by Proposition 1(iii) and [32, (7)], let $N_{\text{av}} = N$, set $\mathbf{s}^{\text{av}}_i \coloneqq \mathbf{s}'_i$, $\forall i \in \{1, \ldots, N\}$, and let (3) take the following special form: $\forall Q \in \mathcal{H}$,

$$T_\mu(Q) = g + \alpha \tfrac{1}{N}\boldsymbol{\Phi}_{\mathcal{T}_N}(\tfrac{1}{N}\mathbf{K}_{\mathcal{T}_N} + \sigma'_N\mathbf{I}_N)^{-1}\boldsymbol{\Phi}^{\text{av}\intercal}_{\mu} Q \tag{20a}$$

$$= g + \alpha(\tfrac{1}{N}\boldsymbol{\Phi}_{\mathcal{T}_N}\boldsymbol{\Phi}^{\intercal}_{\mathcal{T}_N} + \sigma'_N \operatorname{Id})^{-1}\tfrac{1}{\sqrt{N}}\boldsymbol{\Phi}_{\mathcal{T}_N} \cdot \tfrac{1}{\sqrt{N}}\boldsymbol{\Phi}^{\text{av}\intercal}_{\mu} Q\,, \tag{20b}$$

$$T(Q) = g + \alpha \tfrac{1}{N}\boldsymbol{\Phi}_{\mathcal{T}_N}(\tfrac{1}{N}\mathbf{K}_{\mathcal{T}_N} + \sigma'_N\mathbf{I}_N)^{-1}\, \mathbf{inf}_{\mu\in\mathcal{M}}\, \boldsymbol{\Phi}^{\text{av}\intercal}_{\mu} Q \tag{20c}$$

$$= g + \alpha(\tfrac{1}{N}\boldsymbol{\Phi}_{\mathcal{T}_N}\boldsymbol{\Phi}^{\intercal}_{\mathcal{T}_N} + \sigma'_N \operatorname{Id})^{-1}\tfrac{1}{\sqrt{N}}\boldsymbol{\Phi}_{\mathcal{T}_N} \cdot \mathbf{inf}_{\mu\in\mathcal{M}}\, \tfrac{1}{\sqrt{N}}\boldsymbol{\Phi}^{\text{av}\intercal}_{\mu} Q\,, \tag{20d}$$

  where $\sigma'_N \in \mathbb{R}_{++}$ is a regularization coefficient, dependent on N, Id is the identity operator in $\mathcal{H}$, and $\boldsymbol{\Phi}_{\mathcal{T}_N}(\boldsymbol{\Phi}^{\intercal}_{\mathcal{T}_N}\boldsymbol{\Phi}_{\mathcal{T}_N}/N + \sigma'_N\mathbf{I}_N)^{-1} = (\boldsymbol{\Phi}_{\mathcal{T}_N}\boldsymbol{\Phi}^{\intercal}_{\mathcal{T}_N}/N + \sigma'_N \operatorname{Id})^{-1}\boldsymbol{\Phi}_{\mathcal{T}_N}$ was used in (20).
- **(iv)** $\lim_{N\to\infty}\sigma'_N = 0$ and $\lim_{N\to\infty} N\sigma'^3_N = +\infty$; *e.g.,* $\sigma'_N = N^{-\tau}$, with $\tau \in (0, 1/3)$.
- **(v)** The $\inf$ operators in (2b) and (20c) are achievable.
- **(vi)** There exists $\beta_\infty \in (0, 1)$ s.t. $\beta = \beta(N)$ in (19) satisfies $\beta(N) \leq \beta_\infty$, $\forall N$, a.s.

Assumptions 3(i) and 3(ii) follow [32], whose arguments are used to construct B-Maps in [33]; see the discussion regarding Proposition 1(iii). Assumption 3(ii) expresses the need for a sufficiently large number of IID samples of the triplet $(\mathbf{s}, a, \mathbf{s}')$ for the sampling average to approximate arbitrarily well the conditional expectation operator in (2) via the law of large numbers. Recall from Section II-A that this study does not necessitate $\mathbf{s}_{i+1} = \mathbf{s}'_i$ in $\mathcal{T}_N = \{(\mathbf{s}_i, a_i, \mathbf{s}'_i)\}_{i=1}^N$, to allow for looser assumptions than the Markovian one in stochastic analyses, as Assumption 3(ii) demonstrates. Assumption 3(v) holds true trivially in the case where the actions space $\mathfrak{A}$ is of finite cardinality.

Careful design of $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}^{\text{av}}_{\mu}$ is needed for Assumption 3(vi) to hold true; especially if $\beta(N)$ is desired to have values close to 1 for all sufficiently large N. A detailed discussion on how to design such $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}^{\text{av}}_{\mu}$, with theoretical guarantees and numerical tests for validation, is deferred to a future work. Nevertheless, few remarks are necessary here to sketch the guiding theoretical arguments of such a construction. In the light of (20b) and (20d), upon defining the operator $\boldsymbol{\Psi} \coloneqq (\boldsymbol{\Phi}_{\mathcal{T}_N}\boldsymbol{\Phi}^{\intercal}_{\mathcal{T}_N}/N + \sigma'_N \operatorname{Id})^{-1}(\boldsymbol{\Phi}_{\mathcal{T}_N}/\sqrt{N})\colon \mathbb{R}^N \to \mathcal{H}$, and redefining $\boldsymbol{\Phi}^{\text{av}}_{\mu}$ as $\boldsymbol{\Phi}^{\text{av}}_{\mu}/\sqrt{N}$ to include the scaling factor $1/\sqrt{N}$, notice that $\|\mathbf{K}_\Psi\|_2 = \|\boldsymbol{\Psi}^{\intercal}\boldsymbol{\Psi}\|_2 = \|\boldsymbol{\Psi}\|^2$ and $\|\mathbf{K}^{\text{av}}_{\mu}\|_2 = \|\boldsymbol{\Phi}^{\text{av}}_{\mu}\boldsymbol{\Phi}^{\text{av}\intercal}_{\mu}/N\| = \|\sum_{i=1}^N \varphi(\mathbf{s}^{\text{av}}_i, \mu(\mathbf{s}^{\text{av}}_i))\,\varphi^{\intercal}(\mathbf{s}^{\text{av}}_i, \mu(\mathbf{s}^{\text{av}}_i))/N\|$ [56, Thm. 3.9-4(e)]. Motivated now by the celebrated Tikhonov regularization and the result that for a matrix $\mathbf{A}$, $\lim_{\sigma'\to 0}(\mathbf{A}\mathbf{A}^{\intercal} + \sigma'\mathbf{I})^{-1}\mathbf{A} = \mathbf{A}^{\intercal\dagger}$ [57], [58, §3.3], it is conceivable that $\boldsymbol{\Psi}$ converges in some probabilistic sense to an operator $\boldsymbol{\Phi}^{\intercal\dagger}_{\mathcal{T}_\infty}$ as $N \to \infty$, where $\dagger$ denotes the pseudoinverse, provided that the sampling average $\boldsymbol{\Phi}_{\mathcal{T}_N}\boldsymbol{\Phi}^{\intercal}_{\mathcal{T}_N}/N$ converges at a much faster rate than $\sigma'_N \to 0$; *cf.* Assumption 3(iv). It is valid then to anticipate that $\|\mathbf{K}_\Psi\|_2$ converges to $\|\boldsymbol{\Phi}^{\dagger}_{\mathcal{T}_\infty}\|^2$ as $N \to \infty$. Moreover, it is conceivable that the sampling average $\boldsymbol{\Phi}^{\text{av}}_{\mu}\boldsymbol{\Phi}^{\text{av}\intercal}_{\mu}/N$ converges in some probabilistic sense to an operator $\mathbf{C}^{\text{av}}_{\mu}$, so that $\|\mathbf{K}^{\text{av}}_{\mu}\|_2$ converges to $\|\mathbf{C}^{\text{av}}_{\mu}\|$ as $N \to \infty$. To summarize, $\beta(N)$ can be assumed to converge to $\alpha\|\boldsymbol{\Phi}^{\dagger}_{\mathcal{T}_\infty}\| \sup_{\mu\in\mathcal{M}}\|\mathbf{C}^{\text{av}}_{\mu}\|^{1/2}$ as $N \to \infty$. If the column vectors of $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}^{\text{av}}_{\mu}$ are carefully constructed, for example, to be nearly orthonormal in an infinite dimensional RKHS, *e.g.*, Gaussian kernel, for some μ and for all sufficiently large N, then it is valid to anticipate that $\|\boldsymbol{\Phi}^{\dagger}_{\mathcal{T}_\infty}\| \sup_{\mu\in\mathcal{M}}\|\mathbf{C}^{\text{av}}_{\mu}\|^{1/2}$ takes values close to 1.

The following Theorem 4 establishes the asymptotic consistency of the fixed points of the B-Maps introduced in Proposition 1(iii) with those of the classical B-Maps (2), as the number of samples goes to infinity. The discussion is motivated by [32], where conditional expectations are approximated by inner products; see also the related discussion after Proposition 1.

Theorem 4. (Consistency of fixed points) Under Assumptions 3, $T^{\diamond}_{\mu}, T^{\diamond}$ in (2) and T_μ, T in (20) are contractions in the Hilbert space $\mathcal{H}$, and thus possess unique fixed points $Q^{\diamond}_{\mu}, Q^{\diamond}_{*}, Q_\mu, Q_*$, respectively, a.s. Notice that Q_μ, Q_* depend on N, *i.e.*, $Q_\mu = Q_\mu(N)$ and $Q_* = Q_*(N)$. Furthermore,

$$\mathbb{P}\text{-}\lim_{N\to\infty}\|Q^{\diamond}_{\mu} - Q_\mu(N)\|_{\mathcal{H}} = 0\,, \tag{21a}$$

$$\mathbb{P}\text{-}\lim_{N\to\infty}\|Q^{\diamond}_{*} - Q_*(N)\|_{\mathcal{H}} = 0\,, \tag{21b}$$

where $\mathbb{P}$-lim stands for convergence in probability [30].

Proof: See Appendix C. ∎

III. Application to Robust Adaptive Filtering

The following discussion applies the novel B-Maps of Section II to the setting of Section I-A. To abide by the online or time-adaptive premise of Section I-A, the arguments of Section II will be equipped hereafter with a discrete time index $n \in \mathbb{N}$. It is important to note that n serves also as the

Algorithm 1 Approximate policy iteration for LMP

1: Arbitrarily initialize $\boldsymbol{\theta}_0$, Q_0, and $\mathbf{s}_{-1}$.
2: **while** $n \in \mathbb{N}$ **do**
3: Data $(\mathbf{x}_n, y_n)$ become available to the user/agent.
4: New state $\mathbf{s}_n$ is defined by (23).
5: **Policy improvement:** Let $a_n \coloneqq \mu_n(\mathbf{s}_n)$ by (24).
6: Compute $\boldsymbol{\theta}_{n+1}$ by (22), with $p_n \coloneqq a_n$.
7: **Policy evaluation:** Compute Q_{n+1} by (33) and (34).
8: Increase n by one, and go to Section III.
9: **end while**

iteration index of the proposed RL-based Algorithm 1. Index n appears in the following discussion in various forms, such as a sub-/super-script, or as $[n]$. For example, N of Section II-C becomes $N[n]$ from now and on to highlight the fact that N depends on n.

For convenience, navigation directions to key RL quantities are provided here. The state vector is defined by (23), action is defined as any point from a finite grid of the interval $[1, 2]$, while the one-step loss is introduced in (29). Although (3) and Proposition 1 introduce considerable freedom in designing B-Maps, this manuscript focuses on the setting of Proposition 1(iii) to avoid lengthy discussions. The online version $T_{\mu_n}^{(n)}$ of the mapping in Proposition 1(iii) is presented in (26). Variants as well as alternatives of the proposed B-Map are deferred to future publications.

Algorithm 1 offers an RL way to robustify LMP (1) by letting the data themselves select the "optimal" p_n per time n (*cf.* Section III of Algorithm 1), without any assumptions and prior knowledge on the statistics of the outliers. More specifically, instead of (1),

$$\boldsymbol{\theta}_{n+1} \coloneqq \boldsymbol{\theta}_{n+1}(a_n) \coloneqq \boldsymbol{\theta}_n + \rho p_n \operatorname{sgn}(e_n) |e_n|^{p_n - 1} \mathbf{x}_n \,, \quad (22)$$

where e_n is defined after (1). It is clear by (22) and Section III of Algorithm 1 that $\boldsymbol{\theta}_{n+1}$ depends on the action $a_n = p_n$. To highlight this observation, $\boldsymbol{\theta}_{n+1}(a_n)$ is used together with $\boldsymbol{\theta}_{n+1}$ in (22), as well as in the following discussion.

Algorithm 1 belongs to the class of *policy-iteration (PI)* algorithms of RL [18]. More precisely, it is an *approximate* (A)PI algorithm, because the expectation operators in (2) are approximated by sample averaging in Proposition 1(iii). Typically, (A)PI comprises two major steps: policy improvement in Section III and policy evaluation in Section III. The following discussion details Algorithm 1.

A. State-action space and policy improvement

This subsection refers to Section III of Algorithm 1. Action space $\mathfrak{A}$ is defined as any finite grid of the interval $[1, 2]$, and it is the domain p_n in (22) takes values from. The more general case of a continuous action space is currently under study and deferred to a future publication. The state space is the continuous $\mathfrak{S} \coloneqq \mathbb{R}^4$, with the dimension of $\mathfrak{S}$ rendered independent of the filter length L. The state-action space is denoted by $\mathfrak{Z} \coloneqq \mathfrak{S} \times \mathfrak{A} \coloneqq \{\mathbf{z} \coloneqq (\mathbf{s}, a) \,|\, \mathbf{s} \in \mathfrak{S}, a \in \mathfrak{A}\}$.

To help the agent take a meaningful decision/action a_n at state $\mathbf{s}_n$, and transition to the new state $\mathbf{s}'_n =: \mathbf{s}_{n+1}$, sufficiently useful information should be packed into the low-dimensional $\mathbf{s}_n \in \mathbb{R}^4$. To define an adequate $\mathbf{s}_n$, available to the user at time n are considered to be data $\mathfrak{D}_{(n-M_{\text{av}}):n} \coloneqq ((\mathbf{x}_\nu, y_\nu))_{\nu=n-M_{\text{av}}}^{n}$, for some buffer length $M_{\text{av}} \in \mathbb{R}_{++}$, as well as estimates $(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n-1})$. Therefore, after taking action a_{n-1} at state $\mathbf{s}_{n-1}$, state $\mathbf{s}_n$ is defined inductively as:

$$\mathbf{s}_n \coloneqq \mathbf{s}'_{n-1} = [s_{n-1}^{\prime(1)}, s_{n-1}^{\prime(2)}, s_{n-1}^{\prime(3)}, s_{n-1}^{\prime(4)}]^{\mathsf{T}} \,, \quad (23a)$$

$$s_{n-1}^{\prime(1)} \coloneqq \log |e_n|^2 \,, \quad (23b)$$

$$s_{n-1}^{\prime(2)} \coloneqq \tfrac{1}{M_{\text{av}}} \sum_{m=1}^{M_{\text{av}}} \log \frac{|y_{n-m} - \boldsymbol{\theta}_n^{\mathsf{T}}(a_{n-1})\, \mathbf{x}_{n-m}|^2}{\|\mathbf{x}_{n-m}\|_2^2} \,, \quad (23c)$$

$$s_{n-1}^{\prime(3)} \coloneqq \log \|\mathbf{x}_n\|_2 \,, \quad (23d)$$

$$\begin{aligned} s_{n-1}^{\prime(4)} &\coloneqq \varpi s_{n-1}^{(4)} + (1 - \varpi) \log(\tfrac{1}{\rho} \|\boldsymbol{\theta}_n(a_{n-1}) - \boldsymbol{\theta}_{n-1}\|_2) \\ &= \varpi s_{n-1}^{(4)} + (1 - \varpi)(p_{n-1} - 1) \log|e_{n-1}| \\ &\quad + (1 - \varpi) \log \|\mathbf{x}_{n-1}\|_2 \\ &\quad + (1 - \varpi) \log p_{n-1} \,, \end{aligned} \quad (23e)$$

with $\varpi \in (0, 1)$ being a user-defined parameter, while ρ comes from (22). The classical a-priori error in AdaFilt [1, (10.11)] is used in (23b), an M_{av}-length sliding-window sampling average of the a-posteriori error [1, (10.12)] is provided in (23c), normalized by the norm of the input signal to remove as much as possible its effect on the error, the instantaneous norm of the input signal in (23d), and a smoothing auto-regressive process in (23e) to monitor the consecutive displacement of the estimates $(\boldsymbol{\theta}_n)_{n \in \mathbb{N}}$. The reason for including ρ in (23e) is to remove ρ's effect from $s_4^{(n)}$. The $\log(\cdot)$ function is employed in (23) to decrease the dynamic range of the positive values in (23). Any logarithmic function can be used in (23); the 10-base one is used in Section V.

Although the $(2L + 1)$-dimensional vector $(\mathbf{x}_n, \mathbf{y}_n, \boldsymbol{\theta}_n)$ would be a more natural choice for a state vector than the heuristic (23), because it would induce a proper Markov decision process (MDP) and would fall under the umbrella of typical RL designs [18], extensive numerical tests along the lines of Section V have shown that Algorithm 1 with the state vector $(\mathbf{x}_n, \mathbf{y}_n, \boldsymbol{\theta}_n)$ yields slow convergence speed due to the large dimensionality of the state space, especially in cases where the length L of the unknown system $\boldsymbol{\theta}_*$ is large ("curse of dimensionality"). Notice also that the state space in [17] is similarly defined to be $\mathbb{R}^{2L+1}$. Motivated by this observation, the choice of the state vector in (23) reflects the effort to reduce the dimensionality of the state space. Similar approaches, which do not adhere to typical MDP settings but tailor state-action spaces to fit the application at hand and meet design requirements, are not seldom in the literature of RL, *e.g.*, [59]. Other ways than (23) to reduce the dimensionality of the state space are currently under consideration.

With the estimate Q_n available to the user/agent, policy improvement in Section III is achieved by the standard greedy rule [18]

$$\mu_n(\mathbf{s}) \coloneqq \arg\min_{a \in \mathfrak{A}} Q_n(\mathbf{s}, a) \,, \quad \forall \mathbf{s} \in \mathfrak{S} \,. \quad (24)$$

More specifically, the next action a_n for the agent is identified by plugging $\mathbf{s}_n$ in the place of $\mathbf{s}$ in (24). Now that $a_n \coloneqq$

$\mu_n(\mathbf{s}_n)$ is available, recursion (22) is applied with $p_n \coloneqq a_n$ to Section III of Algorithm 1 to obtain the new estimate $\boldsymbol{\theta}_{n+1}$.

B. Defining the one-step loss, $T^{(n)}_{\mu_n}$, and loss $\mathcal{L}^{(n)}_{\mu_n}[\cdot]$

This subsection defines the online version $T^{(n)}_{\mu_n}$ of the B-Map discussed in Proposition 1(iii). State-action pair $\mathbf{z}_n = (\mathbf{s}_n, a_n)$ and stationary policy μ_n are now available to the user/agent by the discussion in Section III-A, and therefore, trajectory samples $\mathcal{T}^{(n)}_{N[n]} \coloneqq \{(\mathbf{s}_i[n], a_i[n], \mathbf{s}'_i[n])\}_{i=1}^{N[n]}$ can be now defined according to Section III-C.

Let $N_{\mathrm{av}}[n] \coloneqq N[n]$, $\mathbf{s}^{\mathrm{av}}_i[n] \coloneqq \mathbf{s}'_i[n]$, $\forall i \in \{1, \ldots, N[n]\}$, and for some $\sigma \in \mathbb{R}_+$,

$$\begin{aligned}
\boldsymbol{\Psi}[n] &\coloneqq [\psi_1[n], \ldots, \psi_{N[n]}[n]] \\
&\coloneqq \boldsymbol{\Phi}_{\mathcal{T}^{(n)}_{N[n]}} (\mathbf{K}_{\mathcal{T}^{(n)}_{N[n]}} + \sigma \mathbf{I}_{N[n]})^{-1}, \qquad (25)\\
\boldsymbol{\Phi}_{\mathcal{T}^{(n)}_{N[n]}} &\coloneqq [\varphi(\mathbf{s}_1[n], a_1[n]), \ldots, \varphi(\mathbf{s}_{N[n]}[n], a_{N[n]}[n])], \\
\boldsymbol{K}_{\mathcal{T}^{(n)}_{N[n]}} &\coloneqq \boldsymbol{\Phi}^{\intercal}_{\mathcal{T}^{(n)}_{N[n]}} \boldsymbol{\Phi}_{\mathcal{T}^{(n)}_{N[n]}}, \\
\boldsymbol{\Phi}^{\mathrm{av}}_{\mu_n}[n] &\coloneqq [\varphi(\mathbf{s}'_1[n], \mu_n(\mathbf{s}'_1[n])), \\
&\qquad \ldots, \varphi(\mathbf{s}'_{N[n]}[n], \mu_n(\mathbf{s}'_{N[n]}[n]))],
\end{aligned}$$

where $\{\mu_n(\mathbf{s}'_i[n])\}_{i=1}^{N[n]}$ are computed by (24). As such, (3a) takes the following form:

$$T^{(n)}_{\mu_n}(Q) \coloneqq g + \alpha \boldsymbol{\Psi}[n] \boldsymbol{\Phi}^{\mathrm{av}\intercal}_{\mu_n}[n] Q, \quad \forall Q \in \mathcal{H}. \qquad (26)$$

Computing $T^{(n)}_{\mu_n}(Q)$ for a given $Q \in \mathcal{H}$ amounts to identifying $(T^{(n)}_{\mu_n}(Q))(\mathbf{s}, a)$ for *all* $\mathbf{z} = (\mathbf{s}, a) \in \mathfrak{S} \times \mathfrak{A}$, which is a computationally infeasible task given that $\mathfrak{S}$ is the *continuous* $\mathbb{R}^4$. To surmount this obstacle, this study uses the point evaluation of $T^{(n)}_{\mu_n}(Q)$ at a *single* state-action vector $\mathbf{z}_{\nu_*} = (\mathbf{s}_{\nu_*}, a_{\nu_*})$, chosen by the user from the history of state-action pairs $\{\mathbf{z}_\nu = (\mathbf{s}_\nu, a_\nu)\}_{\nu=0}^{n-1}$, that is, $\nu_* \in \{0, \ldots, n-1\}$, to define the following superset of $\mathrm{Fix}\, T^{(n)}_{\mu_n}$:

$$\begin{aligned}
&H^{(n)}_{\mu_n}[\mathbf{z}_{\nu_*}] \\
&\coloneqq \{Q \in \mathcal{H} \,|\, (T^{(n)}_{\mu_n}(Q) - Q)(\mathbf{z}_{\nu_*}) = 0\} \\
&= \{Q \in \mathcal{H} \,|\, \langle T^{(n)}_{\mu_n}(Q) - Q \mid \varphi(\mathbf{z}_{\nu_*})\rangle_{\mathcal{H}} = 0\} \qquad (27a)\\
&= \{Q \in \mathcal{H} \,|\, \langle Q \mid h^{(n)}_{\mu_n}[\mathbf{z}_{\nu_*}]\rangle_{\mathcal{H}} = g(\mathbf{z}_{\nu_*})\}, \qquad (27b)
\end{aligned}$$

where (27a) follows by the reproducing property, and (27b) by incorporating (26) with

$$\begin{aligned}
&h^{(n)}_{\mu_n}[\mathbf{z}_{\nu_*}] \\
&\coloneqq \varphi(\mathbf{z}_{\nu_*}) - \alpha \sum\nolimits_{i=1}^{N[n]} (\psi_i[n])(\mathbf{z}_{\nu_*})\, \varphi(\mathbf{s}'_i[n], \mu_n(\mathbf{s}'_i[n])) \qquad (28)
\end{aligned}$$

in (27a). Notice that $H^{(n)}_{\mu_n}[\mathbf{z}_{\nu_*}]$ is a hyperplane of $\mathcal{H}$, with $h^{(n)}_{\mu_n}[\mathbf{z}_{\nu_*}]$ being its normal vector, $\varphi(\mathbf{z}_{\nu_*}) = \kappa(\mathbf{z}_{\nu_*}, \cdot)$, and $(\psi_i[n])(\mathbf{z}_{\nu_*})$ stands for the value of $\psi_i[n]$ at $\mathbf{z}_{\nu_*}$. Hyperplane $H^{(n)}_{\mu_n}[\mathbf{z}_{\nu_*}]$ is well defined and non-empty even if $\mathrm{Fix}\, T^{(n)}_{\mu_n} = \emptyset$. Recall here that the Banach-Picard fixed-point theorem [56] guarantees that $\mathrm{Fix}\, T^{(n)}_{\mu_n}$ is non-empty and a singleton in the case where $T^{(n)}_{\mu_n}$ is a contraction.

It is worth stressing here that exact knowledge of the one-step loss g, as in $g(\mathbf{z})$ for *all* $\mathbf{z} \in \mathfrak{S} \times \mathfrak{A}$, is no longer necessary since the definition of $H^{(n)}_{\mu_n}[\mathbf{z}_{\nu_*}]$ requires only a point evaluation at $\mathbf{z}_{\nu_*}$, which, for the present setting, is set to be

$$g(\mathbf{z}_{\nu_*}) = g(\mathbf{s}_{\nu_*}, a_{\nu_*}) \coloneqq s'^{(2)}_{\nu_*} = s^{(2)}_{\nu_*+1}, \qquad (29)$$

where the rightmost equality in (29) follows by (23). This loss is inspired by the classical a-priori error in AdaFilt [1, (10.11)], with the $\log$ used to reduce as much as possible the dynamic range of the negative effect of the outliers on y_n.

Define also the quadratic loss $\mathcal{L}^{(n)}_{\mu_n}[\mathbf{z}_{\nu_*}](\cdot) \colon \mathcal{H} \to \mathbb{R}_+$ as

$$\begin{aligned}
\mathcal{L}^{(n)}_{\mu_n}[\mathbf{z}_{\nu_*}](Q) &\coloneqq \tfrac{1}{2} \langle T^{(n)}_{\mu_n}(Q) - Q \mid \varphi(\mathbf{z}_{\nu_*})\rangle^2_{\mathcal{H}} \\
&= \tfrac{1}{2} \left[\langle Q \mid h^{(n)}_{\mu_n}[\mathbf{z}_{\nu_*}]\rangle_{\mathcal{H}} - g(\mathbf{z}_{\nu_*})\right]^2. \qquad (30)
\end{aligned}$$

It can be verified by (27) that

$$H^{(n)}_{\mu_n}[\mathbf{z}_{\nu_*}] = \arg\min\nolimits_{Q \in \mathcal{H}} \mathcal{L}^{(n)}_{\mu_n}[\mathbf{z}_{\nu_*}](Q).$$

C. Trajectory samples and policy evaluation

This subsection details the way the trajectory samples $\mathcal{T}^{(n)}_{N[n]} \coloneqq \{(\mathbf{s}_i[n], a_i[n], \mathbf{s}'_i[n])\}_{i=1}^{N[n]}$ are constructed. Instrumental to the construction is the buffer $\mathfrak{B}_n \coloneqq \{\mathfrak{b}_j \coloneqq (\bar{\mathbf{s}}_j, \bar{a}_j, g(\bar{\mathbf{s}}_j, \bar{a}_j), \bar{\mathbf{s}}'_j)\}_{j=1}^{|\mathfrak{B}_n|}$, where $(\bar{\mathbf{s}}_j, \bar{\mathbf{s}}'_j) \in \mathfrak{S}^2$, $\bar{a}_j \in \mathfrak{A}$, and where $\bar{\mathbf{s}}'_j$ is determined by $(\bar{\mathbf{s}}_j, \bar{a}_j)$. The way to update $\mathfrak{B}_n$ is provided first, while the design of trajectory samples follows next by utilizing the strategy of experience replay [27].

1) Updating $\mathfrak{B}_{n-1}$ to $\mathfrak{B}_n$: At time n, buffer $\mathfrak{B}_{n-1}$ and tuple $(\mathbf{s}_{n-1}, a_{n-1}, g(\mathbf{s}_{n-1}, a_{n-1}), \mathbf{s}_n)$ are available to the user. Given a user-defined distance function $\mathrm{dist}_{\mathfrak{S}}(\cdot, \cdot) \colon \mathfrak{S} \times \mathfrak{S} \to \mathbb{R}_+$, for example, $\mathrm{dist}_{\mathfrak{S}}(\cdot, \cdot) \coloneqq 1 - \kappa_{\mathrm{G}}(\cdot, \cdot)$, where $\kappa_{\mathrm{G}}(\cdot, \cdot)$ is a Gaussian kernel, and a threshold $\delta_{\mathfrak{S}} \in \mathbb{R}_{++}$, consider the following criterion:

$$\begin{aligned}
&\mathrm{dist}_{\mathfrak{S}}(\mathbf{s}_{n-1}, \bar{\mathbf{s}}_j) > \delta_{\mathfrak{S}}, \\
&\qquad \forall \mathfrak{b}_j = (\bar{\mathbf{s}}_j, \bar{a}_j, g(\bar{\mathbf{s}}_j, \bar{a}_j), \bar{\mathbf{s}}'_j) \in \mathfrak{B}_{n-1}. \qquad (31)
\end{aligned}$$

If (31) is satisfied, $\mathbf{s}_{n-1}$ is considered to be "different" enough from *all* states $\bar{\mathbf{s}}_j$ which appear in the tuples $\mathfrak{b}_j$ of $\mathfrak{B}_{n-1}$, and to carry "sufficiently novel" information to be included in $\mathfrak{B}_n$. Consequently, generate all tuples $\mathfrak{C}_{n-1} \coloneqq \{(\mathbf{s}_{n-1}, a, g(\mathbf{s}_{n-1}, a), \mathbf{s}'_{n-1}(\mathbf{s}_{n-1}, a, \mathfrak{D}_{(n-M_{\mathrm{av}}):n}, \boldsymbol{\theta}_n, \boldsymbol{\theta}_{n-1}) \,|\, a \in \mathfrak{A}\}$, where $\mathbf{s}'_{n-1}(\mathbf{s}_{n-1}, a, \mathfrak{D}_{(n-M_{\mathrm{av}}):n}, \boldsymbol{\theta}_n, \boldsymbol{\theta}_{n-1})$ is obtained by replacing a_{n-1} with a in (23). Define then

$$\mathfrak{B}_n \coloneqq \mathfrak{B}_{n-1} \cup \{(\mathbf{s}_{n-1}, a_{n-1}, g(\mathbf{s}_{n-1}, a_{n-1}), \mathbf{s}_n)\} \cup \mathfrak{C}_{n-1}.$$

On the other hand, if (31) is not satisfied, then $\mathfrak{B}_n \coloneqq \mathfrak{B}_{n-1}$.

2) Experience replay: Now that buffer $\mathfrak{B}_n$ has been updated and given a user-defined distance function $\mathrm{dist}_{\mathfrak{Z}}(\cdot, \cdot) \colon \mathfrak{Z} \times \mathfrak{Z} \to \mathbb{R}_+$, for example, $\mathrm{dist}_{\mathfrak{Z}}(\cdot, \cdot) \coloneqq 1 - \kappa_{\mathrm{G}}(\cdot, \cdot)$, where $\kappa_{\mathrm{G}}(\cdot, \cdot)$ is a Gaussian kernel, and a threshold $\delta_{\mathfrak{Z}} \in \mathbb{R}_{++}$, define

$$\begin{aligned}
\mathfrak{T}_{\mathbf{z}} \coloneqq \{&\mathfrak{b}_j = (\bar{\mathbf{s}}_j, \bar{a}_j, g(\bar{\mathbf{s}}_j, \bar{a}_j), \bar{\mathbf{s}}'_j) \in \mathfrak{B}_n \\
&\mid \mathrm{dist}_{\mathfrak{Z}}(\mathbf{z}, (\bar{\mathbf{s}}_j, \bar{a}_j)) \le \delta_{\mathfrak{Z}}\}, \quad \forall \mathbf{z} \in \mathfrak{Z}. \qquad (32)
\end{aligned}$$

In other words, $\mathfrak{T}_{\mathbf{z}}$ includes all tuples $\mathfrak{b}_j$ of $\mathfrak{B}_n$ whose state-action pairs $(\bar{\mathbf{s}}_j, \bar{a}_j)$ are "sufficiently similar" with $\mathbf{z}$. Identify then $\mathfrak{T}_{\mathbf{z}_{n-1}}$ by (32), and define the trajectory samples

$$\begin{aligned}
\mathcal{T}^{(n)}_{N[n]} &= \{(\mathbf{s}_i[n], a_i[n], \mathbf{s}'_i[n])\}_{i=1}^{N[n]} \\
&\coloneqq \{(\bar{\mathbf{s}}_j, \bar{a}_j, \bar{\mathbf{s}}'_j) \,|\, \mathfrak{b}_j = (\bar{\mathbf{s}}_j, \bar{a}_j, g(\bar{\mathbf{s}}_j, \bar{a}_j), \bar{\mathbf{s}}'_j) \in \mathfrak{T}_{\mathbf{z}_{n-1}}\} \\
&\qquad \cup \{(\mathbf{s}_{n-1}, a_{n-1}, \mathbf{s}'_{n-1})\},
\end{aligned}$$

where $\mathbf{s}'_{n-1}$ is defined by (23). Consider also $T_{\mu_n}^{(n)}$ and loss $\mathcal{L}_{\mu_n}^{(n)}[\mathbf{z}_{n-1}](\cdot)$ as in (26) and (30), respectively, and apply the SGD rule, with a learning rate $\eta \in \mathbb{R}_{++}$, to form the update:

$$\begin{aligned} Q_{n+1/2} &:= Q_n - \eta \nabla \mathcal{L}_{\mu_n}^{(n)}[\mathbf{z}_{n-1}](Q_n) \\ &= Q_n - \eta \Big[\langle Q_n \mid h_{\mu_n}^{(n)}[\mathbf{z}_{n-1}] \rangle_{\mathcal{H}} \\ &\qquad - g(\mathbf{z}_{n-1}) \Big] \cdot h_{\mu_n}^{(n)}[\mathbf{z}_{n-1}] \,, \end{aligned} \tag{33}$$

where $g(\mathbf{z}_{n-1})$ is computed by (29), with $\mathbf{z}_{n-1}$ taking the place of $\mathbf{z}_{\nu_*}$.

It is worth mentioning here that in the case where the set $\mathcal{T}_{N[n]}^{(n)}$ of trajectory samples is a singleton, more specifically, $\mathcal{T}_{N[n]}^{(n)} = \{(\mathbf{s}_{n-1}, a_{n-1}, \mathbf{s}'_{n-1})\}$, and whenever $\sigma = 0$ in (25), then (33) corresponds to [38].

To exploit also state-action pairs other than $\mathbf{z}_{n-1}$, the strategy of experience replay is adopted to choose a $\mathfrak{b}_{j_*} \in \mathfrak{B}_n \setminus \{(\bar{\mathbf{s}}_j, \bar{a}_j, g(\bar{\mathbf{s}}_j, \bar{a}_j), \bar{\mathbf{s}}'_j) \in \mathfrak{B}_n \,|\, (\bar{\mathbf{s}}_j, \bar{a}_j) = \mathbf{z}_{n-1}\}$ via a probability distribution, whose details are skipped here but can be found in [27]. Having identified such a $\mathfrak{b}_{j_*} = (\bar{\mathbf{s}}_{j_*}, \bar{a}_{j_*}, g(\bar{\mathbf{s}}_{j_*}, \bar{a}_{j_*}), \bar{\mathbf{s}}'_{j_*})$, let $\bar{\mathbf{z}}_{j_*} := (\bar{\mathbf{s}}_{j_*}, \bar{a}_{j_*})$ and define $\mathfrak{T}_{\bar{\mathbf{z}}_{j_*}}$ via (32). Next, let the trajectory samples

$$\begin{aligned} &\mathcal{T}_{N[n+1/2]}^{(n+1/2)} \\ &= \{(\mathbf{s}_i[n+1/2], a_i[n+1/2], \mathbf{s}'_i[n+1/2])\}_{i=1}^{N[n+1/2]} \\ &:= \{(\bar{\mathbf{s}}_j, \bar{a}_j, \bar{\mathbf{s}}'_j) \,|\, \mathfrak{b}_j = (\bar{\mathbf{s}}_j, \bar{a}_j, g(\bar{\mathbf{s}}_j, \bar{a}_j), \bar{\mathbf{s}}'_j) \in \mathfrak{T}_{\bar{\mathbf{z}}_{j_*}}\} \,. \end{aligned}$$

Define also $T_{\mu_n}^{(n+1/2)}$ and loss $\mathcal{L}_{\mu_n}^{(n+1/2)}[\bar{\mathbf{z}}_{j_*}](\cdot)$ as in (26) and (30), respectively, and apply again the SGD rule to obtain the Q-function estimate of Section III in Algorithm 1:

$$\begin{aligned} Q_{n+1} &:= Q_{n+1/2} - \eta \nabla \mathcal{L}_{\mu_n}^{(n+1/2)}[\bar{\mathbf{z}}_{j_*}](Q_{n+1/2}) \\ &= Q_{n+1/2} - \eta \Big[\langle Q_{n+1/2} \mid h_{\mu_n}^{(n+1/2)}[\bar{\mathbf{z}}_{j_*}] \rangle_{\mathcal{H}} \\ &\qquad - g(\bar{\mathbf{z}}_{j_*}) \Big] \cdot h_{\mu_n}^{(n+1/2)}[\bar{\mathbf{z}}_{j_*}] \,. \end{aligned} \tag{34}$$

D. Dimensionality reduction by random Fourier features

At every time instance n, Algorithm 1 adds new features into the representation of Q_{n+1} via (28), (33), and (34), justifying the "nonparametric" characterization of the proposed design. These new features contribute information in Q_{n+1} along novel dimensions of $\mathcal{H}$ which may have not been explored prior to time n. Novel dimensions are welcome since they lead into a "rich" kernel-based representation of the Q-function. However, due to the potentially infinite dimensionality of $\mathcal{H}$, the length of the Q-function representation may grow unbounded as new dimensions/features are added up, raising in turn hardware/computational obstacles due to the need for large storage space and large number of computations to process the long Q-function representations ("curse of dimensionality"). The desire for low hardware/computational footprints calls for dimensionality reduction, which is achieved here by employing random Fourier features (RFF) [60] as follows.

According to Bochner's theorem, there exist pairs $(\kappa, \mathbf{v})$, where κ is a real-valued reproducing kernel and $\mathbf{v}$ an RV, s.t. $\kappa(\mathbf{z}, \mathbf{z}') = \mathbb{E}_{\mathbf{v}}\{\cos[\mathbf{v}^\intercal(\mathbf{z} - \mathbf{z}')]\}$, $\forall \mathbf{z}, \mathbf{z}' \in \mathbb{R}^D$ [60]. An example of such a pair is $(\kappa_{\mathrm{G}}, \mathbf{v}_{\mathrm{G}})$, where κ_{G} is the Gaussian kernel and $\mathbf{v}_{\mathrm{G}}$ follows the Gaussian distribution $\mathcal{N}(0, \mathbf{I}_D)$. It can be verified that $\mathbb{E}_{\mathbf{v},u}\{\cos[\mathbf{v}^\intercal(\mathbf{z}+\mathbf{z}') + 2u]\} = 0$, where u is an RV uniformly distributed over $[0, 2\pi)$ and independent of $\mathbf{v}$. Hence, Bochner's theorem yields:

$$\begin{aligned} \kappa(\mathbf{z}, \mathbf{z}') &= \mathbb{E}_{\mathbf{v}}\{\cos[\mathbf{v}^\intercal(\mathbf{z} - \mathbf{z}')]\} \\ &= \mathbb{E}_{\mathbf{v},u}\{\cos[(\mathbf{v}^\intercal\mathbf{z} + u) - (\mathbf{v}^\intercal\mathbf{z}' + u)] \\ &\quad + \cos[(\mathbf{v}^\intercal\mathbf{z} + u) + (\mathbf{v}^\intercal\mathbf{z}' + u)]\} \\ &= 2\,\mathbb{E}_{\mathbf{v},u}\{\cos(\mathbf{v}^\intercal\mathbf{z} + u) \cdot \cos(\mathbf{v}^\intercal\mathbf{z}' + u)\} \\ &\approx 2\tfrac{1}{D_{\mathrm{RFF}}} \sum_{i=1}^{D_{\mathrm{RFF}}} \cos(\mathbf{v}_i^\intercal\mathbf{z} + u_i) \cdot \cos(\mathbf{v}_i^\intercal\mathbf{z}' + u_i) \\ &= \varphi_{\mathrm{RFF}}^\intercal(\mathbf{z})\, \varphi_{\mathrm{RFF}}(\mathbf{z}') \,, \end{aligned} \tag{35}$$

where $\approx$ in (35) holds true by the law of large numbers [61] for a large user-defined number D_{RFF} of IID copies $\{\mathbf{v}_i\}_{i=1}^{D_{\mathrm{RFF}}}$ and $\{u_i\}_{i=1}^{D_{\mathrm{RFF}}}$ of $\mathbf{v}$ and u, respectively, and φ_{RFF} is the feature mapping defined as

$$\begin{aligned} \varphi_{\mathrm{RFF}} &\colon \mathbb{R}^D \to \mathbb{R}^{D_{\mathrm{RFF}}} \\ &\colon \mathbf{z} \mapsto \sqrt{\tfrac{2}{D_{\mathrm{RFF}}}} [\cos(\mathbf{v}_1^\intercal\mathbf{z} + u_1), \ldots, \cos(\mathbf{v}_{D_{\mathrm{RFF}}}^\intercal\mathbf{z} + u_{D_{\mathrm{RFF}}})]^\intercal \,. \end{aligned} \tag{36}$$

Mapping (36) serves as a low dimensional rendition of the feature mapping $\varphi\colon \mathbb{R}^D \to \mathcal{H}$ introduced in Section II-A, since D_{RFF} can be made smaller than the typically large and potentially infinite $\dim \mathcal{H}$.

Samples $\{\mathbf{v}_i\}_{i=1}^{D_{\mathrm{RFF}}}, \{u_i\}_{i=1}^{D_{\mathrm{RFF}}}$ are taken in advance from the Gaussian and uniform PDFs, respectively, and are used via (35) and (36) to bound the computational complexity of Algorithm 1. The results of Section V are based on these approximations. The approximation accuracy of RFF depends on D_{RFF}: the larger the D_{RFF}, the better the approximation in (35). An analysis on D_{RFF} and its connections with the performance of Algorithm 1, as well as with the choice of the approximating RKHS, is deferred to a future study.

E. Computational complexity

First, the complexity to compute $s_{n-1}^{\prime(2)}$ in (23c) is of order $\mathcal{O}(LM_{\mathrm{av}})$, because $s_{n-1}^{\prime(2)}$ averages over M_{av} number of samples, and for each sample $(\mathbf{x}_{n-m}, y_{n-m})$, it takes $\mathcal{O}(L)$ to compute $|y_{n-m} - \boldsymbol{\theta}_n^\intercal(a_{n-1})\mathbf{x}_{n-m}|$. Second, it takes $\mathcal{O}(\dim(\mathfrak{S}))$ and $\mathcal{O}(\dim(\mathfrak{S}) + 1)$ to compute $\mathrm{dist}_{\mathfrak{S}}(\mathbf{s}_{n-1}, \bar{\mathbf{s}}_j)$ and $\mathrm{dist}_{\mathfrak{Z}}(\mathbf{z}, (\bar{\mathbf{s}}_j, \bar{a}_j))$ to verify the criteria in (31) and (32), respectively. Since all elements of the buffers need to be examined, the previous complexities become in total $\mathcal{O}(|\mathfrak{B}_{n-1}| \dim(\mathfrak{S}))$ and $\mathcal{O}(|\mathfrak{B}_n|(\dim(\mathfrak{S}) + 1))$. Third, to compute $\boldsymbol{\Psi}[n]$ in (25), $\mathcal{O}(N^3[n])$ operations are needed for the computation of the $N[n] \times N[n]$ inverse matrix. Moreover, with regards to (24), the computation of $Q_n(\mathbf{s}'_i[n], a)$ costs $\mathcal{O}(D_{\mathrm{RFF}})$, and because this operation is performed over $\mathfrak{A}$, and thus it is run for $|\mathfrak{A}|$ times, a total number of $\mathcal{O}(D_{\mathrm{RFF}}|\mathfrak{A}|)$ operations is required to compute $\mu_n(\mathbf{s}'_i[n])$ in (25). Hence, it takes $\mathcal{O}(N^3[n] + N[n]D_{\mathrm{RFF}}|\mathfrak{A}|)$ operations to compute $h_{\mu_n}^{(n)}[\mathbf{z}_{\nu_*}]$ in (28).

To summarize, if $\mathcal{C}_0 := \mathcal{O}(LM_{\mathrm{av}} + |\mathfrak{B}_{n-1}| \dim(\mathfrak{S}))$, then the total number of operations for the proposed method is

$\mathcal{C}_{\text{prop}} = \mathcal{C}_0 + \mathcal{O}(|\mathfrak{B}_n|(\dim(\mathfrak{S})+1) + N^3[n] + N[n]D_{\text{RFF}}|\mathfrak{A}|)$, whereas the RFF variation of [41] scores a complexity $\mathcal{C}_{\text{TD(0)}} = \mathcal{C}_0 + \mathcal{O}(D_{\text{RFF}}|\mathfrak{A}|)$, and [26] demonstrates complexity $\mathcal{C}_{\text{KLSPI}} = \mathcal{C}_0 + \mathcal{O}(N_{\text{dic}}^2 + (\dim(\mathfrak{S})+1)N_{\text{dic}}|\mathfrak{A}|)$, with N_{dic} being the size of a dictionary whose construction is inherent in [26]. Typical values of $N[n]$ for the proposed method are shown in the caption of Figure 4.

IV. Performance Analysis of Algorithm 1

In the following discussion, a statement $(\dots(n))$, which depends on the iteration index $n \in \mathbb{N}$, will be said to hold true "for all sufficiently large n," if there exists a large $n_0 \in \mathbb{N}_*$ s.t. $(\dots(n))$ holds true $\forall n \geq n_0$. Moreover, within the context of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ (see the discussion before Assumptions 3), a statement $(\dots)$ will be said to hold true with high probability (*w.h.p.*), if $(\dots)$ holds true on an event $\mathcal{E} \in \mathcal{F}$ with $\mathbb{P}(\mathcal{E}) \geq 1-\varepsilon$, for a sufficiently small $\varepsilon \in \mathbb{R}_{++}$.

Central to the following discussion are the sequence of estimates $(Q_n)_{n\in\mathbb{N}}$ generated by Algorithm 1, the classical B-Maps $T^\diamond, T^\diamond_{\mu_n}$ in (2), as well as the newly proposed $T^{(n)}_{\mu_n}$ one in (3) for a stationary policy $\mu_n(\cdot)\colon \mathfrak{S} \to \mathfrak{A}$. Recall also that $Q^\diamond_*$ stands for a fixed point of mapping $T^\diamond$, *i.e.*, $Q^\diamond_* \in \text{Fix}(T^\diamond) \Leftrightarrow Q^\diamond_* = T^\diamond(Q^\diamond_*)$, that $Q^\diamond_{\mu_n} \in \text{Fix}(T^\diamond_{\mu_n})$ and $Q_{\mu_n} \in \text{Fix}(T^{(n)}_{\mu_n})$.

Theorem 4 asserts that for an arbitrarily fixed $\epsilon \in \mathbb{R}_{++}$ and for any n, $\|Q^\diamond_{\mu_n} - Q_{\mu_n}(N[n])\|_{\mathcal{H}} \leq \epsilon$ holds true *w.h.p.* for all sufficiently large $N[n]$. By this result, it is expected that for sufficiently large $N[n]$, the size of the event

$$E^{(\epsilon)}_{n,N[n]} \coloneqq \left\{\omega \in \Omega \,\middle|\, \|Q^\diamond_{\mu_n} - Q_{\mu_n}(N[n])\|_{\mathcal{H}} \leq \epsilon\right\}. \tag{37}$$

can be considered to be large. This observation serves as the motivation behind the following Assumption 5(i).

Assumptions 5.

- **(i)** Consider an $\epsilon \in \mathbb{R}_{++}$ s.t.
$$E^{(\epsilon)} \coloneqq \liminf_{n\to\infty} \liminf_{N[n]\to\infty} E^{(\epsilon)}_{n,N[n]} \neq \emptyset, \tag{38}$$
where for events $(\mathcal{E}_\nu)_{\nu\in\mathbb{N}}$, the event $\liminf_{\nu\to\infty}\mathcal{E}_\nu \coloneqq \cup_\nu \cap_{\nu'\geq\nu} \mathcal{E}_{\nu'}$ bears the meaning of "$\mathcal{E}_\nu$ eventually" [30, Def. 2.8].
- **(ii)** Presume Assumptions 3.
- **(iii)** In the current online setting, Assumption 3(vi) takes the following form. There exists $\beta_\infty \in (0,1)$ s.t. $\beta_n = \beta_n(N[n]) \leq \beta_\infty$, for all sufficiently large n and $N[n]$, a.s., where β_n is defined according to (19) as
$$\beta_n \coloneqq \alpha\left(\|\mathbf{K}_{\Psi_n}\|_2 \sup_{\mu'\in\mathcal{M}} \|\mathbf{K}^{\text{av},n}_{\mu'}\|_2\right)^{1/2}. \tag{39}$$
- **(iv)** There exists $\Delta_0 \in \mathbb{R}_{++}$ s.t. $\|Q_{\mu_n} - Q_n\|_{\mathcal{H}} \leq \Delta_0$, for all sufficiently large n, a.s.
- **(v)** There exists $\Delta_1 \in \mathbb{R}_{++}$ s.t. $\|T^\diamond_{\mu_{n+1}}(Q_n) - T^\diamond(Q_n)\|_{\mathcal{H}} \leq \Delta_1$, for all sufficiently large n, a.s.
- **(vi)** There exists $\Delta_2 \in \mathbb{R}_{++}$ s.t. $\|(T^\diamond_{\mu_{n+1}} - \text{Id})(Q^\diamond_{\mu_n})\|_{\mathcal{H}} \leq \Delta_2$, for all sufficiently large n, a.s.

Assumptions 5(iv) to 5(vi) are motivated by the presuppositions in [18, (5.11), (5.12)] which are used in the proof of [18, Prop. 5.1.4] on approximate policy iteration. However, the discussion in [18, Prop. 5.1.4] is performed in the Banach space of all essentially bounded functions, and thus the presumed bounds are manifested pointwisely, whereas the present discussion is performed in an RKHS, which justifies in turn the adoption of the Hilbertian norm in Assumptions 5(iv) to 5(vi).

The following theorem bounds the distance of the sequence of estimates $(Q_n)_n$ from the fixed point $Q^\diamond_*$.

Theorem 6. Under Assumptions 5, for every $\omega \in E^{(\epsilon)}$ of (38),

$$\limsup_{n\to\infty} \|Q_n - Q^\diamond_*\|_{\mathcal{H}} \leq \Delta_3,$$

where $\Delta_3 \coloneqq \epsilon + \Delta_0 + [2\beta_\infty(\Delta_0+\epsilon) + \Delta_1 + \Delta_2/(1-\beta_\infty)]/(1-\beta_\infty)$.

Proof: See Appendix D. ∎

If the size of $E^{(\epsilon)}$ is large, which is something anticipated by the discussion around (37), the assertion of Theorem 6 holds true *w.h.p.* Instead of the point-wise, sample-point-based analysis of Theorem 6, the following Theorem 8 provides a bound on the sequence $(Q_n)_n$ via the expectation operator $\mathbb{E}\{\cdot\}$. To this end, the following assumptions are necessary.

Assumptions 7.

- **(i)** To simplify the proofs, policy evaluation in Section III of Algorithm 1 is performed without considering experience replay (34), that is, $Q_{n+1} \coloneqq Q_n - \eta\nabla\mathcal{L}^{(n)}_{\mu_n}[\mathbf{z}_{n-1}](Q_n)$.
- **(ii)** **(Stationary policy)** There exists a policy $\mu\colon \mathfrak{S} \to \mathfrak{A}$ s.t. $\mu_n = \mu$ for all sufficiently large n.
- **(iii)** There exists $\beta_\infty \in (0,1)$ s.t. mapping $T^\diamond_{\mu_n}$ is a β_∞-contraction for all sufficiently large n, a.s.
- **(iv)** **(Independency)** The σ-algebra $\sigma(\{\mathbf{z}_n, \xi_{n+1}\})$, generated by the state-action pair $\mathbf{z}_n$ and ξ_{n+1} (54), is independent of the filtration $\mathcal{F}_n$ for all sufficiently large n, where $\mathcal{F}_n \coloneqq \sigma(\{Q_\nu\}^n_{\nu=0})$ is defined as the σ-algebra generated by the sequence of estimates $\{Q_\nu\}^n_{\nu=0}$ [30].
- **(v)** **(Stationary moment)** There exists $\mathfrak{m}^{(4)}_\xi \in \mathbb{R}_{++}$ s.t. $\mathfrak{m}^{(4)}_\xi = \mathbb{E}\{\|\xi_n\|^4_{\mathcal{H}}\}$, for all sufficiently large n, where ξ_n is defined by (54).
- **(vi)** **(Stationary covariance operators)** There exist bounded linear operators $\Sigma_{zz}, \Sigma_{\xi z}, \Sigma_{\xi\xi}\colon \mathcal{H} \to \mathcal{H}$ s.t. $\Sigma^{(n)}_{zz} = \Sigma_{zz}, \Sigma^{(n)}_{\xi z} = \Sigma_{\xi z}$, and $\Sigma^{(n)}_{\xi\xi} = \Sigma_{\xi\xi}$, for all sufficiently large n, where $\Sigma^{(n)}_{zz}, \Sigma^{(n)}_{\xi z}, \Sigma^{(n)}_{\xi\xi}$ are defined by (55).
- **(vii)** **(Positive definite $\mathcal{A}_{\mu_n}, \Sigma_{zz}$)** For all sufficiently large n, the linear bounded and self-adjoint mappings $\mathcal{A}_{\mu_n}$ (57a) and Σ_{zz} are positive definite, *i.e.*, their minimum spectral values $\sigma_{\min}(\mathcal{A}_{\mu_n}) > 0$ and $\sigma_{\min}(\Sigma_{zz}) > 0$, where $\sigma_{\min}(\cdot)$ is defined by (47a).
- **(viii)** **(Bounded kernel)** There exists $B_\kappa \in \mathbb{R}_{++}$ s.t. $\kappa(\mathbf{z},\mathbf{z}) \leq B_\kappa$, $\forall \mathbf{z} \in \mathfrak{Z}$.

Assumption 7(iv) is introduced to avoid more complicated proofs and lengthier discussions. Stationarity in Assumptions 7(v) and 7(vi) is often met in the stochastic analysis of online-learning algorithms, *e.g.*, [2, Examples 2.1 and 3.4]. Assumption 7(vii) is also met frequently in the literature on covariance matrices, *e.g.*, [1, Chap. 5] and Hessian operators of strongly convex loss/risk functions, *e.g.*, [2, §2.2]. Moreover, Assumption 7(viii) holds true for most of the commonly used kernels, *e.g.*, $B_\kappa = 1$ for the Gaussian and Laplacian ones.

Theorem 8. Under Assumptions 7, for any sufficiently small step size η, there exist $\Delta_4, \Delta_5 \in \mathbb{R}_{++}$ s.t.

$$\begin{aligned}&\limsup_{n\to\infty} \mathbb{E}\{\|Q_n - Q_*^\diamond\|_{\mathcal{H}}^2\}\\ &\leq \overbrace{\Delta_4\eta}^{\mathrm{T}_1} + \overbrace{\Delta_5 \limsup_{n\to\infty} \mathbb{E}\{\|\Sigma_{s'|z}^{\mu_n} - \hat{\Sigma}_{s'|z}^{\mu_n}(N[n])\|^2\}}^{\mathrm{T}_2}\\ &\quad + \underbrace{2\limsup_{n\to\infty} \mathbb{E}\{\|Q_{\mu_n}^\diamond - Q_*^\diamond\|_{\mathcal{H}}^2\}}_{\mathrm{T}_3},\end{aligned}$$

where operator $\Sigma_{s'|z}^{\mu_n}$ is defined by (46c) and $\hat{\Sigma}_{s'|z}^{\mu_n}(N[n])$ is defined by

$$\begin{aligned}\hat{\Sigma}_{s'|z}^{\mu_n}(N[n]) \coloneqq \tfrac{1}{\sqrt{N[n]}}\boldsymbol{\Phi}_{\mathcal{T}_{N[n]}}\big(\tfrac{1}{N[n]}\mathbf{K}_{\mathcal{T}_{N[n]}} + \sigma'_{N[n]}\mathbf{I}_{N[n]}\big)^{-1}\\ \cdot \tfrac{1}{\sqrt{N[n]}}\boldsymbol{\Phi}_\mu^{\mathsf{av}\intercal};\end{aligned}$$

see also (49), Proposition 1(iii), and Assumption 3(iii).

Proof: See Appendix E. ∎

Three terms $\mathrm{T}_1, \mathrm{T}_2, \mathrm{T}_3$ contribute to the bound in Theorem 8: T_1 which stems from the stochastic-gradient-descent recursion of Assumption 7(i), T_2 because of the error in approximating the expectation operator by trajectory sampling (Section III-C), and T_3 which quantifies the disagreement between μ_n and the "optimal" policy through the lenses of the classical B-Maps (2). It is worth mentioning here that under Assumptions 3 and according to Theorem 12 in Appendix C, $\|\Sigma_{s'|z}^{\mu_n} - \hat{\Sigma}_{s'|z}^{\mu_n}(N[n])\|$ can be made arbitrarily small *w.h.p.* for sufficiently large $N[n]$.

To establish bounds on the distance $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\|$ between the estimate $\boldsymbol{\theta}$ and the estimandum $\boldsymbol{\theta}_*$ (*cf.* Section I-A), standard arguments from the analysis of stochastic gradient descent (SGD) can be applied, because (22) is nothing but SGD on the convex p-power loss ($p \in [1, 2]$). For example, the discussion of [62, §8.2] can be followed by employing also the minimizer of the p-power loss as an auxiliary quantity to facilitate the analysis. Because of space limitations, rather than such a straightforward but tedious performance analysis, this study prefers to offer Theorem 10 instead, which showcases the connection between $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\|$ and the main mathematical object of this work, the Q-functions. To this end, Assumptions 9 are needed.

Assumptions 9.

- **(i)** Presume Assumption 7(ii).
- **(ii)** Presume Assumption 7(iii).
- **(iii)** Let $\alpha \in (0, 1)$.
- **(iv)** **(Bounded errors)** There exist $\Delta_6, \Delta_7 \in \mathbb{R}_{++}$ s.t.

$$\Delta_6 < |y_{n-m} - \boldsymbol{\theta}_n^\intercal \mathbf{x}_{n-m}| < \Delta_7\,,$$

  $\forall m \in \{0, \ldots, M_{\mathsf{av}} - 1\}$ and for all sufficiently large n, a.s.
- **(v)** **(Independency)** The input-signal $\mathbf{x}_n$ is independent of the outlier/noise o_n for all sufficiently large n. Moreover, the σ-algebra generated by $\{\mathbf{x}_m \mid m \in \{n - M_{\mathsf{av}} + 1, \ldots, n\}\}$ and $\{o_m \mid m \in \{n - M_{\mathsf{av}} + 1, \ldots, n\}\}$ is independent of the σ-algebra generated by $\boldsymbol{\theta}_n$ for all sufficiently large n.
- **(vi)** **(Stationary moment of o_n)** There exists $\sigma_o \in \mathbb{R}_{++}$ s.t. $\mathbb{E}\{o_n^2\} = \sigma_o^2$ and $\mathbb{E}\{o_n\} = 0$, for all sufficiently large n.
- **(vii)** **(Stationary covariance operator of $\mathbf{x}_n$)** There exists a positive definite matrix $\Sigma_{xx} \in \mathbb{R}^{L\times L}$, with $\sigma_{\min}(\Sigma_{xx}) = \lambda_{\min}(\Sigma_{xx}) > 0$ and where $\lambda_{\min}(\Sigma_{xx})$ stands for the minimum eigenvalue of Σ_{xx}, s.t. $\mathbb{E}\{\mathbf{x}_n\mathbf{x}_n^\intercal\} = \Sigma_{xx}$, for all sufficiently large n. Moreover, $\liminf_{n\to\infty} \mathbb{E}\{\log\|\mathbf{x}_n\|_2^2\} > -\infty$.
- **(viii)** For the sequence of states $(\mathbf{s}_n)_n$ in Algorithm 1, $\limsup_{n\to\infty} |\mathbb{E}\{Q_\mu^\diamond(\mathbf{s}_n, \mu(\mathbf{s}_n))\}| < +\infty$.

Bounds Δ_6, Δ_7 in Assumption 9(iv) are introduced to simplify proofs and avoid lengthier discussions on convergence of RVs *w.h.p.* Bound Δ_6 in Assumption 9(iv) is motivated by the well-known fact that whenever the distribution function of the RV $y_{n-m} - \boldsymbol{\theta}_n^\intercal \mathbf{x}_{n-m}$ is continuous, then $\mathbb{P}\{|y_{n-m} - \boldsymbol{\theta}_n^\intercal \mathbf{x}_{n-m}| = 0\} = 0$ [30, §3.10]. Independency between the input signals and outlier/noise in Assumption 9(v) is a standard argument in estimation theory, *e.g.*, [1, Chap. 5], while independency between $\{\mathbf{x}_m\}$, $\{o_m\}$ and $\boldsymbol{\theta}_n$ is introduced in Assumption 9(v) to avoid more complicated proofs. Stationarity in Assumptions 9(vi) and 9(vii) is a classical hypothesis in the stochastic analysis of online-learning algorithms [2], while boundedness of the sequences $\mathbb{E}\{\log\|\mathbf{x}_n\|_2^2\}$ and $\mathbb{E}\{Q_\mu^\diamond(\mathbf{s}_n, \mu(\mathbf{s}_n))\}$ in Assumptions 9(vii) and 9(viii) are weak presuppositions stated here explicitly for mathematical rigor and to avoid any unclear points in the proof of Theorem 10.

Theorem 10. Under Assumptions 9, there exist $\Delta_8 \in \mathbb{R}_{++}$ and $\Delta_9 \in \mathbb{R}_+$ s.t. for all sufficiently large n,

$$\begin{aligned}&\mathbb{E}\{\|\boldsymbol{\theta}_* - \boldsymbol{\theta}_n(\mu(\mathbf{s}_{n-1}))\|_2^2\}\\ &\leq \frac{1-\alpha}{\Delta_8\lambda_{\min}(\Sigma_{xx})}\,\mathbb{E}\{Q_\mu^\diamond(\mathbf{s}_n, \mu(\mathbf{s}_n))\}\\ &\quad + \frac{1}{\Delta_8\lambda_{\min}(\Sigma_{xx})}[\log\operatorname{trace}(\Sigma_{xx}) - \Delta_8\sigma_o^2 - \Delta_9]\,.\end{aligned}$$

Proof: See Appendix F. ∎

In the light of Theorem 10, (24) makes now sense in the context of AdaFilt, because, choosing a policy by the greedy rule $\arg\min_{\mu(\cdot)\in\mathcal{M}} \mathbb{E}\{Q_\mu^\diamond(\mathbf{s}_n, \mu(\mathbf{s}_n))\}$ pushes the upper bound of Theorem 10 to lower levels, and, thus, potentially forces the estimate $\boldsymbol{\theta}_n$ to approach $\boldsymbol{\theta}_*$.

V. Numerical Tests

This section follows the standard route in AdaFilt and validates Algorithm 1, as well as several state-of-the-art methods, on synthetic data [1, 8–16]. Heavy-tailed α-stable and sparse outliers are considered because they are widely used to model realistic problems in AdaFilt [5]. Heavy-tailed α-stable outliers are generated by setting their parameters $\alpha_{\text{stable}} = 1$, $\beta_{\text{stable}} = 0.5$, $\sigma_{\text{stable}} = 1$ [6]. "Sparse" outliers appear in 10% of the data, while Gaussian noise with SNR $= 30$dB is added to the rest 90% of the data. Sparse outliers are generated by the uniform distribution and take values from the interval $[-100, 100]$.

Algorithm 1 competes against

1) **Non-RL-based methods:**
   - **(i)** LMP (1), for values of $p \in \mathfrak{A} \coloneqq \{1, 1.25, 1.5, 1.75, 2\}$ which are kept fixed throughout all iterations;

(ii) [12], which uses a combination of adaptive filters with different forgetting factors but with the same p-power loss;

(iii) [9], where two LMP recursions (1), with different p, are combined to tackle outliers;

(iv) the variable-kernel-width and correntropy-based VKW-MCC [16];

2) **RL-based methods:**

(v) the kernel-based TD(0) [41], equipped with RFF (Section III-D);

(vi) the popular kernel (K)LSPI [26]; and

(vii) the predecessor [17] of this work which is based on RKHS arguments.

Action space is defined as $\mathfrak{A} := \{1, 1.25, 1.5, 1.75, 2\}$ for all employed RL methods, including Algorithm 1. Computing $\arg\min_{a\in\mathfrak{A}}$ in (24) is not an "innocent" task when the range of p is the continuous interval $[1, 2]$, and not only a finite grid on the interval. Clearly, the finer the granularity of the finite grid on $[1, 2]$, and thus the larger the cardinality of $\mathfrak{A}$, the larger the computational complexity required to identify $\arg\min_{a\in\mathfrak{A}}$ in (24). Not only Algorithm 1, but all tested RL methods other than [17] define state vectors by (23) and are equipped with experience replay (Section III-C2). On the other hand, [17] defines the state space as $\mathfrak{S} \subset \mathbb{R}^{2L+1}$ and utilizes no experience replay, but uses rollout instead [18]. Notice also according to the discussion which follows (33) that Algorithm 1 with only one trajectory sample, *i.e.*, $N[n] = 1$ per n, corresponds to [38]. Nevertheless, Algorithm 1 is equipped with RFF (Section III-D) which is not available in [38].

The performance metric is the normalized deviation of the estimate $\boldsymbol{\theta}_n$ from the estimandum $\boldsymbol{\theta}_*$; see the vertical axes in all figures. The classical Gaussian kernel [22] was employed to define $\mathcal{H}$, approximated by RFF with $D_{\text{RFF}} = 500$ as described in Section III-D. The dimension of $\mathbf{x}_n, \boldsymbol{\theta}_*$ is set to $L = 100$, where $\mathbf{x}_n$ and $\boldsymbol{\theta}_*$ are generated by the Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_L)$, with $(\mathbf{x}_n)_{n\in\mathbb{N}}$ designed to be IID. The learning rate in (22) is $\rho = 10^{-3}$. Moreover, $M_{\text{av}} = 300$ and $\varpi = 0.3$ in (23), while $\eta = 0.1$ in (33) and (34). In (31), $\delta_{\mathfrak{S}} = 1-0.99$.

To study the effect of size $N[n]$, controlled by $\delta_{\mathfrak{Z}}$ in (32), the following sets of parameters were tested: **(i)** $\delta_{\mathfrak{Z}} = 1-0.98$ in (32), which yields $N[n] \geq 1$, and $\sigma = 10^{-1}$ in (25); and **(ii)** $\delta_{\mathfrak{Z}} = 0$, which forces $N[n] = 1$, and $\sigma = 0$, corresponding thus to [38]. Typical values of $N[n] > 1$ for the proposed method are shown in the caption of Figure 4. Additionally, to study the effect of the long-term loss, $\alpha = 0$ was also tested. Note that $\alpha = 0$ suggests that Algorithm 1 uses no trajectory samples.

In Algorithm 1, policy improvement and evaluation are scheduled to be run at every iteration n. Nevertheless, to promote stability and allow the policy-iteration step reach a "steady state" between two consecutive invocations of the policy-improvement step, a less greedy approach is followed here and Section III of Algorithm 1 is not run at every n, but it is invoked periodically, every other $N_p = 500$ iterations, *i.e.*, at iterations $\{n = N_p k \,|\, k \in \mathbb{N}\}$. Between two consecutive policy-improvement steps, that is, during iterations $\{N_p k, \ldots, N_p(k+1) - 1\}$, the policy stays fixed to $\mu_{N_p k}(\cdot)$. The same strategy is also followed for KLSPI [26].

KLSPI [26] was originally designed to generate trajectory samples offline by using training data. However, since this work considers the online setting, where no training data are available and only test data are considered, and to ensure fairness among all competing methods, matrix $\mathbf{A}$ and vector $\mathbf{b}$, which appear in [26] and are learned from training data, are substituted by the following online versions $\mathbf{A}_n$ and $\mathbf{b}_n$: upon defining $\mathbf{k}_n := \boldsymbol{\Phi}_{\text{KLSPI}}^{(n)\intercal} \varphi(\mathbf{s}_n, \mu_n(\mathbf{s}_n))$, where $\boldsymbol{\Phi}_{\text{KLSPI}}^{(n)}$ is a time-varying basis of RKHS vectors, constructed online by an approximate-linear-dependency criterion [26], let

$$\mathbf{A}_n := \mathbf{A}_{n-1} + \mathbf{k}_{n-1}(\mathbf{k}_{n-1} - \alpha\mathbf{k}_n)^\intercal\,,$$
$$\mathbf{b}_n := \mathbf{b}_{n-1} + g(\mathbf{s}_n, \mu_n(\mathbf{s}_n))\,\mathbf{k}_n\,,$$

where $\mathbf{A}_1 := \mathbf{k}_0(\mathbf{k}_0 - \alpha\mathbf{k}_1)^\intercal$ and $\mathbf{b}_0 := g(\mathbf{s}_0, \mu_0(\mathbf{s}_0))\,\mathbf{k}_0$.

All curves in the subsequent figures are the uniformly averaged results of 100 independent tests. Moreover, all competing methods were finely tuned to show their best performance per setting of the environment. Finely tuned hyper-parameters include the learning rate for Q-function estimation in RL-based methods, as well as the learning rate, forgetting factor, *etc.*, in the AdaFilt methods. Only the most crucial parameters per method are stated in the captions of Figures 3 and 6, while an exhaustive description of all the finely tuned parameters per method is avoided to improve the flow of the manuscript.

A. Scenario 1

Figures 2 to 4 refer to the scenario where the statistics (PDF) of the outliers stay fixed throughout all iterations. Moreover, as it is customary in the AdaFilt literature, system $\boldsymbol{\theta}_*$ changes randomly at a specific time index (here at time #20 000) to test the tracking ability of all competing methods.

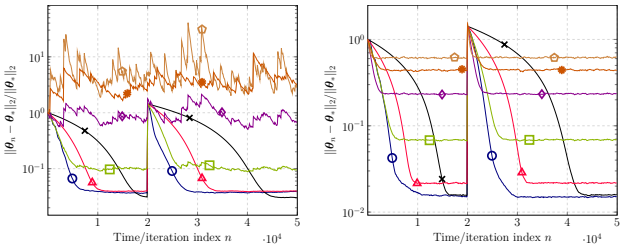
Figures 2 to 4 demonstrate that Algorithm 1 shows high estimation accuracy while tracking swiftly the estimandum $\boldsymbol{\theta}_*$. In Figure 2, Algorithm 1 reaches steady state faster than the "best" versions of LMP ($p = 1, 1.25$). An inspection of Figure 2 suggests that Algorithm 1 selects large values of p, within $\{1, 1.25, 1.5, 1.75, 2\}$, in the beginning state of learning to speed up convergence and then changes to select small values of p to score high accuracy in the steady state.

Figure 3 shows that Algorithm 1 outperforms the kernel-based TD(0) [41] and KLSPI [26]. The predecessor [17] scores an almost identical steady-state performance with Algorithm 1, but with a slower convergence speed. VKW-MCC [16] shows excellent performance under sparse outliers. However, under α-stable outliers, it is outperformed by Algorithm 1.

In Figure 4, where several parameters of Algorithm 1 are validated, $\alpha = 0.75$ shows the best estimation accuracy among $\alpha \in \{0, 0.75, 0.9\}$. Among them, $\alpha = 0$ does not perform well, which strongly suggests that the long-term loss is crucial in choosing p. The number of samples $N[n]$ does not seem to affect performance significantly in this specific AdaFilt application.

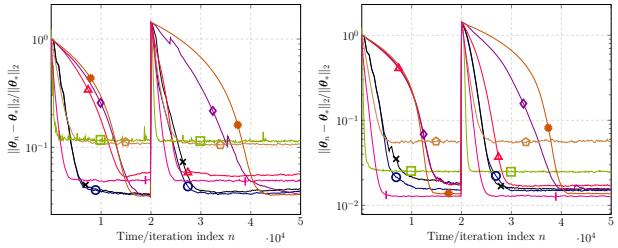
B. Scenario 2

Figures 5 to 7 refer to the scenario where the system $\boldsymbol{\theta}_*$ stays fixed but the statistics (PDF) of the outliers change



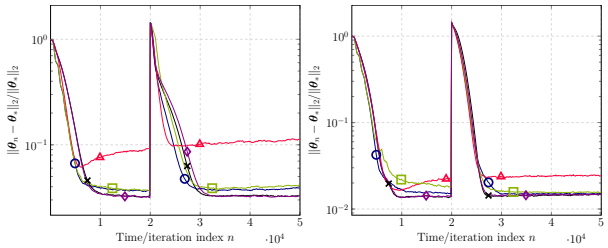
(a) α-stable outliers (b) Sparse outliers

Fig. 2. Scenario 1 (Section V-A): Algorithm 1 against LMP. ◯: Algorithm 1 with $\alpha = 0.9, N[n] \geq 1$. Marks $\times$, △, □, ◇, ⬠ correspond to (1) with $p = 1, 1.25, 1.5, 1.75, 2$, respectively. Mark ✳ denotes an algorithm which randomly chooses p, $\forall n$.



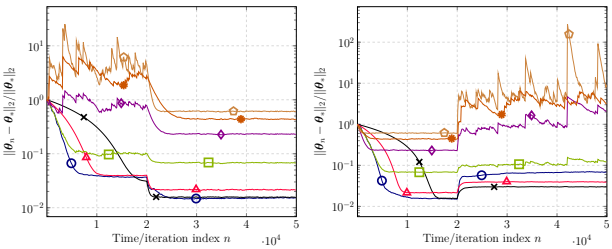
(a) α-stable outliers (b) Sparse outliers

Fig. 3. Scenario 1 (Section V-A): Algorithm 1 against non-RL- and RL-based methods. ◯: Algorithm 1 with $\alpha = 0.9, N[n] \geq 1$. $\times$: Algorithm 1 with $\alpha = 0.9, N[n] = 1$. △: Kernel-based TD(0) with $\alpha = 0.9$ [41]. □: [12] with $p = 1, \gamma_1 = 0.9, \gamma_2 = 0.99$. ◇: KLSPI with $\alpha = 0.9$ [26]. ⬠: mixed norm [9]. ✳: the predecessor [17] of the current work. |: VKW-MCC [16].



(a) α-stable outliers (b) Sparse outliers

Fig. 4. Scenario 1 (Section V-A): Versions of Algorithm 1 under several parameter settings. ◯: $\alpha = 0.9, N[n] \geq 1$. $\times$: $\alpha = 0.75, N[n] \geq 1$. △: $\alpha = 0$. □: $\alpha = 0.9, N[n] = 1$. ◇: $\alpha = 0.75, N[n] = 1$. In (b) and in the $\times$ case, the mean value of the time varying $N[n]$ ($\pm$ standard deviation) is 6.65 ($\pm$ 2.53).



(a) α-stable $\rightarrow$ sparse (b) Sparse $\rightarrow$ α-stable

Fig. 5. Scenario 2 (Section V-B): Algorithm 1 against LMP. ◯: Algorithm 1 with $\alpha = 0.9, N[n] \geq 1$. Marks $\times$, △, □, ◇, ⬠ correspond to (1) with $p = 1, 1.25, 1.5, 1.75, 2$, respectively. Mark ✳ denotes an algorithm which randomly chooses p, $\forall n$.

at a specific time instance (here, at iteration #20 000). Both sparse and α-stable outliers are considered in the following two dynamic sub-scenarios: **(i)** α-stable outliers appear at $n \in \{1, \ldots, 20\,000\}$, followed by sparse ones at $n \in \{20\,001, \ldots, 50\,000\}$; and **(ii)** sparse outliers contaminate signals whenever $n \in \{1, \ldots, 20\,000\}$, while α-stable ones appear at $n \in \{20\,001, \ldots, 50\,000\}$.

Figures 2(a) and 5(b) show different steady-state performances of Algorithm 1 after iteration #20 000, despite the fact that the statistics of the α-stable outliers are the same. More specifically, in Figure 2(a), the steady-state performance level of Algorithm 1 is almost identical to that of LMP for $p = 1.25$, whereas LMP with $p = 1.25$ scores a lower steady-state level than Algorithm 1 after iteration #20 000 in Figure 5(b). On the other hand, Algorithm 1 shows excellent performance in Figure 5(a) after the sudden transition to sparse outliers. The previous discussion concludes that Algorithm 1 appears to underperform in cases where there is a sudden change from light-tailed outliers to the heavy-tailed α-stable ones. In contrast, notice the excellent performance of Algorithm 1 under α-stable outliers in Figure 3(a). Moreover, Figure 6(b) demonstrates that kernel-based TD(0) [41] deteriorates significantly whenever the outlier PDF suddenly changes to α-stable outliers. The rest of the RL-based methods exhibit more or less robust performance against the abrupt change to α-stable outliers in Figure 6(b).

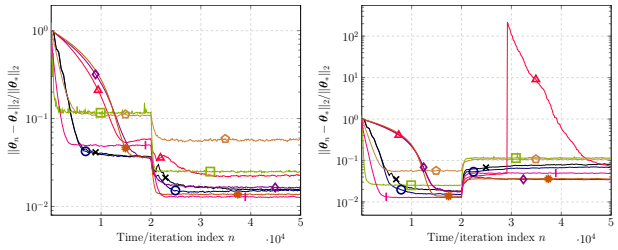
Finally, notice that under the heavy-tailed α-stable outliers, versions of Algorithm 1 with $\alpha > 0$ perform better than version with $\alpha = 0$ in Figure 4(a), whereas versions $\alpha > 0$ and $\alpha = 0$ perform similarly after iteration #20 000 in Figure 7(b).

VI. Conclusions

This paper designed novel nonparametric Bellman mappings (B-Maps) in reproducing kernel Hilbert spaces (RKHSs) for reinforcement learning (RL). The new B-Maps exhibit several desirable features (see Section I-B), with ample degrees of freedom. To benefit from that freedom, a variational framework (Proposition 1) was provided to identify the free parameters of the B-Maps. As a side effect, it was demonstrated that several state-of-the-art designs become special cases of the proposed B-Maps. Other non-trivial designs of B-Maps are deferred to a future work. On the application front, the manuscript considered the problem of selecting online, per time instance, the "optimal" coefficient p in the least-mean-p-power method, with no prior information on the outlier statistics and no training data. The application of the proposed B-Maps to automatically choose "optimal" hyperparameters of correntropy-based adaptive-filtering (AdaFilt) algorithms is also currently under consideration. The proposed B-Maps are general enough to be applied also to domains other than AdaFilt. Such application domains, together with their RL designs, are currently under consideration and will be presented soon at other publication venues.

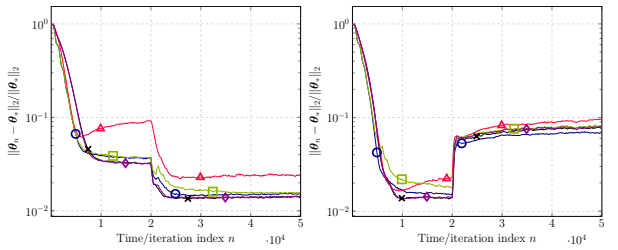
References

[1] A. H. Sayed, *Adaptive Filters*. Hoboken, New Jersey: John Wiley & Sons, 2008.



(a) α-stable $\rightarrow$ sparse (b) Sparse $\rightarrow$ α-stable

Fig. 6. Scenario 2 (Section V-B): Algorithm 1 against non-RL- and RL-based methods. ○: Algorithm 1 with $\alpha = 0.9, N[n] \geq 1$. ×: Algorithm 1 with $\alpha = 0.9, N[n] = 1$. △: Kernel-based TD(0) with $\alpha = 0.9$ [41]. □: [12] with $p = 1, \gamma_1 = 0.9, \gamma_2 = 0.99$. ◇: KLSPI with $\alpha = 0.9$ [26]. ⬠: mixed norm [9]. ✳: the predecessor [17] of the current work. |: VKW-MCC [16].



(a) α-stable $\rightarrow$ sparse (b) Sparse $\rightarrow$ α-stable

Fig. 7. Scenario 2 (Section V-B): Versions of Algorithm 1 under several parameter settings. ○: $\alpha = 0.9, N[n] \geq 1$. ×: $\alpha = 0.75, N[n] \geq 1$. △: $\alpha = 0$. □: $\alpha = 0.9, N[n] = 1$. ◇: $\alpha = 0.75, N[n] = 1$.

[2] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Foundations and Trends in Machine Learning*, vol. 7, no. 4-5, pp. 311–801, 2014.

[3] S. Theodoridis, *Machine Learning—A Bayesian and Optimization Perspective*, 2nd. Elsevier, 2020.

[4] P. J. Rousseeuw and A. Leroy, *Robust Regression and Outlier Detection*. Wiley, 1987.

[5] M. Shao and C. L. Nikias, "Signal processing with fractional lower order moments: Stable processes and their applications," *Proc. IEEE*, vol. 81, no. 7, pp. 986–1010, 1993.

[6] J. M. Miotto, *Pylevy*, https://github.com/josemiotto/pylevy, 2020.

[7] C. Gentile, "The robustness of the p-norm algorithms," *Machine Learning*, vol. 53, pp. 265–299, 2003.

[8] S.-C. Pei and C.-C. Tseng, "Least mean p-power error criterion for adaptive FIR filter," *IEEE Journal on Selected Areas in Communications*, vol. 12, no. 9, pp. 1540–1547, 1994.

[9] J. Chambers and A. Avlonitis, "A robust mixed-norm adaptive filter algorithm," *IEEE Signal Processing Letters*, vol. 4, no. 2, pp. 46–48, 1997.

[10] Y. Xiao, Y. Tadokoro, and K. Shida, "Adaptive algorithm based on least mean p-power error criterion for Fourier analysis in additive noise," *IEEE Trans. Signal Process.*, vol. 47, no. 4, pp. 1172–1181, 1999.

[11] E. E. Kuruoğlu, "Nonlinear least ℓ_p-norm filters for nonlinear autoregressive α-stable processes," *Digital Signal Processing*, vol. 12, no. 1, pp. 119–142, 2002.

[12] A. Navia-Vazquez and J. Arenas-Garcia, "Combination of recursive least p-norm algorithms for robust adaptive filtering in alpha-stable noise," *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1478–1482, 2012.

[13] B. Chen, L. Xing, Z. Wu, J. Liang, J. C. Príncipe, and N. Zheng, "Smoothed least mean p-power error criterion for adaptive filtering," *Digital Signal Processing*, vol. 40, no. C, pp. 154–163, May 2015.

[14] K. Slavakis and M. Yukawa, "Outlier-robust kernel hierarchical-optimization RLS on a budget with affine constraints," in *IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5335–5339.

[15] A. Singh and J. C. Príncipe, "Using correntropy as a cost function in linear adaptive filters," in *International Joint Conference on Neural Networks*, 2009, pp. 2950–2955.

[16] F. Huang, J. Zhang, and S. Zhang, "Adaptive filtering under a variable kernel width maximum correntropy criterion," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 64, no. 10, pp. 1247–1251, 2017.

[17] M. Vu, Y. Akiyama, and K. Slavakis, "Dynamic selection of p-norm in linear adaptive filtering via online kernel-based reinforcement learning," in *IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[18] D. Bertsekas, *Reinforcement Learning and Optimal Control*. Athena Scientific, 2019.

[19] D. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[20] R. E. Bellman, *Dynamic Programming*. Dover Publications, 2003.

[21] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, pp. 337–404, 1950.

[22] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (Adaptive computation and machine learning). MIT Press, 2002.

[23] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer, 2010.

[24] D. Ormoneit and Ś. Sen, "Kernel-based reinforcement learning," *Machine Learning*, vol. 49, pp. 161–178, 2002.

[25] D. Ormoneit and P. Glynn, "Kernel-based reinforcement learning in average-cost problems," *IEEE Transactions on Automatic Control*, vol. 47, no. 10, pp. 1624–1636, Oct. 2002.

[26] X. Xu, D. Hu, and X. Lu, "Kernel-based least squares policy iteration for reinforcement learning," *IEEE Transactions on Neural Networks*, vol. 18, no. 4, pp. 973–992, 2007.

[27] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," in *International Conference on Learning Representations*, 2016.

[28] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *AAAI Conference on Artificial Intelligence*, vol. 30, 2016.

[29] M. G. Bellemare, G. Ostrovski, A. Guez, P. Thomas, and R. Munos, "Increasing the action gap: New operators for reinforcement learning," *AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[30] D. Williams, *Probability with Martingales*. Cambridge University Press, 1991.

[31] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. New York: Springer, 2011.

[32] L. Song, A. Gretton, and C. Guestrin, "Nonparametric tree graphical models via kernel embeddings," in *AISTATS*, JMLR Workshop and Conference Proceedings, 2010, pp. 765–772.

[33] S. Grunewalder, G. Lever, L. Baldassarre, M. Pontil, and A. Gretton, "Modeling transition dynamics in MDPs with RKHS embeddings," in *Intern. Conf. on Machine Learning (ICML)*, 2012.

[34] A. Nedić and D. P. Bertsekas, "Least squares policy evaluation algorithms with linear function approximation," *Discrete Event Dynamic Systems*, vol. 13, no. 1, pp. 79–110, Jan. 2003.

[35] D. P. Bertsekas, V. S. Borkar, and A. Nedić, "Improved temporal difference methods with linear function approximation," *Learning and Approximate Dynamic Programming*, pp. 231–255, 2004.

[36] T. Jung and D. Polani, "Kernelizing LSPE(λ)," in *IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, 2007, pp. 338–345.

[37] P. J. Schweitzer and A. Seidmann, "Generalized polynomial approximations in Markovian decision processes," *Journal of Mathematical Analysis and Applications*, vol. 110, no. 2, pp. 568–582, Sep. 1985.

[38] W. Sun and J. A. Bagnell, "Online Bellman residual and temporal difference algorithms with predictive error guarantees," in *International Joint Conference on Artificial Intelligence*, 2016, pp. 4213–4217.

[39] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine Learning*, vol. 3, no. 1, pp. 9–44, Aug. 1988.

[40] J. N. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," *IEEE Transactions on Automatic Control*, vol. 42, no. 5, pp. 674–690, 1997.

[41] J. Bae, P. Chhatbar, J. T. Francis, J. C. Sanchez, and J. C. Príncipe, "Reinforcement learning via kernel temporal difference," in *IEEE EMBS*, 2011, pp. 5662–5665.

[42] S. J. Bradtke and A. G. Barto, "Linear least-squares algorithms for temporal difference learning," *Machine Learning*, vol. 22, no. 1, pp. 33–57, Mar. 1996.

[43] J. A. Boyan, "Technical update: Least-squares temporal difference learning," *Machine Learning*, vol. 49, no. 2, pp. 233–246, Nov. 2002.

[44] M. G. Lagoudakis and R. Parr, "Least-squares policy iteration," *J. Mach. Learn. Res.*, vol. 4, pp. 1107–1149, Dec. 2003.

[45] A.-M. Farahmand, M. Ghavamzadeh, C. Szepesvári, and S. Mannor, "Regularized policy iteration with nonparametric function spaces," *J. Machine Learning Research*, vol. 17, no. 1, pp. 4809–4874, 2016.

[46] Z. Qin, W. Li, and F. Janoos, "Sparse reinforcement learning via convex optimization," in *Intern. Conf. on Machine Learning (ICML)*, E. P. Xing and T. Jebara, Eds., ser. Proceedings of Machine Learning Research, vol. 32, Bejing, China: PMLR, 2014, pp. 424–432.

[47] S. Mahadevan, B. Liu, P. S. Thomas, W. Dabney, S. Giguere, N. Jacek, I. Gemp, and J. Liu, "Proximal reinforcement learning: A new theory of sequential decision making in primal-dual spaces," *arXiv:1405.6757*, 2014.

[48] B. Liu, I. Gemp, M. Ghavamzadeh, J. Liu, S. Mahadevan, and M. Petrik, "Proximal gradient temporal difference learning: Stable reinforcement learning with polynomial sample complexity," *J. Artif. Int. Res.*, vol. 63, no. 1, pp. 461–494, Sep. 2018.

[49] L. Song, J. Huang, A. Smola, and K. Fukumizu, "Hilbert space embeddings of conditional distributions with applications to dynamical systems," in *Intern. Conf. on Machine Learning (ICML)*, Montreal, Quebec, Canada, 2009, pp. 961–968.

[50] L. Song, B. Boots, S. M. Siddiqi, G. Gordon, and A. Smola, "Hilbert space embeddings of hidden Markov models," in *Intern. Conf. on Machine Learning (ICML)*, Haifa, Israel, 2010, pp. 991–998.

[51] A. Barreto, D. Precup, and J. Pineau, "Reinforcement learning using kernel-based stochastic factorization," in *NIPS*, vol. 24, 2011.

[52] A. Barreto, D. Precup, and J. Pineau, "On-line reinforcement learning using incremental kernel-based stochastic factorization," in *NIPS*, vol. 25, 2012.

[53] B. Kveton and G. Theocharous, "Structured kernel-based reinforcement learning," in *AAAI Conference on Artificial Intelligence*, vol. 27, Jun. 2013, pp. 569–575.

[54] B. Kveton and G. Theocharous, "Kernel-based reinforcement learning on representative states," in *AAAI Conference on Artificial Intelligence*, vol. 26, Sep. 2021, pp. 977–983.

[55] K. Slavakis and I. Yamada, "Fejér-monotone hybrid steepest descent method for affinely constrained and composite convex minimization tasks," *Optimization*, vol. 67, no. 11, pp. 1963–2001, Nov. 2018.

[56] E. Kreyszig, *Introductory Functional Analysis with Applications* (Wiley Classics Library). Wiley, 1991.

[57] C. G. den Broeder Jr. and A. Charnes, "Contributions to the theory of generalized inverses for matrices," Purdue University, Lafayette: IN, Tech. Rep., 1957.

[58] A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications* (CMS Books in Mathematics), 2nd. New York: Springer, 2003.

[59] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence*, vol. 101, no. 1, pp. 99–134, 1998.

[60] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *NIPS*, vol. 20, 2007.

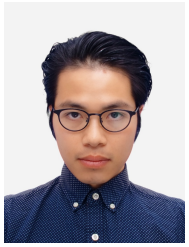
[61] R. B. Ash and C. A. Doléans-Dade, *Probability and Measure Theory*, 2nd ed. Academic Press, 2000.

[62] D. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex Analysis and Optimization*. Athena Scientific, 2003.

[63] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces," *J. Machine Learning Research*, vol. 5, no. Jan, pp. 73–99, 2004.

[64] J. B. Conway, *A Course in Functional Analysis*, 2nd ed. New York: Springer, 1990.

Yuki Akiyama received the B.E. in information and communication from Tokyo Institute of Technology (now the Institute of Science Tokyo), Japan in 2023 and is currently pursuing the M.E. there. His research interests include signal processing and reinforcement learning.



Minh Vu received the B.E. and M.E. in information and communications from the Tokyo Institute of Technology (now the Institute of Science Tokyo), Japan, in 2022 and 2024, respectively and now is currently pursuing the Ph.D degree there. His research interests include reinforcement learning, signal processing and convex optimization. From 2018, he is a recipient of the Japanese Government (MEXT) Scholarship.

Konstantinos Slavakis (M '08, SM '12) earned his PhD degree in Electrical & Electronic Eng. from Tokyo Institute of Technology, Japan. He joined Institute of Science Tokyo, Japan, as a professor in 2021. He has served IEEE Transactions on Signal Processing as an associate editor ['09–'13] and as an area editor ['10–'15]. He is currently a subject area editor of Signal Processing and a member of the senior-editorial board of the IEEE Signal Processing Magazine. He has also been a member of the IEEE Signal Processing Theory and Methods (SPTM) technical committee ['12–'17]. His research interests are in the areas of signal processing and machine learning.

Supplementary File

Appendix A
Proof of Proposition 1

First, recall that the orthogonal projection mapping $P_{\mathbf{\Phi}_{\mathcal{T}_N}}$ onto the linear span of $\{\varphi(\mathbf{z}_i)\}_{i=1}^N$ is provided by $P_{\mathbf{\Phi}_{\mathcal{T}_N}} = \mathbf{\Phi}_{\mathcal{T}_N}\mathbf{K}_{\mathcal{T}_N}^\dagger\mathbf{\Phi}_{\mathcal{T}_N}^\intercal$. Moreover, let $P_{\mathbf{\Phi}_{\mathcal{T}_N}^\perp}$ denote the orthogonal projection mapping onto the orthogonal complement of the linear span of $\{\varphi(\mathbf{z}_i)\}_{i=1}^N$. The Pythagoras theorem states that $\mathrm{Id} = P_{\mathbf{\Phi}_{\mathcal{T}_N}} + P_{\mathbf{\Phi}_{\mathcal{T}_N}^\perp}$, where Id is the identity operator in $\mathcal{H}$.

Define $\mathbf{c} \coloneqq \boldsymbol{\gamma} + \alpha\boldsymbol{\Upsilon}\mathbf{\Phi}_{\mu_\star}^{\mathrm{av}\intercal}Q$ for convenience, and observe that

$$\begin{aligned}
&(\mathbf{c} - \mathbf{K}_{\mathcal{T}_N}^\dagger\mathbf{\Phi}_{\mathcal{T}_N}^\intercal Q)^\intercal\mathbf{K}_{\mathcal{T}_N}(\mathbf{c} - \mathbf{K}_{\mathcal{T}_N}^\dagger\mathbf{\Phi}_{\mathcal{T}_N}^\intercal Q)\\
&= (\mathbf{c} - \mathbf{K}_{\mathcal{T}_N}^\dagger\mathbf{\Phi}_{\mathcal{T}_N}^\intercal Q)^\intercal\mathbf{\Phi}_{\mathcal{T}_N}^\intercal\mathbf{\Phi}_{\mathcal{T}_N}(\mathbf{c} - \mathbf{K}_{\mathcal{T}_N}^\dagger\mathbf{\Phi}_{\mathcal{T}_N}^\intercal Q)\\
&= \langle\mathbf{\Phi}_{\mathcal{T}_N}(\mathbf{c} - \mathbf{K}_{\mathcal{T}_N}^\dagger\mathbf{\Phi}_{\mathcal{T}_N}^\intercal Q) \mid \mathbf{\Phi}_{\mathcal{T}_N}(\mathbf{c} - \mathbf{K}_{\mathcal{T}_N}^\dagger\mathbf{\Phi}_{\mathcal{T}_N}^\intercal Q)\rangle_{\mathcal{H}}\\
&= \|\mathbf{\Phi}_{\mathcal{T}_N}\mathbf{c} - \mathbf{\Phi}_{\mathcal{T}_N}\mathbf{K}_{\mathcal{T}_N}^\dagger\mathbf{\Phi}_{\mathcal{T}_N}^\intercal Q\|_{\mathcal{H}}^2\\
&= \|\mathbf{\Phi}_{\mathcal{T}_N}\mathbf{c} - P_{\mathbf{\Phi}_{\mathcal{T}_N}}Q\|_{\mathcal{H}}^2\,. \qquad (40)
\end{aligned}$$

A change of variables and the Pythagoras theorem suggest that

$$\begin{aligned}
&T_{\mathrm{LSPE},\mu}(Q)\\
&= \arg\min_{Q'\in\mathcal{H}}\ \|\mathbf{\Phi}_{\mathcal{T}_N}^\intercal Q' - \mathbf{g} - \alpha\mathbf{\Phi}_\mu'^\intercal Q\|_{\mathbb{R}^N}^2\\
&\qquad\qquad + \sigma\|Q' - Q\|_{\mathcal{H}}^2\\
&= \arg\min_{Q'\in\mathcal{H}}\ \|\mathbf{\Phi}_{\mathcal{T}_N}^\intercal(Q' - P_{\mathbf{\Phi}_{\mathcal{T}_N}^\perp}Q) - \mathbf{g} - \alpha\mathbf{\Phi}_\mu'^\intercal Q\|_{\mathbb{R}^N}^2\\
&\qquad\qquad + \sigma\|Q' - Q\|_{\mathcal{H}}^2\\
&= \arg\min_{Q''\in\mathcal{H}}\ \|\mathbf{\Phi}_{\mathcal{T}_N}^\intercal Q'' - \mathbf{g} - \alpha\mathbf{\Phi}_\mu'^\intercal Q\|_{\mathbb{R}^N}^2\\
&\qquad\qquad + \sigma\|Q'' + P_{\mathbf{\Phi}_{\mathcal{T}_N}^\perp}Q - Q\|_{\mathcal{H}}^2 + P_{\mathbf{\Phi}_{\mathcal{T}_N}^\perp}Q\\
&= \arg\min_{Q''\in\mathcal{H}}\ \|\mathbf{\Phi}_{\mathcal{T}_N}^\intercal Q'' - \mathbf{g} - \alpha\mathbf{\Phi}_\mu'^\intercal Q\|_{\mathbb{R}^N}^2\\
&\qquad\qquad + \sigma\|Q'' - P_{\mathbf{\Phi}_{\mathcal{T}_N}}Q\|_{\mathcal{H}}^2 + P_{\mathbf{\Phi}_{\mathcal{T}_N}^\perp}Q \qquad (41a)\\
&= \mathbf{\Phi}_{\mathcal{T}_N}(\mathbf{K}_{\mathcal{T}_N} + \sigma\mathbf{I}_N)^{-1}\mathbf{g}\\
&\qquad\qquad + \sigma\mathbf{\Phi}_{\mathcal{T}_N}(\mathbf{K}_{\mathcal{T}_N} + \sigma\mathbf{I}_N)^{-1}\mathbf{K}_{\mathcal{T}_N}^\dagger\mathbf{\Phi}_{\mathcal{T}_N}^\intercal Q\\
&\qquad\qquad + \alpha\mathbf{\Phi}_{\mathcal{T}_N}(\mathbf{K}_{\mathcal{T}_N} + \sigma\mathbf{I}_N)^{-1}\mathbf{\Phi}_\mu'^\intercal Q\\
&\qquad\qquad + P_{\mathbf{\Phi}_{\mathcal{T}_N}^\perp}Q\,, \qquad (41b)
\end{aligned}$$

where $\mathbf{\Phi}_{\mathcal{T}_N}(\mathbf{K}_{\mathcal{T}_N} + \sigma\mathbf{I}_N)^{-1} = (\mathbf{\Phi}_{\mathcal{T}_N}\mathbf{\Phi}_{\mathcal{T}_N}^\intercal + \sigma\,\mathrm{Id})^{-1}\mathbf{\Phi}_{\mathcal{T}_N}$ was used in the last equality. Because of the assumption on Q, $P_{\mathbf{\Phi}_{\mathcal{T}_N}^\perp}Q = 0$.

Notice again by the Pythagoras theorem and $P_{\mathbf{\Phi}_{\mathcal{T}_N}}^2 = P_{\mathbf{\Phi}_{\mathcal{T}_N}}$ that

$$\begin{aligned}
&\|\mathbf{\Phi}_{\mathcal{T}_N}^\intercal Q'' - \mathbf{g} - \alpha\mathbf{\Phi}_\mu'^\intercal Q\|_{\mathbb{R}^N}^2 + \sigma\|Q'' - P_{\mathbf{\Phi}_{\mathcal{T}_N}}Q\|_{\mathcal{H}}^2\\
&\geq \|\mathbf{\Phi}_{\mathcal{T}_N}^\intercal Q'' - \mathbf{g} - \alpha\mathbf{\Phi}_\mu'^\intercal Q\|_{\mathbb{R}^N}^2 + \sigma\|P_{\mathbf{\Phi}_{\mathcal{T}_N}}(Q'' - P_{\mathbf{\Phi}_{\mathcal{T}_N}}Q)\|_{\mathcal{H}}^2\\
&= \|\mathbf{\Phi}_{\mathcal{T}_N}^\intercal P_{\mathbf{\Phi}_{\mathcal{T}_N}}Q'' - \mathbf{g} - \alpha\mathbf{\Phi}_\mu'^\intercal Q\|_{\mathbb{R}^N}^2\\
&\quad + \sigma\|P_{\mathbf{\Phi}_{\mathcal{T}_N}}Q'' - P_{\mathbf{\Phi}_{\mathcal{T}_N}}Q\|_{\mathcal{H}}^2\,,
\end{aligned}$$

which clearly suggests that the minimizer in (41a) lies in the linear span of $\{\varphi(\mathbf{z}_i)\}_{i=1}^N \Leftrightarrow (Q'' = \mathbf{\Phi}_{\mathcal{T}_N}\mathbf{c},\ \text{for } \exists\mathbf{c}\in\mathbb{R}^N) \Leftrightarrow (Q'' = P_{\mathbf{\Phi}_{\mathcal{T}_N}}Q'')$. Hence,

$$\begin{aligned}
&T_{\mathrm{LSPE},\mu}Q\\
&= \arg\min_{Q''\in\mathcal{H}\,|\,Q''=P_{\mathbf{\Phi}_{\mathcal{T}_N}}Q''}\|\mathbf{\Phi}_{\mathcal{T}_N}^\intercal Q'' - \mathbf{g} - \alpha\mathbf{\Phi}_\mu'^\intercal Q\|_{\mathbb{R}^N}^2\\
&\qquad\qquad + \sigma\|Q'' - P_{\mathbf{\Phi}_{\mathcal{T}_N}}Q\|_{\mathcal{H}}^2\\
&= \mathbf{\Phi}_{\mathcal{T}_N}\mathbf{c}_\star\,,
\end{aligned}$$

where $\mathbf{c}_\star$ satisfies

$$\begin{aligned}
&\arg\min_{\mathbf{c}\in\mathbb{R}^N}\|\mathbf{\Phi}_{\mathcal{T}_N}^\intercal\mathbf{\Phi}_{\mathcal{T}_N}\mathbf{c} - \mathbf{g} - \alpha\mathbf{\Phi}_\mu'^\intercal Q\|_{\mathbb{R}^N}^2\\
&\qquad + \sigma\|\mathbf{\Phi}_{\mathcal{T}_N}\mathbf{c} - P_{\mathbf{\Phi}_{\mathcal{T}_N}}Q\|_{\mathcal{H}}^2 \qquad (42)\\
&= \arg\min_{\mathbf{c}\in\mathbb{R}^N}\overbrace{\|\mathbf{K}_{\mathcal{T}_N}\mathbf{c} - \mathbf{g} - \alpha\mathbf{\Phi}_\mu'^\intercal Q\|_{\mathbb{R}^N}^2}^{\mathcal{L}(\boldsymbol{\gamma},\boldsymbol{\Upsilon})}\\
&\qquad + \sigma\underbrace{(\mathbf{c} - \mathbf{K}_{\mathcal{T}_N}^\dagger\mathbf{\Phi}_{\mathcal{T}_N}^\intercal Q)^\intercal\mathbf{K}_{\mathcal{T}_N}(\mathbf{c} - \mathbf{K}_{\mathcal{T}_N}^\dagger\mathbf{\Phi}_{\mathcal{T}_N}^\intercal Q)}_{\mathcal{R}(\boldsymbol{\gamma},\boldsymbol{\Upsilon})}\\
&\ni\ \mathbf{c}_\star\\
&\coloneqq (\mathbf{K}_{\mathcal{T}_N} + \sigma\mathbf{I}_N)^{-1}\big(\mathbf{g} + \sigma\mathbf{K}_{\mathcal{T}_N}^\dagger\mathbf{\Phi}_{\mathcal{T}_N}^\intercal Q + \alpha\mathbf{\Phi}_\mu'^\intercal Q\big)\\
&= (\mathbf{K}_{\mathcal{T}_N} + \sigma\mathbf{I}_N)^{-1}\mathbf{g}\\
&\qquad + (\mathbf{K}_{\mathcal{T}_N} + \sigma\mathbf{I}_N)^{-1}[\sigma\mathbf{K}_{\mathcal{T}_N}^\dagger, \alpha\mathbf{I}_N]\begin{bmatrix}\mathbf{\Phi}_{\mathcal{T}_N}^\intercal Q\\ \mathbf{\Phi}_\mu'^\intercal Q\end{bmatrix}\\
&= (\mathbf{K}_{\mathcal{T}_N} + \sigma\mathbf{I}_N)^{-1}\mathbf{g}\\
&\qquad + \alpha(\mathbf{K}_{\mathcal{T}_N} + \sigma\mathbf{I}_N)^{-1}[(\sigma/\alpha)\mathbf{K}_{\mathcal{T}_N}^\dagger, \mathbf{I}_N]\mathbf{\Phi}_{\mu_\star}^{\mathrm{av}\intercal}Q\\
&= \boldsymbol{\gamma}_\star + \alpha\boldsymbol{\Upsilon}_\star\mathbf{\Phi}_{\mu_\star}^{\mathrm{av}\intercal}Q\,,
\end{aligned}$$

and where $\boldsymbol{\gamma}_\star \coloneqq (\mathbf{K}_{\mathcal{T}_N} + \sigma\mathbf{I}_N)^{-1}\mathbf{g}$ and $\boldsymbol{\Upsilon}_\star \coloneqq (\mathbf{K}_{\mathcal{T}_N} + \sigma\mathbf{I}_N)^{-1}[(\sigma/\alpha)\mathbf{K}_{\mathcal{T}_N}^\dagger, \mathbf{I}_N]$.

Under the light of (40), the previous findings are summarized as follows: $(\boldsymbol{\gamma}_\star, \boldsymbol{\Upsilon}_\star)$ satisfies (4) iff $\mathbf{c}_\star = \boldsymbol{\gamma}_\star + \alpha\boldsymbol{\Upsilon}_\star\mathbf{\Phi}_{\mu_\star}^{\mathrm{av}\intercal}Q$ is one of the minimizers in (42) iff $T_{\mathrm{LSPE},\mu}(Q) = \mathbf{\Phi}_{\mathcal{T}_N}\mathbf{c}_\star = T_{\mu_\star}(Q)$. These equivalences establish the claim of Proposition 1(i). As an additional remark, notice that $T_{\mathrm{LSPE},\mu}(Q) = Q$ in (41b) yields (14).

The proofs of Propositions 1(ii) and 1(iii) follow similar steps with the proof of Proposition 1(i) and are thus skipped.

Appendix B
Proof of Theorem 2

First, the following lemma and its proof are in order.

Lemma 11. For any $Q_1, Q_2 \in \mathcal{H}$, $\mathbf{s}\in\mathfrak{S}$, there exists $\delta \neq 0$ s.t. for any $\epsilon \in (0, \sqrt{|\delta|})$ an $a' \in \mathfrak{A}$ can be always selected s.t.

$$\begin{aligned}
&\big[\inf_{a\in\mathfrak{A}}Q_1(\mathbf{s},a) - \inf_{a\in\mathfrak{A}}Q_2(\mathbf{s},a)\big]^2\\
&\qquad\leq [Q_1(\mathbf{s},a') - Q_2(\mathbf{s},a')]^2 + \epsilon\,. \qquad (43)
\end{aligned}$$

Proof: Define $\delta \coloneqq \inf_{a\in\mathfrak{A}}Q_1(\mathbf{s},a) - \inf_{a\in\mathfrak{A}}Q_2(\mathbf{s},a)$. Whenever $\delta = 0$, (43) holds true trivially. Consider now the case where $\delta > 0$, take any $\epsilon \in (0, \sqrt{\delta})$, and define $\epsilon' \coloneqq \delta - \sqrt{\delta^2 - \epsilon} > 0$ so that $\epsilon = 2\epsilon'\delta - \epsilon'^2$ and $\delta - \epsilon' > 0$. Then, $\exists a' \in \mathfrak{A}$ s.t. $Q_2(\mathbf{s},a') \leq \inf_{a\in\mathfrak{A}}Q_2(\mathbf{s},a) + \epsilon'$. This result together with $\inf_{a\in\mathfrak{A}}Q_1(\mathbf{s},a) \leq Q_1(\mathbf{s},a')$ suggest:

$$Q_1(\mathbf{s},a') - Q_2(\mathbf{s},a') \geq \inf_{a\in\mathfrak{A}}Q_1(\mathbf{s},a) - \inf_{a\in\mathfrak{A}}Q_2(\mathbf{s},a) - \epsilon'\,,$$

where the right-hand side is positive, and

$$\begin{aligned}&[Q_1(\mathbf{s},a')-Q_2(\mathbf{s},a')]^2\\&\geq [\inf_{a\in\mathfrak{A}}Q_1(\mathbf{s},a)-\inf_{a\in\mathfrak{A}}Q_2(\mathbf{s},a)]^2-2\epsilon'\delta+\epsilon'^2\\&=[\inf_{a\in\mathfrak{A}}Q_1(\mathbf{s},a)-\inf_{a\in\mathfrak{A}}Q_2(\mathbf{s},a)]^2-\epsilon\,,\end{aligned}$$

which establishes (43). The proof for the case $\delta<0$ follows exactly the previous steps by using $-\delta$ instead of δ and by interchanging Q_1 with Q_2. ∎

Property (18a) is established as follows: $\forall Q_1,Q_2\in\mathcal{H}$,

$$\begin{aligned}&\|T_\mu(Q_1)-T_\mu(Q_2)\|_{\mathcal{H}}^2\\&=\alpha^2\langle\boldsymbol{\Psi}\boldsymbol{\Phi}_\mu^{\text{av}\intercal}(Q_1-Q_2)\mid\boldsymbol{\Psi}\boldsymbol{\Phi}_\mu^{\text{av}\intercal}(Q_1-Q_2)\rangle_{\mathcal{H}}\\&=\alpha^2\langle\boldsymbol{\Phi}_\mu^{\text{av}\intercal}(Q_1-Q_2)\mid\boldsymbol{\Psi}^\intercal\boldsymbol{\Psi}\boldsymbol{\Phi}_\mu^{\text{av}\intercal}(Q_1-Q_2)\rangle\\&\leq\alpha^2\|\boldsymbol{\Psi}^\intercal\boldsymbol{\Psi}\|_2\langle Q_1-Q_2\mid\boldsymbol{\Phi}_\mu^{\text{av}}\boldsymbol{\Phi}_\mu^{\text{av}\intercal}(Q_1-Q_2)\rangle_{\mathcal{H}}\\&\leq\alpha^2\|\mathbf{K}_\Psi\|_2\cdot\|\boldsymbol{\Phi}_\mu^{\text{av}}\boldsymbol{\Phi}_\mu^{\text{av}\intercal}\|\cdot\|Q_1-Q_2\|_{\mathcal{H}}^2\\&\leq\beta^2\|Q_1-Q_2\|_{\mathcal{H}}^2\,,\end{aligned}$$

where observation $\|\boldsymbol{\Phi}_\mu^{\text{av}}\boldsymbol{\Phi}_\mu^{\text{av}\intercal}\|=\|\boldsymbol{\Phi}_\mu^{\text{av}\intercal}\boldsymbol{\Phi}_\mu^{\text{av}}\|_2=\|\mathbf{K}_\mu^{\text{av}}\|_2$ and (19) were used to obtain the last inequality.

The proof of (18b) follows. For any $Q_1,Q_2\in\mathcal{H}$,

$$\begin{aligned}&\|T(Q_1)-T(Q_2)\|_{\mathcal{H}}^2\\&=\alpha^2\|\boldsymbol{\Psi}(\mathbf{inf}_{\mu\in\mathcal{M}}\boldsymbol{\Phi}_\mu^{\text{av}\intercal}Q_1-\mathbf{inf}_{\mu\in\mathcal{M}}\boldsymbol{\Phi}_\mu^{\text{av}\intercal}Q_2)\|_{\mathcal{H}}^2\\&\leq\alpha^2\|\mathbf{K}_\Psi\|_2\,\|\mathbf{inf}_{\mu\in\mathcal{M}}\boldsymbol{\Phi}_\mu^{\text{av}\intercal}Q_1-\mathbf{inf}_{\mu\in\mathcal{M}}\boldsymbol{\Phi}_\mu^{\text{av}\intercal}Q_2\|^2\\&=\alpha^2\|\mathbf{K}_\Psi\|_2\sum_{i=1}^{N_{\text{av}}}\big[\inf_{a_i\in\mathfrak{A}}Q_1(\mathbf{s}_i^{\text{av}},a_i)-\inf_{a_i\in\mathfrak{A}}Q_2(\mathbf{s}_i^{\text{av}},a_i)\big]^2.\end{aligned}\tag{44}$$

According to Lemma 11, there exist $\{\delta_i\neq 0\}_{i=1}^{N_{\text{av}}}$ s.t. for any $\epsilon\in(0,\min_{i\in\{1,\ldots,N_{\text{av}}\}}\{\sqrt{|\delta_i|}\})$ actions $\{a_i'\}_{i=1}^{N_{\text{av}}}$ can be selected so that (44) yields

$$\begin{aligned}&\|T(Q_1)-T(Q_2)\|_{\mathcal{H}}^2\\&\leq\alpha^2\|\mathbf{K}_\Psi\|_2\sum_{i=1}^{N_{\text{av}}}\big[Q_1(\mathbf{s}_i^{\text{av}},a_i')-Q_2(\mathbf{s}_i^{\text{av}},a_i')\big]^2\\&\quad+\epsilon\alpha^2N_{\text{av}}\|\mathbf{K}_\Psi\|_2\,.\end{aligned}\tag{45}$$

Take now any stationary policy $\mu'\in\mathcal{M}$ s.t. $\mu'(\mathbf{s}_i^{\text{av}})=a_i'$, $\forall i$. Then, by $\boldsymbol{\Phi}_{\mu'}^{\text{av}}\coloneqq[\varphi(\mathbf{s}_1^{\text{av}},\mu'(\mathbf{s}_1^{\text{av}})),\ldots,\varphi(\mathbf{s}_{N_{\text{av}}}^{\text{av}},\mu'(\mathbf{s}_{N_{\text{av}}}^{\text{av}}))]$,

$$\begin{aligned}&\sum\nolimits_{i=1}^{N_{\text{av}}}\big[Q_1(\mathbf{s}_i^{\text{av}},a_i')-Q_2(\mathbf{s}_i^{\text{av}},a_i')\big]^2\\&=\sum\nolimits_{i=1}^{N_{\text{av}}}\big[Q_1(\mathbf{s}_i^{\text{av}},\mu'(\mathbf{s}_i^{\text{av}}))-Q_2(\mathbf{s}_i^{\text{av}},\mu'(\mathbf{s}_i^{\text{av}}))\big]^2\\&=\|\boldsymbol{\Phi}_{\mu'}^{\text{av}\intercal}(Q_1-Q_2)\|^2\\&=\langle\boldsymbol{\Phi}_{\mu'}^{\text{av}\intercal}(Q_1-Q_2)\mid\boldsymbol{\Phi}_{\mu'}^{\text{av}\intercal}(Q_1-Q_2)\rangle\\&=\langle Q_1-Q_2\mid\boldsymbol{\Phi}_{\mu'}^{\text{av}}\boldsymbol{\Phi}_{\mu'}^{\text{av}\intercal}(Q_1-Q_2)\rangle_{\mathcal{H}}\\&\leq\|\mathbf{K}_{\mu'}^{\text{av}}\|_2\,\|Q_1-Q_2\|_{\mathcal{H}}^2\,,\end{aligned}$$

so that (45) results in

$$\begin{aligned}&\|T(Q_1)-T(Q_2)\|_{\mathcal{H}}^2\\&\leq\alpha^2\|\mathbf{K}_\Psi\|_2\sup_{\mu'\in\mathcal{H}}\|\mathbf{K}_{\mu'}^{\text{av}}\|_2\,\|Q_1-Q_2\|_{\mathcal{H}}^2+\epsilon\alpha^2N_{\text{av}}\|\mathbf{K}_\Psi\|_2\\&=\beta^2\|Q_1-Q_2\|_{\mathcal{H}}^2+\epsilon\alpha^2N_{\text{av}}\|\mathbf{K}_\Psi\|_2\,.\end{aligned}$$

Since ϵ can be made arbitrarily small, this last inequality establishes (18b).

Appendix C
Proof of Theorem 4

The proof is built on arguments of [32, 49, 63]. In particular, following [32, §3, §4.1] and for an arbitrarily fixed policy $\mu(\cdot)\in\mathcal{M}$, define the linear covariance operators $\Sigma_{zz},\Sigma_{s'z}^\mu,\Sigma_{s'|z}^\mu\colon\mathcal{H}\to\mathcal{H}$ by

$$\Sigma_{zz}(Q)\coloneqq\mathbb{E}_{\mathbf{z}}\{\langle\varphi(\mathbf{z})\mid Q\rangle_{\mathcal{H}}\,\varphi(\mathbf{z})\}\,,\tag{46a}$$

$$\Sigma_{s'z}^\mu(Q)\coloneqq\mathbb{E}_{(\mathbf{s}',\mathbf{z})}\{\langle\varphi(\mathbf{s}',\mu(\mathbf{s}'))\mid Q\rangle_{\mathcal{H}}\,\varphi(\mathbf{z})\}\,,\tag{46b}$$

$$\Sigma_{s'|z}^\mu(Q)\coloneqq\Sigma_{zz}^{-1}\Sigma_{s'z}^\mu(Q)\,,\tag{46c}$$

where $\mathbb{E}_{\mathbf{z}}\{\cdot\}$ denotes expectation with respect to (w.r.t.) the σ-subalgebra generated by $\mathbf{z}$ [30], similarly for $\mathbb{E}_{(\mathbf{s}',\mathbf{z})}\{\cdot\}$, and Σ_{zz}^{-1} stands for the inverse of Σ_{zz}. It is important to stress here that the expectation symbols in (46) are considered in the following sense. In (46a), for example, for arbitrarily fixed $Q\in\mathcal{H}$, $\Sigma_{zz}(Q)$ stands for the unique point of $\mathcal{H}$ which satisfies, according to the Riesz representation theorem [64], $\langle\Sigma_{zz}(Q)\mid h\rangle=L_{zz}(h)$, $\forall h\in\mathcal{H}$, where $L_{zz}(\cdot)$ is the linear continuous operator defined by $L_{zz}(\cdot)\colon\mathcal{H}\to\mathbb{R}\colon h\mapsto L_{zz}(h)\coloneqq\mathbb{E}_{\mathbf{z}}\{\langle\langle Q\mid\varphi(\mathbf{z})\rangle_{\mathcal{H}}\cdot\varphi(\mathbf{z})\mid h\rangle_{\mathcal{H}}\}$, and where expectation is taken here in the usual sense [30]. Operators (46b) and (46c) are defined in a similar way.

Recall here also the definitions for the minimum $\sigma_{\min}(\mathcal{A})$ and maximum $\sigma_{\max}(\mathcal{A})$ spectral values of a linear bounded and self-adjoint operator $\mathcal{A}\colon\mathcal{H}\to\mathcal{H}$ [56, Thm. 9.2-3]:

$$\sigma_{\min}(\mathcal{A})\coloneqq\inf\nolimits_{\|h\|_{\mathcal{H}}=1}\langle h\mid\mathcal{A}(h)\rangle_{\mathcal{H}}\,,\tag{47a}$$

$$\sigma_{\max}(\mathcal{A})\coloneqq\sup\nolimits_{\|h\|_{\mathcal{H}}=1}\langle h\mid\mathcal{A}(h)\rangle_{\mathcal{H}}=\|\mathcal{A}\|\,.\tag{47b}$$

It can be verified by [63, Thm. 2] and [32, §4.1] that $\mathbb{E}_{\mathbf{s}'|\mathbf{z}}\{Q(\mathbf{s}',\mu(\mathbf{s}'))\}=\langle\Sigma_{s'|z}^{\mu\,*}(\varphi(\mathbf{z}))\mid Q\rangle_{\mathcal{H}}$, where $*$ denotes the adjoint of a linear operator [64]. Hence, by the reproducing property of the kernel in $\mathcal{H}$, $\mathbb{E}_{\mathbf{s}'|\mathbf{z}}\{Q(\mathbf{s}',\mu(\mathbf{s}'))\}=\langle\varphi(\mathbf{z})\mid\Sigma_{s'|z}^\mu(Q)\rangle_{\mathcal{H}}=\Sigma_{s'|z}^\mu(Q)(\mathbf{z})$, and the classical B-Maps (2) take the following equivalent form:

$$(T_\mu^\diamond Q)(\mathbf{z})\coloneqq g(\mathbf{z})+\alpha\Sigma_{s'|z}^\mu(Q)(\mathbf{z})\,,\tag{48a}$$

$$(T^\diamond Q)(\mathbf{z})\coloneqq g(\mathbf{z})+\alpha\Sigma_{s'|z}^{\mu_Q}(Q)(\mathbf{z})\,,\tag{48b}$$

where the stationary policy $\mu_Q(\cdot)$ is defined as in (2c). Along the lines of Assumption 3(iii) and [32, (7)], define

$$\begin{aligned}\hat{\Sigma}_{s'|z}^\mu&\coloneqq\hat{\Sigma}_{s'|z}^\mu(N)\\&\coloneqq\boldsymbol{\Phi}_{\mathcal{T}_N}(\mathbf{K}_{\mathcal{T}_N}+N\sigma_N'\mathbf{I}_N)^{-1}\boldsymbol{\Phi}_\mu^{\text{av}\intercal}\\&=\tfrac{1}{\sqrt{N}}\boldsymbol{\Phi}_{\mathcal{T}_N}\left(\tfrac{1}{N}\mathbf{K}_{\mathcal{T}_N}+\sigma_N'\mathbf{I}_N\right)^{-1}\tfrac{1}{\sqrt{N}}\boldsymbol{\Phi}_\mu^{\text{av}\intercal}\,,\end{aligned}\tag{49}$$

so that (20a) is recast as

$$(T_\mu Q)(\mathbf{z})\coloneqq g(\mathbf{z})+\alpha\hat{\Sigma}_{s'|z}^\mu(Q)(\mathbf{z})\,.\tag{50}$$

The following theorem asserts that the previous quantity is a consistent finite-sample estimate of (46c).

Theorem 12. Assumption 3(i) means that

$$\sum\nolimits_{i=1}^{+\infty}\|\Sigma_{zz}^{-3/2}\Sigma_{s'z}^\mu e_i\|_{\mathcal{H}}^2<+\infty,$$

for a countable orthonormal basis $(e_i)_{i=1}^{+\infty}$ of $\mathcal{H}$ [64, p. 267]. Under also Assumption 3(iv), $\mathbb{P}\text{-}\lim_{N\to\infty}\|\Sigma_{s'|z}^\mu-\hat{\Sigma}_{s'|z}^\mu(N)\|=0$.

Proof: [32, Thm. 1] yields $\mathbb{P}\text{-}\lim_{N\to\infty}\|\Sigma^{\mu}_{s'|z} - \hat{\Sigma}^{\mu}_{s'|z}(N)\|_{\text{HS}} = 0$, where $\|\cdot\|_{\text{HS}}$ stands for the Hilbert-Schmidt norm of an operator [64, p. 267]. Consequently, the fact $\|\cdot\| \leq \|\cdot\|_{\text{HS}}$ [64, p. 267] establishes the claim of the theorem. ∎

The claim of Theorem 4 that T_μ and T are contractions follows directly by (18) and assumption $\beta(N) \leq \beta_\infty < 1$, a.s. Now, the triangle inequality and Assumptions 3 suggest

$$\begin{aligned}
&\|T^{\diamond}_{\mu}(Q_1) - T^{\diamond}_{\mu}(Q_2)\|_{\mathcal{H}} && \text{(51a)}\\
&\leq \|T^{\diamond}_{\mu}(Q_1) - T_{\mu}(Q_1)\|_{\mathcal{H}} + \|T^{\diamond}_{\mu}(Q_2) - T_{\mu}(Q_2)\|_{\mathcal{H}}\\
&\quad + \|T_{\mu}(Q_1) - T_{\mu}(Q_2)\|_{\mathcal{H}}\\
&\leq \alpha\|\Sigma^{\mu}_{s'|z} - \hat{\Sigma}^{\mu}_{s'|z}(N)\|\,\|Q_1\|_{\mathcal{H}} && \text{(51b)}\\
&\quad + \alpha\|\Sigma^{\mu}_{s'|z} - \hat{\Sigma}^{\mu}_{s'|z}(N)\|\,\|Q_2\|_{\mathcal{H}} && \text{(51c)}\\
&\quad + \beta_\infty\|Q_1 - Q_2\|_{\mathcal{H}}\,. && \text{(51d)}
\end{aligned}$$

By applying $\mathbb{P}\text{-}\lim_{N\to\infty}$ to (51) and by Theorem 12, it can be verified that $\|T^{\diamond}_{\mu}(Q_1) - T^{\diamond}_{\mu}(Q_2)\|_{\mathcal{H}} \leq \beta_\infty\|Q_1 - Q_2\|_{\mathcal{H}}$, $\forall Q_1, Q_2 \in \mathcal{H}$, a.s. The claim $\|T^{\diamond}(Q_1) - T^{\diamond}(Q_2)\|_{\mathcal{H}} \leq \beta_\infty\|Q_1 - Q_2\|_{\mathcal{H}}$, $\forall Q_1, Q_2 \in \mathcal{H}$, a.s., can be established in a similar way to (51), but with μ_{Q_1} and μ_{Q_2}, whose definitions are provided below (2c), taking the place of μ in (51b) and (51c), respectively.

Consider now the fixed points $Q^{\diamond}_{\mu}$ and Q_μ of $T^{\diamond}_{\mu}$ and T_μ, respectively. Notice now that

$$\begin{aligned}
&\|Q^{\diamond}_{\mu} - Q_{\mu}\|_{\mathcal{H}}\\
&= \|T^{\diamond}_{\mu}(Q^{\diamond}_{\mu}) - T_{\mu}(Q_{\mu})\|_{\mathcal{H}}\\
&\leq \|T^{\diamond}_{\mu}(Q^{\diamond}_{\mu}) - T_{\mu}(Q^{\diamond}_{\mu})\|_{\mathcal{H}} + \|T_{\mu}(Q^{\diamond}_{\mu}) - T_{\mu}(Q_{\mu})\|_{\mathcal{H}}\\
&\leq \alpha\|(\Sigma^{\mu}_{s'|z} - \hat{\Sigma}^{\mu}_{s'|z}(N))(Q^{\diamond}_{\mu})\|_{\mathcal{H}} + \beta\|Q^{\diamond}_{\mu} - Q_{\mu}\|_{\mathcal{H}}\\
&\leq \alpha\|\Sigma^{\mu}_{s'|z} - \hat{\Sigma}^{\mu}_{s'|z}(N)\|\,\|Q^{\diamond}_{\mu}\|_{\mathcal{H}} + \beta_\infty\|Q^{\diamond}_{\mu} - Q_{\mu}\|_{\mathcal{H}}\,, \quad (52)
\end{aligned}$$

which yields

$$\|Q^{\diamond}_{\mu} - Q_{\mu}(N)\|_{\mathcal{H}} \leq \frac{\alpha\|Q^{\diamond}_{\mu}\|_{\mathcal{H}}}{1-\beta_\infty}\|\Sigma^{\mu}_{s'|z} - \hat{\Sigma}^{\mu}_{s'|z}(N)\|\,,$$

that establishes in turn (21a) by Theorem 12.

Notice now that $Q^{\diamond}_{*} - Q_* = T^{\diamond}(Q^{\diamond}_{*}) - T(Q_*) = T^{\diamond}(Q^{\diamond}_{*}) - T(Q^{\diamond}_{*}) + T(Q^{\diamond}_{*}) - T(Q_*)$, that

$$T^{\diamond}(Q^{\diamond}_{*}) - T(Q^{\diamond}_{*}) = (\Sigma^{\mu_{Q^{\diamond}_{*}}}_{s'|z} - \hat{\Sigma}^{\mu_{Q^{\diamond}_{*}}}_{s'|z}(N))(Q^{\diamond}_{*})\,,$$

and follow steps like those in (52) to establish (21b) by Theorem 12.

Appendix D
Proof of Theorem 6

The following discussion is performed for any ω chosen arbitrarily from $E^{(\epsilon)}$ of Assumption 5(i), after possibly excluding from $E^{(\epsilon)}$ the union of zero-probability events which appear via the qualifier "a.s." in Assumptions 5(iii) to 5(vi). By the definition of $E^{(\epsilon)}$, $\omega \in E^{(\epsilon)}$ implies that there exists a sufficiently large n_0 s.t. for any $n \geq n_0$, there exists a sufficiently large $N[n]$ with $\omega \in E^{(\epsilon)}_{n,N[n]}$.

By Assumptions 3, $T^{\diamond}_{\mu_n}$ is β_∞-Lipschitz continuous, and thus a contraction for all sufficiently large n by Assumption 5(iii) and the discussion after (51). Recall then the Banach-Picard fixed-point theorem [31], which guarantees that $\forall Q \in \mathcal{H}$, $\lim_{K\to\infty}(T^{\diamond}_{\mu_{n+1}})^K(Q) = Q^{\diamond}_{\mu_{n+1}}$, with $Q^{\diamond}_{\mu_{n+1}}$ being the unique fixed point of $T^{\diamond}_{\mu_{n+1}}$.

Lemma 13. For all sufficiently large n, a.s.,

$$\|Q^{\diamond}_{\mu_{n+1}} - T^{\diamond}_{\mu_{n+1}}(Q^{\diamond}_{\mu_n})\|_{\mathcal{H}} \leq \frac{\Delta_2}{1-\beta_\infty}\,.$$

Proof: For any $k \in \mathbb{N}_*$,

$$\begin{aligned}
&\|(T^{\diamond}_{\mu_{n+1}})^k(Q^{\diamond}_{\mu_n}) - (T^{\diamond}_{\mu_{n+1}})^{k-1}(Q^{\diamond}_{\mu_n})\|_{\mathcal{H}}\\
&\leq \beta_\infty\|(T^{\diamond}_{\mu_{n+1}})^{k-1}(Q^{\diamond}_{\mu_n}) - (T^{\diamond}_{\mu_{n+1}})^{k-2}(Q^{\diamond}_{\mu_n})\|_{\mathcal{H}}\\
&\leq \beta^{k-1}_\infty\|(T^{\diamond}_{\mu_{n+1}} - \text{Id})(Q^{\diamond}_{\mu_n})\|_{\mathcal{H}} \leq \beta^{k-1}_\infty\Delta_2\,,
\end{aligned}$$

by Assumption 5(vi). Hence, for any $K \in \mathbb{N}_*$,

$$\begin{aligned}
&\|(T^{\diamond}_{\mu_{n+1}})^K(Q^{\diamond}_{\mu_n}) - T^{\diamond}_{\mu_{n+1}}(Q^{\diamond}_{\mu_n})\|_{\mathcal{H}}\\
&\leq \sum\nolimits_{k=1}^{K}\|(T^{\diamond}_{\mu_{n+1}})^k(Q^{\diamond}_{\mu_n}) - (T^{\diamond}_{\mu_{n+1}})^{k-1}(Q^{\diamond}_{\mu_n})\|_{\mathcal{H}}\\
&\leq \sum\nolimits_{k=1}^{K}\beta^{k-1}_\infty\Delta_2 \leq \Delta_2\sum\nolimits_{k=0}^{\infty}\beta^{k}_\infty = \tfrac{\Delta_2}{1-\beta_\infty}\,. \quad (53)
\end{aligned}$$

Since $\lim_{K\to\infty}(T^{\diamond}_{\mu_{n+1}})^K(Q^{\diamond}_{\mu_n}) = Q^{\diamond}_{\mu_{n+1}}$, the application of $\lim_{K\to\infty}$ to (53) establishes Lemma 13. ∎

Lemma 14. For all sufficiently large n,

$$\begin{aligned}
&\|Q^{\diamond}_{\mu_{n+1}} - Q^{\diamond}_{*}\|_{\mathcal{H}}\\
&\leq \beta_\infty\|Q^{\diamond}_{\mu_n} - Q^{\diamond}_{*}\|_{\mathcal{H}} + 2\beta_\infty(\Delta_0 + \epsilon)\\
&\quad + \Delta_1 + \frac{\Delta_2}{1-\beta_\infty}\,.
\end{aligned}$$

Proof: By Assumption 5(iii) and by following again the discussion after (51), it can be verified that $T^{\diamond}$ is a β_∞-contraction. Observe also by Assumptions 5 and Lemma 13 that

$$\begin{aligned}
&\|Q^{\diamond}_{\mu_{n+1}} - Q^{\diamond}_{*}\|_{\mathcal{H}}\\
&\leq \|T^{\diamond}(Q_n) - Q^{\diamond}_{*}\|_{\mathcal{H}} + \|Q^{\diamond}_{\mu_{n+1}} - T^{\diamond}(Q_n)\|_{\mathcal{H}}\\
&= \|T^{\diamond}(Q_n) - T^{\diamond}(Q^{\diamond}_{*})\|_{\mathcal{H}} + \|Q^{\diamond}_{\mu_{n+1}} - T^{\diamond}(Q_n)\|_{\mathcal{H}}\\
&\leq \beta_\infty\|Q_n - Q^{\diamond}_{*}\|_{\mathcal{H}} + \|Q^{\diamond}_{\mu_{n+1}} - T^{\diamond}_{\mu_{n+1}}(Q_n)\|_{\mathcal{H}}\\
&\quad + \|T^{\diamond}_{\mu_{n+1}}(Q_n) - T^{\diamond}(Q_n)\|_{\mathcal{H}}\\
&\leq \beta_\infty\|Q^{\diamond}_{\mu_n} - Q^{\diamond}_{*}\|_{\mathcal{H}} + \beta_\infty\|Q_n - Q_{\mu_n}(N[n])\|_{\mathcal{H}}\\
&\quad + \beta_\infty\|Q_{\mu_n}(N[n]) - Q^{\diamond}_{\mu_n}\|_{\mathcal{H}}\\
&\quad + \|Q^{\diamond}_{\mu_{n+1}} - T^{\diamond}_{\mu_{n+1}}(Q^{\diamond}_{\mu_n})\|_{\mathcal{H}}\\
&\quad + \|T^{\diamond}_{\mu_{n+1}}(Q^{\diamond}_{\mu_n}) - T^{\diamond}_{\mu_{n+1}}(Q_n)\|_{\mathcal{H}}\\
&\quad + \|T^{\diamond}_{\mu_{n+1}}(Q_n) - T^{\diamond}(Q_n)\|_{\mathcal{H}}\\
&\leq \beta_\infty\|Q^{\diamond}_{\mu_n} - Q^{\diamond}_{*}\|_{\mathcal{H}} + \beta_\infty(\Delta_0 + \epsilon)\\
&\quad + \tfrac{\Delta_2}{1-\beta_\infty} + \beta_\infty\|Q^{\diamond}_{\mu_n} - Q_n\|_{\mathcal{H}} + \Delta_1\\
&\leq \beta_\infty\|Q^{\diamond}_{\mu_n} - Q^{\diamond}_{*}\|_{\mathcal{H}} + \beta_\infty(\Delta_0 + \epsilon)\\
&\quad + \tfrac{\Delta_2}{1-\beta_\infty} + \beta_\infty\|Q_{\mu_n}(N[n]) - Q_n\|_{\mathcal{H}}\\
&\quad + \beta_\infty\|Q^{\diamond}_{\mu_n} - Q_{\mu_n}(N[n])\|_{\mathcal{H}} + \Delta_1\\
&\leq \beta_\infty\|Q^{\diamond}_{\mu_n} - Q^{\diamond}_{*}\|_{\mathcal{H}} + 2\beta_\infty(\Delta_0 + \epsilon) + \Delta_1 + \tfrac{\Delta_2}{1-\beta_\infty}\,,
\end{aligned}$$

which establishes Lemma 14. ∎

For a sufficiently large n, the application of Lemma 14 recursively for K times yields

$$\begin{aligned}
&\|Q^\diamond_{\mu_{n+K}} - Q^\diamond_*\|_{\mathcal{H}} \\
&\le \beta_\infty^K \|Q^\diamond_{\mu_n} - Q^\diamond_*\|_{\mathcal{H}} \\
&\quad + \sum\nolimits_{k=0}^{K-1} \beta_\infty^k \left(2\beta_\infty(\Delta_0+\epsilon) + \Delta_1 + \tfrac{\Delta_2}{1-\beta_\infty}\right) \\
&\le \beta_\infty^K \|Q^\diamond_{\mu_n} - Q^\diamond_*\|_{\mathcal{H}} \\
&\quad + \underbrace{\tfrac{1}{1-\beta_\infty}\left(2\beta_\infty(\Delta_0+\epsilon) + \Delta_1 + \tfrac{\Delta_2}{1-\beta_\infty}\right)}_{\Delta'},
\end{aligned}$$

and because of $\beta_\infty < 1$,

$$\limsup_{n\to\infty} \|Q^\diamond_{\mu_n} - Q^\diamond_*\|_{\mathcal{H}} = \limsup_{K\to\infty} \|Q^\diamond_{\mu_{n+K}} - Q^\diamond_*\|_{\mathcal{H}} \le \Delta' .$$

Now, the triangle inequality suggests

$$\begin{aligned}
&\|Q_n - Q^\diamond_*\|_{\mathcal{H}} \\
&\le \|Q_n - Q_{\mu_n}(N[n])\|_{\mathcal{H}} + \|Q_{\mu_n}(N[n]) - Q^\diamond_{\mu_n}\|_{\mathcal{H}} \\
&\quad + \|Q^\diamond_{\mu_n} - Q^\diamond_*\|_{\mathcal{H}} \\
&\le \Delta_0 + \epsilon + \|Q^\diamond_{\mu_n} - Q^\diamond_*\|_{\mathcal{H}} ,
\end{aligned}$$

and an application of $\limsup_{n\to\infty}$ to the previous inequality yields $\limsup_{n\to\infty}\|Q_n - Q^\diamond_*\|_{\mathcal{H}} \le \Delta_0 + \epsilon + \Delta'$, which establishes Theorem 6.

Appendix E
Proof of Theorem 8

First, recall operators (46c) and (49). Define then

$$\xi_n \coloneqq (\hat{\Sigma}^{\mu_n}_{s'|z}(N) - \Sigma^{\mu_n}_{s'|z})^*(\varphi(\mathbf{z}_{n-1})) , \tag{54}$$

where superscript $*$ over a bounded linear operator denotes its adjoint [64].

The linear covariance operators $\Sigma^{(n)}_{zz}, \Sigma^{(n)}_{\xi z}, \Sigma^{(n)}_{\xi\xi} \colon \mathcal{H} \to \mathcal{H}$ are introduced next; $\forall Q \in \mathcal{H}$,

$$\Sigma^{(n)}_{zz}(Q) \coloneqq \mathbb{E}\{\langle \varphi(\mathbf{z}_n) \mid Q\rangle_{\mathcal{H}} \cdot \varphi(\mathbf{z}_n)\} , \tag{55a}$$

$$\begin{aligned}
\Sigma^{(n)}_{\xi z}(Q) &\coloneqq \mathbb{E}\{\langle \varphi(\mathbf{z}_{n-1}) \mid Q\rangle_{\mathcal{H}} \cdot \xi_n\} , \\
&= \mathbb{E}\{\langle \varphi(\mathbf{z}_{n-1}) \mid Q\rangle_{\mathcal{H}} \\
&\qquad \cdot (\hat{\Sigma}^{\mu_n}_{s'|z}(N) - \Sigma^{\mu_n}_{s'|z})^*(\varphi(\mathbf{z}_{n-1}))\} ,
\end{aligned} \tag{55b}$$

$$\begin{aligned}
\Sigma^{(n)}_{\xi\xi}(Q) &\coloneqq \mathbb{E}\{\langle \xi_n \mid Q\rangle_{\mathcal{H}} \cdot \xi_n\} \\
&= \mathbb{E}\{\langle \varphi(\mathbf{z}_{n-1}) \mid (\hat{\Sigma}^{\mu_n}_{s'|z}(N) - \Sigma^{\mu_n}_{s'|z})(Q)\rangle_{\mathcal{H}} \\
&\qquad \cdot (\hat{\Sigma}^{\mu_n}_{s'|z}(N) - \Sigma^{\mu_n}_{s'|z})^*(\varphi(\mathbf{z}_{n-1}))\} ,
\end{aligned} \tag{55c}$$

where expectations in (55) are considered along the lines of (46). It can be verified that $\Sigma^{(n)}_{zz}, \Sigma^{(n)}_{\xi\xi}$ are self adjoint, *i.e.*, $\Sigma^{(n)*}_{zz} = \Sigma^{(n)}_{zz}$ and $\Sigma^{(n)*}_{\xi\xi} = \Sigma^{(n)}_{\xi\xi}$.

Proposition 15. With $\mathcal{L}^{(n)}_{\mu_n}[\mathbf{z}_{n-1}](\cdot)$ defined in (30), its expected loss $G_{\mu_n}(\cdot) \coloneqq \mathbb{E}\{\mathcal{L}^{(n)}_{\mu_n}[\mathbf{z}_{n-1}](\cdot)\}$ takes the following form for all sufficiently large n:

$$\begin{aligned}
G_{\mu_n}(Q) = \tfrac{1}{2}\langle Q \mid \mathcal{A}_{\mu_n}(Q)\rangle_{\mathcal{H}} + \langle Q \mid \mathcal{B}_{\mu_n}(g)\rangle_{\mathcal{H}} \\
+ \tfrac{1}{2}\langle g \mid \Sigma_{zz}(g)\rangle_{\mathcal{H}} , \quad \forall Q \in \mathcal{H} ,
\end{aligned} \tag{56}$$

where the linear $\mathcal{A}_{\mu_n}, \mathcal{B}_{\mu_n} \colon \mathcal{H} \to \mathcal{H}$ are defined by

$$\begin{aligned}
\mathcal{A}_{\mu_n} &\coloneqq (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\Sigma_{zz}(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id}) + \alpha^2\Sigma_{\xi\xi} \\
&\quad + \alpha\Sigma_{\xi z}(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id}) + \alpha(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\Sigma^*_{\xi z} ,
\end{aligned} \tag{57a}$$

$$\mathcal{B}_{\mu_n} \coloneqq (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\Sigma_{zz} + \alpha\Sigma_{\xi z} . \tag{57b}$$

Proof: Loss (30) takes the following form: $\forall Q \in \mathcal{H}$,

$$\begin{aligned}
\mathcal{L}^{(n)}_{\mu_n}[\mathbf{z}_{n-1}](Q) &= \tfrac{1}{2}\langle T^{(n)}_{\mu_n}(Q) - Q \mid \varphi(\mathbf{z}_{n-1})\rangle^2_{\mathcal{H}} \\
&= \tfrac{1}{2}\Big[\langle (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}} \\
&\qquad + \alpha\langle \xi_n \mid Q\rangle_{\mathcal{H}} + \langle g \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\Big]^2 \\
&= \mathrm{term}_1 + \mathrm{term}_2 + \mathrm{term}_3 ,
\end{aligned} \tag{58}$$

where

$$\begin{aligned}
&\mathrm{term}_1 \\
&\coloneqq \tfrac{1}{2}\Big[\langle (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \varphi(\mathbf{z}_{n-1})\rangle^2_{\mathcal{H}} + \alpha^2\langle \xi_n \mid Q\rangle^2_{\mathcal{H}} \\
&\qquad + 2\alpha\langle (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}} \cdot \langle \xi_n \mid Q\rangle_{\mathcal{H}}\Big] , \\
&\mathrm{term}_2 \\
&\coloneqq \langle g \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}} \cdot \langle (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}} \\
&\quad + \alpha\langle g \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}} \cdot \langle \xi_n \mid Q\rangle_{\mathcal{H}} , \\
&\mathrm{term}_3 \coloneqq \tfrac{1}{2}\langle g \mid \varphi(\mathbf{z}_{n-1})\rangle^2_{\mathcal{H}} .
\end{aligned}$$

A closer look at term_1 via (46a) suggests that

$$\begin{aligned}
&\mathbb{E}\{\langle (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \varphi(\mathbf{z}_{n-1})\rangle^2_{\mathcal{H}}\} \\
&= \mathbb{E}\{\langle (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \langle (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}} \\
&\qquad \cdot \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\} \\
&= \langle (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \mathbb{E}\{\langle (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}} \\
&\qquad \cdot \varphi(\mathbf{z}_{n-1})\}\rangle_{\mathcal{H}} \\
&= \langle Q \mid (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\Sigma_{zz}(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q)\rangle_{\mathcal{H}} ,
\end{aligned}$$

that

$$\begin{aligned}
\alpha^2\,\mathbb{E}\{\langle \xi_n \mid Q\rangle^2_{\mathcal{H}}\} &= \alpha^2\langle Q \mid \mathbb{E}\{\langle \xi_n \mid Q\rangle_{\mathcal{H}}\xi_n\}\rangle_{\mathcal{H}} \\
&= \langle Q \mid \alpha^2\Sigma_{\xi\xi}(Q)\rangle_{\mathcal{H}} ,
\end{aligned}$$

and

$$\begin{aligned}
&2\alpha\,\mathbb{E}\{\langle (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\langle \xi_n \mid Q\rangle_{\mathcal{H}}\} \\
&= 2\alpha\langle Q \mid \mathbb{E}\{\langle (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\,\xi_n\}\rangle_{\mathcal{H}} \\
&= \alpha\langle Q \mid \Sigma_{\xi z}(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q)\rangle_{\mathcal{H}} \\
&\quad + \alpha\langle Q \mid \Sigma_{\xi z}(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q)\rangle_{\mathcal{H}} \\
&= \alpha\langle Q \mid \Sigma_{\xi z}(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q)\rangle_{\mathcal{H}} \\
&\quad + \alpha\langle Q \mid (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\Sigma^*_{\xi z}(Q)\rangle_{\mathcal{H}} \\
&= \langle Q \mid (\alpha\Sigma_{\xi z}(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id}) + \alpha(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\Sigma^*_{\xi z})(Q)\rangle_{\mathcal{H}} .
\end{aligned}$$

Hence, $\mathbb{E}\{\text{term}_1\} = (1/2)\langle Q \mid \mathcal{A}_{\mu_n}(Q)\rangle_{\mathcal{H}}$. Moreover,

$$\begin{aligned}
&\mathbb{E}\{\text{term}_2\}\\
&= \mathbb{E}\{\langle g \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}} \cdot \langle (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\\
&\qquad + \alpha\langle g \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}} \cdot \langle \xi_n \mid Q\rangle_{\mathcal{H}}\}\\
&= \langle (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \mathbb{E}\{\langle g \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\, \varphi(\mathbf{z}_{n-1})\}\rangle_{\mathcal{H}}\\
&\quad + \alpha\langle Q \mid \mathbb{E}\{\langle g \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\, \xi_n\}\rangle_{\mathcal{H}}\\
&= \langle Q \mid (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\Sigma_{zz}(g)\rangle_{\mathcal{H}} + \alpha\langle Q \mid \Sigma_{\xi z}(g)\rangle_{\mathcal{H}}\\
&= \langle Q \mid \mathcal{B}_{\mu_n}(g)\rangle_{\mathcal{H}}\,,
\end{aligned}$$

and finally,

$$\begin{aligned}
\mathbb{E}\{\text{term}_3\} &= \tfrac{1}{2}\,\mathbb{E}\{\langle g \mid \varphi(\mathbf{z}_{n-1})\rangle^2_{\mathcal{H}}\}\\
&= \tfrac{1}{2}\langle g \mid \mathbb{E}\{\langle g \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\varphi(\mathbf{z}_{n-1})\}\rangle_{\mathcal{H}}\\
&= \tfrac{1}{2}\langle g \mid \Sigma_{zz}(g)\rangle_{\mathcal{H}}\,,
\end{aligned}$$

which completes the proof of Proposition 15. ∎

It is worth noting here that by the adopted assumptions and the observation $\mathcal{A}_{\mu_n} = \mathcal{A}^*_{\mu_n}$, operator $\mathcal{A}_{\mu_n}$ turns out to be bounded linear and self-adjoint.

Lemma 16. The expected loss $G_{\mu_n}(\cdot)$ (56) is $\sigma_{\min}(\mathcal{A}_{\mu_n})$-strongly convex for all sufficiently large n.

Proof: Verify that $\forall Q_1, Q_2 \in \mathcal{H}$, $\forall \gamma \in (0,1)$,

$$\begin{aligned}
&\gamma G_{\mu_n}(Q_1) + (1-\gamma)G_{\mu_n}(Q_2) - G_{\mu_n}(\gamma Q_1 + (1-\gamma)Q_2)\\
&= \tfrac{1}{2}\gamma(1-\gamma)[\langle Q_1 \mid \mathcal{A}_{\mu_n}(Q_1)\rangle_{\mathcal{H}} + \langle Q_2 \mid \mathcal{A}_{\mu_n}(Q_2)\rangle_{\mathcal{H}}\\
&\qquad - 2\langle Q_1 \mid \mathcal{A}_{\mu_n}(Q_2)\rangle_{\mathcal{H}}]\\
&= \tfrac{1}{2}\gamma(1-\gamma)\langle Q_1 - Q_2 \mid \mathcal{A}_{\mu_n}(Q_1 - Q_2)\rangle_{\mathcal{H}}\\
&\geq \tfrac{1}{2}\gamma(1-\gamma)\,\sigma_{\min}(\mathcal{A}_{\mu_n})\,\|Q_1 - Q_2\|_{\mathcal{H}}\,,
\end{aligned}$$

and recall that $\sigma_{\min}(\mathcal{A}_{\mu_n})$ is assumed to be positive for all sufficiently large n. ∎

Given the assertion of Lemma 16, define the minimizer

$$\breve{Q}^{\diamond}_{\mu_n} \coloneqq \operatorname{argmin}_{Q\in\mathcal{H}} G_{\mu_n}(Q)\,, \tag{59}$$

which is well defined and unique because of the coercivity and strongly convexity of G_{μ_n} [31].

Lemma 17. For any $h_1, h_2 \in \mathcal{H}$ and any $Q \in \mathcal{H}$,

$$\nabla\left(\langle \cdot \mid \langle h_1 \mid \cdot\rangle_{\mathcal{H}}\, h_2\rangle_{\mathcal{H}}\right)(Q) = \langle Q \mid h_1\rangle_{\mathcal{H}}\, h_2 + \langle Q \mid h_2\rangle_{\mathcal{H}}\, h_1\,,$$

where ∇ stands for the Fréchet gradient [31].

Proof: Notice that for any $q \in \mathcal{H}$,

$$\begin{aligned}
&\langle Q+q \mid \langle h_1 \mid Q+q\rangle_{\mathcal{H}}\, h_2\rangle_{\mathcal{H}} - \langle Q \mid \langle h_1 \mid Q\rangle_{\mathcal{H}}\, h_2\rangle_{\mathcal{H}}\\
&= \langle q \mid \langle h_1 \mid Q\rangle_{\mathcal{H}}\, h_2\rangle_{\mathcal{H}} + \langle Q \mid \langle h_1 \mid q\rangle_{\mathcal{H}}\, h_2\rangle_{\mathcal{H}}\\
&\quad + \langle q \mid \langle h_1 \mid q\rangle_{\mathcal{H}}\, h_2\rangle_{\mathcal{H}}\\
&= \langle q \mid \langle Q \mid h_1\rangle_{\mathcal{H}}\, h_2 + \langle Q \mid h_2\rangle_{\mathcal{H}}\, h_1\rangle_{\mathcal{H}}\\
&\quad + \langle q \mid \langle h_1 \mid q\rangle_{\mathcal{H}}\, h_2\rangle_{\mathcal{H}}\,,
\end{aligned}$$

that $\lim_{0\neq\|q\|_{\mathcal{H}}\to 0}\langle q \mid \langle h_1 \mid q\rangle_{\mathcal{H}}\, h_2\rangle_{\mathcal{H}}/\|q\|_{\mathcal{H}} = 0$, and recall the definition of the Fréchet gradient [31] to establish Lemma 17. ∎

Lemma 18.

(i) For all sufficiently large n and for any $Q \in \mathcal{F}_{n-1}$, a.s.,

$$\nabla G_{\mu_n}(Q) = \mathbb{E}_{|\mathcal{F}_{n-1}}\{\nabla\mathcal{L}^{(n)}_{\mu_n}[\mathbf{z}_{n-1}](Q)\}\,,$$

where $\mathbb{E}_{|\mathcal{F}_{n-1}}\{\cdot\}$ stands for the conditional expectation, conditioned on the filtration $\mathcal{F}_{n-1}$.

(ii) ∇G_{μ_n} is $\|\mathcal{A}_{\mu_n}\|$-Lipschitz continuous.

Proof: Because of (56), $\nabla G_{\mu_n}(Q) = \mathcal{A}_{\mu_n}(Q) + \mathcal{B}_{\mu_n}(g)$. By (58), $\nabla\mathcal{L}^{(n)}_{\mu_n}[\mathbf{z}_{n-1}](Q) = \nabla\text{term}_1 + \nabla\text{term}_2$. Following the lines of the proof of Proposition 15, notice by Lemma 17 that

$$\begin{aligned}
&\mathbb{E}_{|\mathcal{F}_{n-1}}\{\nabla\langle(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \langle(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\\
&\qquad \cdot \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\}\\
&= \mathbb{E}_{|\mathcal{F}_{n-1}}\{\nabla\langle Q \mid \langle Q \mid (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\\
&\qquad \cdot (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\}\\
&= \mathbb{E}_{|\mathcal{F}_{n-1}}\{2\langle Q \mid (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\\
&\qquad \cdot \varphi(\mathbf{z}_{n-1})\}\\
&= 2(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\\
&\qquad \cdot \mathbb{E}_{|\mathcal{F}_{n-1}}\{\langle(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\varphi(\mathbf{z}_{n-1})\}\\
&= 2(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\\
&\qquad \cdot \mathbb{E}\{\langle(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\varphi(\mathbf{z}_{n-1})\}\\
&= 2(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\Sigma_{zz}(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q)\,,
\end{aligned} \tag{60}$$

where Assumption 7(iv) was used in (60). Furthermore,

$$\begin{aligned}
&\alpha^2\,\mathbb{E}_{|\mathcal{F}_{n-1}}\{\nabla\langle Q \mid \langle Q \mid \xi_n\rangle_{\mathcal{H}}\xi_n\rangle_{\mathcal{H}}\}\\
&= \alpha^2\,\mathbb{E}_{|\mathcal{F}_{n-1}}\{2\langle Q \mid \xi_n\rangle_{\mathcal{H}}\xi_n\}\\
&= \alpha^2\,\mathbb{E}\{2\langle Q \mid \xi_n\rangle_{\mathcal{H}}\xi_n\}\\
&= 2\alpha^2\Sigma_{\xi\xi}(Q)\,,
\end{aligned}$$

and

$$\begin{aligned}
&\alpha\,\mathbb{E}_{|\mathcal{F}_{n-1}}\{\nabla\langle Q \mid \langle(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\xi_n\rangle_{\mathcal{H}}\}\\
&= \alpha\,\mathbb{E}_{|\mathcal{F}_{n-1}}\{\nabla\langle Q \mid \langle Q \mid (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\xi_n\rangle_{\mathcal{H}}\}\\
&= \alpha\,\mathbb{E}_{|\mathcal{F}_{n-1}}\{\langle Q \mid (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*(\varphi(\mathbf{z}_{n-1}))\rangle_{\mathcal{H}}\xi_n\\
&\qquad + \langle Q \mid \xi_n\rangle_{\mathcal{H}}(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*(\varphi(\mathbf{z}_{n-1}))\}\\
&= \alpha\,\mathbb{E}_{|\mathcal{F}_{n-1}}\{\langle(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\xi_n\}\\
&\qquad + \alpha(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})\,\mathbb{E}_{|\mathcal{F}_{n-1}}\{\langle Q \mid \xi_n\rangle_{\mathcal{H}}\varphi(\mathbf{z}_{n-1})\}\\
&= \alpha\,\mathbb{E}\{\langle(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\xi_n\}\\
&\qquad + \alpha(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})\,\mathbb{E}\{\langle Q \mid \xi_n\rangle_{\mathcal{H}}\varphi(\mathbf{z}_{n-1})\}\\
&= \left(\alpha\Sigma_{\xi z}(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id}) + \alpha(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\Sigma^*_{\xi z}\right)(Q)\,,
\end{aligned}$$

where Assumption 7(iv) was used again as in (60) to replace conditional expectations by $\mathbb{E}\{\cdot\}$. The previous derivations suggest $\mathbb{E}_{|\mathcal{F}_{n-1}}\{\nabla\text{term}_1\} = \mathcal{A}_{\mu_n}(Q)$. Moreover,

$$\begin{aligned}
&\mathbb{E}_{|\mathcal{F}_{n-1}}\{\nabla\text{term}_2\}\\
&= \mathbb{E}_{|\mathcal{F}_{n-1}}\{\nabla\langle g \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\langle(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q) \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\\
&\qquad + \alpha\nabla\langle g \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\langle\xi_n \mid Q\rangle_{\mathcal{H}}\}\\
&= (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\,\mathbb{E}_{|\mathcal{F}_{n-1}}\{\langle g \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\varphi(\mathbf{z}_{n-1})\}\\
&\qquad + \alpha\,\mathbb{E}_{|\mathcal{F}_{n-1}}\{\langle g \mid \varphi(\mathbf{z}_{n-1})\rangle\xi_n\}\\
&= (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\,\mathbb{E}\{\langle g \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\varphi(\mathbf{z}_{n-1})\}\\
&\qquad + \alpha\,\mathbb{E}\{\langle g \mid \varphi(\mathbf{z}_{n-1})\rangle\xi_n\}\\
&= (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\Sigma_{zz}(g) + \alpha\Sigma_{\xi z}(g) = \mathcal{B}_{\mu_n}(g)\,.
\end{aligned}$$

Gathering all of the previous results, $\mathbb{E}_{|\mathcal{F}_{n-1}}\{\nabla\text{term}_1 + \nabla\text{term}_2\} = \mathcal{A}_{\mu_n}(Q)+\mathcal{B}_{\mu_n}(g) = \nabla G_{\mu_n}(Q)$, which establishes the proof of Lemma 18(i).

The proof of Lemma 18(ii) follows directly from the observation that $\forall Q_1, Q_2 \in \mathcal{H}$,

$$\begin{aligned}\|\nabla G_{\mu_n}(Q_1) - \nabla G_{\mu_n}(Q_2)\|_{\mathcal{H}} &= \|\mathcal{A}_{\mu_n}(Q_1) - \mathcal{A}_{\mu_n}(Q_2)\|_{\mathcal{H}} \\ &\leq \|\mathcal{A}_{\mu_n}\| \, \|Q_1 - Q_2\|_{\mathcal{H}} \,,\end{aligned}$$

where it can be also observed by (57) that

$$\begin{aligned}\|\mathcal{A}_{\mu_n}\| \leq \|\Sigma_{zz}\| \, \|\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id}\|^2 + \alpha^2\|\Sigma_{\xi\xi}\| \\ + 2\alpha\|\Sigma_{\xi z}\| \, \|\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id}\| \,.\end{aligned}$$

■

Lemma 19. There exist $c_1, c_2 \in \mathbb{R}_{++}$ s.t. for all sufficiently large n and for any $Q \in \mathcal{F}_{n-1}$, a.s.,

$$\mathbb{E}_{|\mathcal{F}_{n-1}}\{\|\nabla\mathcal{L}^{(n)}_{\mu_n}[\mathbf{z}_{n-1}](Q) - \nabla G(Q)\|^2_{\mathcal{H}}\} \leq c_1\|Q\|^2_{\mathcal{H}} + c_2 \,.$$

Proof: Because of Lemma 18(i),

$$\begin{aligned}&\mathbb{E}_{|\mathcal{F}_{n-1}}\{\|\nabla\mathcal{L}^{(n)}_{\mu_n}[\mathbf{z}_{n-1}](Q) - \nabla G(Q)\|^2_{\mathcal{H}}\} \\ &= \mathbb{E}_{|\mathcal{F}_{n-1}}\{\|\nabla\mathcal{L}^{(n)}_{\mu_n}[\mathbf{z}_{n-1}](Q)\|^2_{\mathcal{H}}\} - \|\nabla G(Q)\|^2_{\mathcal{H}} \\ &\leq \mathbb{E}_{|\mathcal{F}_{n-1}}\{\|\nabla\mathcal{L}^{(n)}_{\mu_n}[\mathbf{z}_{n-1}](Q)\|^2_{\mathcal{H}}\} \,.\end{aligned}$$

By following the proof of Lemma 18(i), it can be readily verified that $\nabla\mathcal{L}^{(n)}_{\mu_n}[\mathbf{z}_{n-1}](Q) = \mathcal{A}^{(n)}_{\mu_n}[\mathbf{z}_{n-1}](Q)+\mathcal{B}^{(n)}_{\mu_n}[\mathbf{z}_{n-1}](g)$, where the mappings

$$\begin{aligned}\mathcal{A}^{(n)}_{\mu_n}[\mathbf{z}_{n-1}](\cdot) &\coloneqq \langle\cdot \mid (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}} \\ &\qquad \cdot (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\varphi(\mathbf{z}_{n-1}) \\ &\quad + \alpha\langle\cdot \mid (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\xi_n \\ &\quad + \alpha\langle\cdot \mid \xi_n\rangle_{\mathcal{H}}(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\varphi(\mathbf{z}_{n-1}) \\ &\quad + \alpha^2\langle\cdot \mid \xi_n\rangle_{\mathcal{H}}\xi_n \,, \\ \mathcal{B}^{(n)}_{\mu_n}[\mathbf{z}_{n-1}](\cdot) &\coloneqq (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\langle\cdot \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\varphi(\mathbf{z}_{n-1}) \\ &\quad + \alpha\langle\cdot \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\xi_n \,.\end{aligned}$$

Notice now by Assumption 7(iii) that for any $Q \in \mathcal{H}$,

$$\begin{aligned}\|\Sigma^{\mu_n}_{s'|z}(Q)\|_{\mathcal{H}} &= \|\Sigma^{\mu_n}_{s'|z}(Q) - \Sigma^{\mu_n}_{s'|z}(0)\|_{\mathcal{H}} \\ &= \tfrac{1}{\alpha}\|T^{\diamond}_{\mu_n}(Q) - T^{\diamond}_{\mu_n}(0)\|_{\mathcal{H}} \\ &\leq \tfrac{1}{\alpha}\beta_\infty\|Q - 0\|_{\mathcal{H}} = \tfrac{1}{\alpha}\beta_\infty\|Q\|_{\mathcal{H}} \,,\end{aligned}$$

which suggests that $\|\Sigma^{\mu_n}_{s'|z}\| \leq \beta_\infty/\alpha$. This implies in turn $\|\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id}\|_{\mathcal{H}} \leq \alpha\|\Sigma^{\mu_n}_{s'|z}\| + \|\mathrm{Id}\| \leq \alpha\beta_\infty/\alpha + 1 \leq \beta_\infty + 1$. Moreover, the reproducing property of the kernel κ yields $\|\varphi(\mathbf{z}_{n-1})\|^2_{\mathcal{H}} = \kappa(\mathbf{z}_{n-1}, \mathbf{z}_{n-1}) \leq B_\kappa$. Notice also that

$$\begin{aligned}&\|\mathcal{A}^{(n)}_{\mu_n}[\mathbf{z}_{n-1}](Q)\|_{\mathcal{H}} \\ &\leq \Big[\|\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id}\|^2\|\varphi(\mathbf{z}_{n-1})\|^2_{\mathcal{H}} \\ &\quad + 2\alpha\|\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id}\| \, \|\varphi(\mathbf{z}_{n-1})\|_{\mathcal{H}} \, \|\xi_n\|_{\mathcal{H}} \\ &\quad + \alpha^2\|\xi_n\|^2_{\mathcal{H}}\Big] \|Q\|_{\mathcal{H}} \\ &\leq \Big[(\beta_\infty + 1)^2 B_\kappa + 2\alpha(\beta_\infty + 1)\sqrt{B_\kappa}\|\xi_n\|_{\mathcal{H}} \\ &\quad + \alpha^2\|\xi_n\|^2_{\mathcal{H}}\Big] \|Q\|_{\mathcal{H}} \,, \\ &\|\mathcal{B}^{(n)}_{\mu_n}[\mathbf{z}_{n-1}](g)\|_{\mathcal{H}} \\ &\leq \|\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id}\| \, \|\varphi(\mathbf{z}_{n-1})\|^2_{\mathcal{H}} + \alpha\|\xi_n\|_{\mathcal{H}} \, \|g\|_{\mathcal{H}} \\ &\leq (\beta_\infty + 1)B_\kappa + \alpha\|\xi_n\|_{\mathcal{H}} \, \|g\|_{\mathcal{H}} \,.\end{aligned}$$

Observe that function $(\cdot)^{4/i} \colon \mathbb{R} \to \mathbb{R}$ is convex $\forall i \in \{1, 2, 3\}$, and recall Jensen's inequality for conditional expectation [30] to verify that $(\mathbb{E}_{|\mathcal{F}_{n-1}}\{\|\xi_n\|^i_{\mathcal{H}}\})^{4/i} \leq \mathbb{E}_{|\mathcal{F}_{n-1}}\{\|\xi_n\|^{i\cdot 4/i}_{\mathcal{H}}\} = \mathbb{E}\{\|\xi_n\|^4_{\mathcal{H}}\} = \mathfrak{m}^{(4)}_\xi$ for all sufficiently large n, $\forall i \in \{1, 2, 3\}$. These arguments suggest that there exist $\{\varrho_i\}^4_{i=0} \subset \mathbb{R}_+$ s.t.

$$\begin{aligned}&\mathbb{E}_{|\mathcal{F}_{n-1}}\{\|\mathcal{A}^{(n)}_{\mu_n}[\mathbf{z}_{n-1}](Q)\|^2_{\mathcal{H}}\} \\ &\leq \mathbb{E}_{|\mathcal{F}_{n-1}}\Big\{\Big[(\beta_\infty + 1)^2 B_\kappa + 2\alpha(\beta_\infty + 1)\sqrt{B_\kappa}\|\xi_n\|_{\mathcal{H}} \\ &\qquad + \alpha^2\|\xi_n\|^2_{\mathcal{H}}\Big]^2\Big\}\|Q\|^2_{\mathcal{H}} \\ &= \mathbb{E}_{|\mathcal{F}_{n-1}}\left\{\sum\nolimits_{i=0}^4 \varrho_i\|\xi_n\|^i_{\mathcal{H}}\right\}\|Q\|^2_{\mathcal{H}} \\ &\leq \left[\sum\nolimits_{i=0}^4 \varrho_i \left(\mathbb{E}_{|\mathcal{F}_{n-1}}\left\{\|\xi_n\|^4_{\mathcal{H}}\right\}\right)^{i/4}\right]\|Q\|^2_{\mathcal{H}} \\ &= \underbrace{\left(\sum\nolimits_{i=0}^4 \varrho_i(\mathfrak{m}^{(4)}_\xi)^{i/4}\right)}_{c'_1}\|Q\|^2_{\mathcal{H}} \leq c'_1\|Q\|^2_{\mathcal{H}} \,, \\ &\mathbb{E}_{|\mathcal{F}_{n-1}}\{\|\mathcal{B}^{(n)}_{\mu_n}[\mathbf{z}_{n-1}](g)\|^2_{\mathcal{H}}\} \\ &\leq (\beta_\infty + 1)^2 B^2_\kappa \\ &\quad + 2\alpha(\beta_\infty + 1)B_\kappa\|g\|_{\mathcal{H}} \, \mathbb{E}_{|\mathcal{F}_{n-1}}\{\|\xi_n\|_{\mathcal{H}}\} \\ &\quad + \alpha^2\|g\|^2_{\mathcal{H}} \, \mathbb{E}_{|\mathcal{F}_{n-1}}\{\|\xi_n\|^2_{\mathcal{H}}\} \\ &\leq (\beta_\infty + 1)^2 B^2_\kappa \\ &\quad + 2\alpha(\beta_\infty + 1)B_\kappa\|g\|_{\mathcal{H}} \, (\mathbb{E}_{|\mathcal{F}_{n-1}}\{\|\xi_n\|^4_{\mathcal{H}}\})^{1/4} \\ &\quad + \alpha^2\|g\|^2_{\mathcal{H}} \, (\mathbb{E}_{|\mathcal{F}_{n-1}}\{\|\xi_n\|^4_{\mathcal{H}}\})^{2/4} \\ &= (\beta_\infty + 1)^2 B^2_\kappa + 2\alpha(\beta_\infty + 1)B_\kappa\|g\|_{\mathcal{H}} \, (\mathfrak{m}^{(4)}_\xi)^{1/4} \\ &\quad + \alpha^2\|g\|^2_{\mathcal{H}} \, (\mathfrak{m}^{(4)}_\xi)^{1/2} \\ &=: c'_2 \,.\end{aligned}$$

Consequently,

$$\begin{aligned}\mathbb{E}_{|\mathcal{F}_{n-1}}\{\|\nabla\mathcal{L}^{(n)}_{\mu_n}[\mathbf{z}_{n-1}](Q)\|^2_{\mathcal{H}}\} &\leq 2{c'_1}^2\|Q\|^2_{\mathcal{H}} + 2{c'_2}^2 \\ &= c_1\|Q\|^2_{\mathcal{H}} + c_2 \,,\end{aligned}$$

where $c_1 \coloneqq 2{c'_1}^2$ and $c_2 \coloneqq 2{c'_2}^2$. This completes the proof. ■

An inspection of [2, Lemma 3.1], under the light of Lemmata 16, 18 and 19, suggests that for any sufficiently small step size η, or more specifically, for any

$$\eta < \frac{2\sigma_{\min}(\mathcal{A}_{\mu_n})}{\sigma^2_{\max}(\mathcal{A}_{\mu_n}) + 2c_1}\,,$$

there exists $c_3 \in \mathbb{R}_{++}$ s.t.

$$\limsup_{n\to\infty} \mathbb{E}\{\|Q_n - \breve{Q}^{\diamond}_{\mu_n}\|^2_{\mathcal{H}}\} \leq c_3\eta\,. \tag{61}$$

Now, because $T^{\diamond}_{\mu_n}$ is a contraction, its fixed point $Q^{\diamond}_{\mu_n}$ is unique and satisfies $(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q^{\diamond}_{\mu_n}) = -g$. This implies that the linear $(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})$ is non-singular, because any Q in the null space of $(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})$ satisfies $(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q^{\diamond}_{\mu_n} - Q) = -g$, and thus $Q = 0$ [56, Thm. 2.6-10(a)]. Therefore, $(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})$ is also non-singular. Consequently, with $\sigma_{\min}(\cdot)$ defined by (47a), $\sigma_{\min}((\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})) > 0$, because otherwise, and by (47a), there would exist a sequence $(h_k)_{k\in\mathbb{N}} \subset \mathcal{H}$, with $\|h_k\|_{\mathcal{H}} = 1$, s.t. $\lim_{k\to\infty}\|(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})h_k\|^2_{\mathcal{H}} = 0 \Leftrightarrow \lim_{k\to\infty}(h'_k \coloneqq (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})h_k) = 0 \Leftrightarrow \lim_{k\to\infty}(h_k = (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^{-1}h'_k) = 0 \Rightarrow 0 = \lim_{k\to\infty}\|h_k\|_{\mathcal{H}} = 1$, which is absurd. Moreover, notice that because of Lemma 16, G_{μ_n} is coercive [31], and there exists thus $B_{\breve{q}} \in \mathbb{R}_{++}$ s.t. $\|\breve{Q}_{\mu_n}\|_{\mathcal{H}} \leq B_{\breve{q}}$, for all sufficiently large n, via Assumption 7(ii).

Notice that there exists $\mathfrak{m}^{(2)}_{\xi} \in \mathbb{R}_{++}$ s.t.

$$\begin{aligned}
\|\Sigma_{\xi\xi}\| &= \sup_{\|h\|_{\mathcal{H}}=1} \langle h \mid \mathbb{E}_{\xi_n}\{\langle h \mid \xi_n\rangle_{\mathcal{H}}\, \xi_n\}\rangle_{\mathcal{H}} \\
&= \sup_{\|h\|_{\mathcal{H}}=1} \mathbb{E}_{\xi_n}\{\langle h \mid \langle h \mid \xi_n\rangle_{\mathcal{H}}\, \xi_n\rangle_{\mathcal{H}}\} \\
&= \mathbb{E}\{\|\xi_n\|^2_{\mathcal{H}}\} =: \mathfrak{m}^{(2)}_{\xi}\,,
\end{aligned} \tag{62a}$$

and

$$\begin{aligned}
\|\Sigma_{\xi z}\| &= \sup_{\|h\|_{\mathcal{H}}=1} \langle h \mid \mathbb{E}_{\mathbf{z}_{n-1},\xi_n}\{\langle h \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\, \xi_n\}\rangle_{\mathcal{H}} \\
&= \sup_{\|h\|_{\mathcal{H}}=1} \mathbb{E}_{\mathbf{z}_{n-1},\xi_n}\{\langle h \mid \langle h \mid \varphi(\mathbf{z}_{n-1})\rangle_{\mathcal{H}}\, \xi_n\rangle_{\mathcal{H}}\} \\
&\leq \sup_{\|h\|_{\mathcal{H}}=1} \|h\|^2_{\mathcal{H}}\, \mathbb{E}_{\mathbf{z}_{n-1},\xi_n}\{\|\xi_n\|_{\mathcal{H}}\, \|\varphi(\mathbf{z}_{n-1})\|_{\mathcal{H}}\} \\
&\leq \sqrt{B_\kappa}\, \mathbb{E}\{\|\xi_n\|_{\mathcal{H}}\} \leq (B_\kappa \mathfrak{m}^{(2)}_{\xi})^{1/2}\,,
\end{aligned} \tag{62b}$$

where Jensen's inequality $\mathbb{E}\{\|\xi_n\|_{\mathcal{H}}\} \leq (\mathfrak{m}^{(2)}_{\xi})^{1/2}$ [30], propelled by the convexity of the function $(\cdot)^2$, was used in (62b). Moreover, define

$$\begin{aligned}
\tilde{\mathcal{A}}_{\mu_n} &\coloneqq \mathcal{A}_{\mu_n} - (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\Sigma_{zz}(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id}) \\
&= \alpha\Sigma_{\xi z}(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id}) + \alpha(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\Sigma^*_{\xi z} + \alpha^2\Sigma_{\xi\xi}\,, \\
\tilde{\mathcal{B}}_{\mu_n} &\coloneqq \mathcal{B}_{\mu_n} - (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\Sigma_{zz} = \alpha\Sigma_{\xi z}\,.
\end{aligned}$$

Via (62),

$$\begin{aligned}
\|\tilde{\mathcal{A}}_{\mu_n}\| &\leq 2\alpha(B_\kappa\mathfrak{m}^{(2)}_{\xi})^{1/2}\|(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})\| + \alpha^2\mathfrak{m}^{(2)}_{\xi} \\
&\leq 2\alpha(B_\kappa\mathfrak{m}^{(2)}_{\xi})^{1/2}(\beta_\infty + 1) + \alpha^2\mathfrak{m}^{(2)}_{\xi}\,, \\
\|\tilde{\mathcal{B}}_{\mu_n}\| &\leq \alpha(B_\kappa\mathfrak{m}^{(2)}_{\xi})^{1/2}\,.
\end{aligned}$$

Recall $(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q^{\diamond}_{\mu_n}) + g = 0$, and observe $\nabla G(\breve{Q}^{\diamond}_{\mu_n}) = 0$, because of the convexity of $G_{\mu_n}(\cdot)$. Hence,

$$\begin{aligned}
0 &= (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\Sigma_{zz}(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q^{\diamond}_{\mu_n}) \\
&\quad + (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\Sigma_{zz}(g)\,, \\
0 &= \mathcal{A}_{\mu_n}(\breve{Q}^{\diamond}_{\mu_n}) + \mathcal{B}_{\mu_n}(g)\,,
\end{aligned}$$

which lead in turn to

$$\begin{aligned}
0 &= \|(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\Sigma_{zz}(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q^{\diamond}_{\mu_n}) \\
&\quad + (\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\Sigma_{zz}(g) - [\mathcal{A}_{\mu_n}(\breve{Q}^{\diamond}_{\mu_n}) + \mathcal{B}(g)]\|_{\mathcal{H}} \\
&= \|(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\Sigma_{zz}(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q^{\diamond}_{\mu_n} - \breve{Q}^{\diamond}_{\mu_n}) \\
&\quad - [\tilde{\mathcal{A}}_{\mu_n}(\breve{Q}^{\diamond}_{\mu_n}) + \tilde{\mathcal{B}}_{\mu_n}(g)]\|_{\mathcal{H}} \\
&\geq \|(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*\Sigma_{zz}(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})(Q^{\diamond}_{\mu_n} - \breve{Q}^{\diamond}_{\mu_n})\|_{\mathcal{H}} \\
&\quad - \|\tilde{\mathcal{A}}_{\mu_n}(\breve{Q}^{\diamond}_{\mu_n}) + \tilde{\mathcal{B}}_{\mu_n}(g)\|_{\mathcal{H}} \\
&\geq \sigma_{\min}((\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id}))\, \sigma_{\min}(\Sigma_{zz}) \\
&\quad \cdot \|Q^{\diamond}_{\mu_n} - \breve{Q}^{\diamond}_{\mu_n}\|_{\mathcal{H}} - \|\tilde{\mathcal{A}}_{\mu_n}\|\, \|\breve{Q}^{\diamond}_{\mu_n}\|_{\mathcal{H}} - \|\tilde{\mathcal{B}}_{\mu_n}\|\, \|g\|_{\mathcal{H}}\,.
\end{aligned}$$

The previous discussion suggests that there exists $c_4 \in \mathbb{R}_{++}$ s.t.

$$\begin{aligned}
&\|Q^{\diamond}_{\mu_n} - \breve{Q}^{\diamond}_{\mu_n}\|_{\mathcal{H}} \\
&\leq \frac{\|\tilde{\mathcal{A}}_{\mu_n}\|\, \|\breve{Q}^{\diamond}_{\mu_n}\|_{\mathcal{H}} + \|\tilde{\mathcal{B}}_{\mu_n}\|\, \|g\|_{\mathcal{H}}}{\sigma_{\min}((\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id}))\, \sigma_{\min}(\Sigma_{zz})} \\
&\leq \frac{\left[2\alpha\sqrt{B_\kappa}(\beta_\infty + 1) + \alpha^2(\mathfrak{m}^{(2)}_{\xi})^{1/2}\right] B_{\breve{q}}}{\sigma_{\min}((\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id}))\, \sigma_{\min}(\Sigma_{zz})}(\mathfrak{m}^{(2)}_{\xi})^{1/2} \\
&\quad + \frac{\alpha\sqrt{B_\kappa}\|g\|_{\mathcal{H}}}{\sigma_{\min}((\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id})^*(\alpha\Sigma^{\mu_n}_{s'|z} - \mathrm{Id}))\, \sigma_{\min}(\Sigma_{zz})}(\mathfrak{m}^{(2)}_{\xi})^{1/2} \\
&\leq c_4(\mathfrak{m}^{(2)}_{\xi})^{1/2}\,.
\end{aligned}$$

Notice also that

$$\begin{aligned}
\mathfrak{m}^{(2)}_{\xi} &= \mathbb{E}\{\|\xi_n\|^2_{\mathcal{H}}\} \\
&= \mathbb{E}\{\langle(\hat{\Sigma}^{\mu_n}_{s'|z} - \Sigma^{\mu_n}_{s'|z})^*(\varphi(\mathbf{z}_{n-1})) \\
&\qquad \mid (\hat{\Sigma}^{\mu_n}_{s'|z} - \Sigma^{\mu_n}_{s'|z})^*(\varphi(\mathbf{z}_{n-1}))\rangle_{\mathcal{H}}\} \\
&= \mathbb{E}\{\langle\varphi(\mathbf{z}_{n-1}) \\
&\qquad \mid (\hat{\Sigma}^{\mu_n}_{s'|z} - \Sigma^{\mu_n}_{s'|z})(\hat{\Sigma}^{\mu_n}_{s'|z} - \Sigma^{\mu_n}_{s'|z})^*(\varphi(\mathbf{z}_{n-1}))\rangle_{\mathcal{H}}\} \\
&\leq \mathbb{E}\{\sigma_{\max}((\hat{\Sigma}^{\mu_n}_{s'|z} - \Sigma^{\mu_n}_{s'|z})(\hat{\Sigma}^{\mu_n}_{s'|z} - \Sigma^{\mu_n}_{s'|z})^*) \\
&\qquad \cdot \kappa(\mathbf{z}_{n-1}, \mathbf{z}_{n-1})\} \\
&\leq B_\kappa\, \mathbb{E}\{\|(\hat{\Sigma}^{\mu_n}_{s'|z} - \Sigma^{\mu_n}_{s'|z})(\hat{\Sigma}^{\mu_n}_{s'|z} - \Sigma^{\mu_n}_{s'|z})^*\|\} \\
&= B_\kappa\, \mathbb{E}\{\|\hat{\Sigma}^{\mu_n}_{s'|z}(N[n]) - \Sigma^{\mu_n}_{s'|z}\|^2\}\,,
\end{aligned} \tag{63}$$

so that

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{E}\{\|Q^{\diamond}_{\mu_n} - \breve{Q}^{\diamond}_{\mu_n}\|^2_{\mathcal{H}}\} \\
&= \limsup_{n\to\infty} \|Q^{\diamond}_{\mu_n} - \breve{Q}^{\diamond}_{\mu_n}\|^2_{\mathcal{H}} \\
&\leq c_4^2\mathfrak{m}^{(2)}_{\xi} \\
&\leq c_4^2 B_\kappa \limsup_{n\to\infty} \mathbb{E}\{\|\hat{\Sigma}^{\mu_n}_{s'|z}(N[n]) - \Sigma^{\mu_n}_{s'|z}\|^2\}\,.
\end{aligned} \tag{64}$$

Combine now (61) with (64) to obtain

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{E}\{\|Q_n - Q^\diamond_{\mu_n}\|^2_{\mathcal{H}}\} \\
&\le 2\limsup_{n\to\infty} \mathbb{E}\{\|Q_n - \breve{Q}^\diamond_{\mu_n}\|^2_{\mathcal{H}}\} \\
&\quad + 2\limsup_{n\to\infty} \mathbb{E}\{\|Q^\diamond_{\mu_n} - \breve{Q}^\diamond_{\mu_n}\|^2_{\mathcal{H}}\} \\
&\le 2c_3\eta + 2c_4^2 B_\kappa \limsup_{n\to\infty} \mathbb{E}\{\|\hat{\Sigma}^{\mu_n}_{s'|z}(N[n]) - \Sigma^{\mu_n}_{s'|z}\|^2\}\,,
\end{aligned}$$

and hence,

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{E}\{\|Q_n - Q^\diamond_*\|^2_{\mathcal{H}}\} \\
&\le 2\limsup_{n\to\infty} \mathbb{E}\{\|Q_n - Q^\diamond_{\mu_n}\|^2_{\mathcal{H}}\} \\
&\quad + 2\limsup_{n\to\infty} \mathbb{E}\{\|Q^\diamond_{\mu_n} - Q^\diamond_*\|^2_{\mathcal{H}}\} \\
&\le 4c_3\eta + 4c_4^2 B_\kappa \limsup_{n\to\infty} \mathbb{E}\{\|\hat{\Sigma}^{\mu_n}_{s'|z}(N[n]) - \Sigma^{\mu_n}_{s'|z}\|^2\} \\
&\quad + 2\limsup_{n\to\infty} \mathbb{E}\{\|Q^\diamond_{\mu_n} - Q^\diamond_*\|^2_{\mathcal{H}}\}\,,
\end{aligned}$$

which establishes Theorem 8 with $\Delta_4 := 4c_3$ and $\Delta_5 := 4c_4^2 B_\kappa$.

Appendix F
Proof of Theorem 10

Lemma 20. There exist $c_1 \in \mathbb{R}_{++}$ and $c_2 \in \mathbb{R}_+$ s.t. the following inequalities hold true for all $m \in \{n - M_{\text{av}} + 1, \ldots, n\}$ and for all sufficiently large n, a.s.,

$$\begin{aligned}
c_1(y_m - \boldsymbol{\theta}^\intercal_{n+1}\mathbf{x}_m)^2 + c_2 &\le \log|y_m - \boldsymbol{\theta}^\intercal_{n+1}\mathbf{x}_m|^2 \\
&\le 1 + (y_m - \boldsymbol{\theta}^\intercal_{n+1}\mathbf{x}_m)^2\,.
\end{aligned}$$

Proof: By the concavity of $\log(\cdot)$, $\log\varpi \le 1 + \varpi$, $\forall \varpi \in \mathbb{R}_{++}$. Hence, $\log|y_m - \boldsymbol{\theta}^\intercal_{n+1}\mathbf{x}_m|^2 \le 1 + (y_m - \boldsymbol{\theta}^\intercal_{n+1}\mathbf{x}_m)^2$. The concavity of $\log(\cdot)$ suggests also that $\forall \varpi \in (\Delta_6^2, \Delta_7^2)$,

$$\log\varpi \ge \frac{\log\Delta_7^2 - \log\Delta_6^2}{\Delta_7^2 - \Delta_6^2}(\varpi - \Delta_6^2) + \log\Delta_6^2 = c_1\varpi + c_2\,,$$

where $c_1 := (\log\Delta_7^2 - \log\Delta_6^2)/(\Delta_7^2 - \Delta_6^2)$ and $c_2 := \log\Delta_6^2 - \Delta_6^2(\log\Delta_7^2 - \log\Delta_6^2)/(\Delta_7^2 - \Delta_6^2)$. Substituting ϖ in the previous inequality by $|y_m - \boldsymbol{\theta}^\intercal_{n+1}\mathbf{x}_m|^2$ establishes Lemma 20. ∎

Recall now the data model in Section I-A to verify that $y_m - \boldsymbol{\theta}^\intercal_{n+1}\mathbf{x}_m = \boldsymbol{\theta}^\intercal_*\mathbf{x}_m + o_m - \boldsymbol{\theta}^\intercal_{n+1}\mathbf{x}_m = (\boldsymbol{\theta}_* - \boldsymbol{\theta}_{n+1})^\intercal\mathbf{x}_m + o_m$. Moreover, recall (23a) and (29) to verify that the chosen one-step loss satisfies

$$\begin{aligned}
g(\mathbf{s}_n, \mu(\mathbf{s}_n)) &= g(\mathbf{s}_n, a_n) = g(\mathbf{z}_n) \\
&= \tfrac{1}{M_{\text{av}}} \sum_{m=n-M_{\text{av}}+1}^{n} \log\frac{|y_m - \boldsymbol{\theta}^\intercal_{n+1}\mathbf{x}_m|^2}{\|\mathbf{x}_m\|_2^2}\,.
\end{aligned}$$

Observe then via Lemma 20 that

$$\begin{aligned}
&M_{\text{av}} + \sum_{m=n-M_{\text{av}}+1}^{n} (y_m - \boldsymbol{\theta}^\intercal_{n+1}\mathbf{x}_m)^2 \\
&\quad - \sum_{m=n-M_{\text{av}}+1}^{n} \log\|\mathbf{x}_m\|_2^2 \qquad (65a) \\
&\ge \sum_{m=n-M_{\text{av}}+1}^{n} \log|y_m - \boldsymbol{\theta}^\intercal_{n+1}\mathbf{x}_m|^2 \\
&\quad - \sum_{m=n-M_{\text{av}}+1}^{n} \log\|\mathbf{x}_m\|_2^2 \\
&= \sum_{m=n-M_{\text{av}}+1}^{n} \log\frac{|y_m - \boldsymbol{\theta}^\intercal_{n+1}\mathbf{x}_m|^2}{\|\mathbf{x}_m\|_2^2} \\
&= M_{\text{av}} g(\mathbf{s}_n, \mu(\mathbf{s}_n)) \qquad (65b) \\
&\ge c_1 \sum_{m=n-M_{\text{av}}+1}^{n} [(\boldsymbol{\theta}_* - \boldsymbol{\theta}_{n+1})^\intercal\mathbf{x}_m + o_m]^2 \\
&\quad + M_{\text{av}}c_2 - \sum_{m=n-M_{\text{av}}+1}^{n} \log\|\mathbf{x}_m\|_2^2 \\
&= c_1 \sum_{m=n-M_{\text{av}}+1}^{n} (\boldsymbol{\theta}_* - \boldsymbol{\theta}_{n+1})^\intercal\mathbf{x}_m\mathbf{x}_m^\intercal(\boldsymbol{\theta}_* - \boldsymbol{\theta}_{n+1}) \\
&\quad + 2c_1 \sum_{m=n-M_{\text{av}}+1}^{n} (\boldsymbol{\theta}_* - \boldsymbol{\theta}_{n+1})^\intercal\mathbf{x}_m o_m \\
&\quad + c_1 \sum_{m=n-M_{\text{av}}+1}^{n} o_m^2 + M_{\text{av}}c_2 \\
&\quad - \sum_{m=n-M_{\text{av}}+1}^{n} \log\|\mathbf{x}_m\|_2^2\,. \qquad (65c)
\end{aligned}$$

Notice also by Assumptions 9(v) and 9(vii) that there exists $c_3 \in \mathbb{R}$ s.t. $\mathbb{E}_{|\boldsymbol{\theta}_{n+1}}\{\log\|\mathbf{x}_m\|_2^2\} = \mathbb{E}\{\log\|\mathbf{x}_m\|_2^2\} \ge c_3$, for all sufficiently large m, a.s. Observe also that $\mathbb{E}\{\|\mathbf{x}_m\|_2^2\} = \mathbb{E}\{\text{trace}(\mathbf{x}_m\mathbf{x}_m^\intercal)\} = \text{trace}(\mathbb{E}\{\mathbf{x}_m\mathbf{x}_m^\intercal\}) = \text{trace}(\Sigma_{xx})$. Hence, by (65), a.s.,

$$\begin{aligned}
&M_{\text{av}} + \sum_{m=n-M_{\text{av}}+1}^{n} \mathbb{E}_{|\boldsymbol{\theta}_{n+1}}\{(y_m - \boldsymbol{\theta}^\intercal_{n+1}\mathbf{x}_m)^2\} \\
&\quad - M_{\text{av}}c_3 \\
&\ge M_{\text{av}} + \sum_{m=n-M_{\text{av}}+1}^{n} \mathbb{E}_{|\boldsymbol{\theta}_{n+1}}\{(y_m - \boldsymbol{\theta}^\intercal_{n+1}\mathbf{x}_m)^2\} \\
&\quad - \sum_{m=n-M_{\text{av}}+1}^{n} \mathbb{E}_{|\boldsymbol{\theta}_{n+1}}\{\log\|\mathbf{x}_m\|_2^2\} \\
&\ge M_{\text{av}}\,\mathbb{E}_{|\boldsymbol{\theta}_{n+1}}\{g(\mathbf{s}_n, \mu(\mathbf{s}_n))\} \\
&\ge c_1 \sum_{m=n-M_{\text{av}}+1}^{n} (\boldsymbol{\theta}_* - \boldsymbol{\theta}_{n+1})^\intercal\,\mathbb{E}_{|\boldsymbol{\theta}_{n+1}}\{\mathbf{x}_m\mathbf{x}_m^\intercal\}(\boldsymbol{\theta}_* - \boldsymbol{\theta}_{n+1}) \\
&\quad + 2c_1 \sum_{m=n-M_{\text{av}}+1}^{n} (\boldsymbol{\theta}_* - \boldsymbol{\theta}_{n+1})^\intercal\,\mathbb{E}_{|\boldsymbol{\theta}_{n+1}}\{\mathbf{x}_m o_m\} \\
&\quad + c_1 \sum_{m=n-M_{\text{av}}+1}^{n} \mathbb{E}_{|\boldsymbol{\theta}_{n+1}}\{o_m^2\} + M_{\text{av}}c_2 \\
&\quad - \sum_{m=n-M_{\text{av}}+1}^{n} \mathbb{E}_{|\boldsymbol{\theta}_{n+1}}\{\log\|\mathbf{x}_m\|_2^2\} \\
&\ge c_1 \sum_{m=n-M_{\text{av}}+1}^{n} (\boldsymbol{\theta}_* - \boldsymbol{\theta}_{n+1})^\intercal\,\mathbb{E}\{\mathbf{x}_m\mathbf{x}_m^\intercal\}(\boldsymbol{\theta}_* - \boldsymbol{\theta}_{n+1}) \\
&\quad + 2c_1 \sum_{m=n-M_{\text{av}}+1}^{n} (\boldsymbol{\theta}_* - \boldsymbol{\theta}_{n+1})^\intercal\,\mathbb{E}\{\mathbf{x}_m\}\,\mathbb{E}\{o_m\} \\
&\quad + c_1 \sum_{m=n-M_{\text{av}}+1}^{n} \mathbb{E}\{o_m^2\} + M_{\text{av}}c_2 \\
&\quad - \sum_{m=n-M_{\text{av}}+1}^{n} \log\mathbb{E}_{|\boldsymbol{\theta}_{n+1}}\{\|\mathbf{x}_m\|_2^2\} \\
&\ge c_1\lambda_{\min}(\Sigma_{xx}) \sum_{m=n-M_{\text{av}}+1}^{n} \|\boldsymbol{\theta}_* - \boldsymbol{\theta}_{n+1}\|_2^2 \\
&\quad + c_1 M_{\text{av}}\sigma_o^2 + M_{\text{av}}c_2 \\
&\quad - \sum_{m=n-M_{\text{av}}+1}^{n} \log\mathbb{E}\{\|\mathbf{x}_m\|_2^2\} \\
&= c_1 M_{\text{av}}\lambda_{\min}(\Sigma_{xx})\|\boldsymbol{\theta}_* - \boldsymbol{\theta}_{n+1}\|_2^2 + c_1 M_{\text{av}}\sigma_o^2 \\
&\quad + M_{\text{av}}c_2 - M_{\text{av}}\log\text{trace}(\Sigma_{xx})\,,
\end{aligned}$$

which yield in turn

$$\begin{aligned}
&M_{\text{av}}(1-c_3)+M_{\text{av}}\Delta_7^2\\
&\geq M_{\text{av}}+\sum\nolimits_{m=n-M_{\text{av}}+1}^{n}\mathbb{E}\{(y_m-\boldsymbol{\theta}_{n+1}^{\intercal}\mathbf{x}_m)^2\}\\
&\geq M_{\text{av}}\,\mathbb{E}\{g(\mathbf{s}_n,\mu(\mathbf{s}_n))\}\\
&\geq c_1 M_{\text{av}}\lambda_{\min}(\Sigma_{xx})\,\mathbb{E}\{\|\boldsymbol{\theta}_*-\boldsymbol{\theta}_{n+1}\|_2^2\}+c_1M_{\text{av}}\sigma_o^2\\
&\quad+M_{\text{av}}c_2-M_{\text{av}}\log\operatorname{trace}(\Sigma_{xx})\,,
\end{aligned}$$

or equivalently,

$$\begin{aligned}
&1-c_3+\Delta_7^2 &\text{(66a)}\\
&\geq\ \mathbb{E}\{g(\mathbf{s}_n,\mu(\mathbf{s}_n))\}\\
&\geq c_1\lambda_{\min}(\Sigma_{xx})\,\mathbb{E}\{\|\boldsymbol{\theta}_*-\boldsymbol{\theta}_{n+1}\|_2^2\}+c_1\sigma_o^2\\
&\quad+c_2-\log\operatorname{trace}(\Sigma_{xx})\,. &\text{(66b)}
\end{aligned}$$

Recall now that $Q_\mu^\diamond=T_\mu^\diamond(Q_\mu^\diamond)$. Hence, by (2a), $\forall n\geq n_0$,

$$\begin{aligned}
&Q_\mu^\diamond(\mathbf{s}_n,\mu(\mathbf{s}_n))\\
&=T_\mu^\diamond(Q_\mu^\diamond)(\mathbf{s}_n,\mu(\mathbf{s}_n))\\
&=g(\mathbf{s}_n,\mu(\mathbf{s}_n))+\alpha\,\mathbb{E}_{\mathbf{s}_{n+1}|\mathbf{s}_n}\{Q_\mu^\diamond(\mathbf{s}_{n+1},\mu(\mathbf{s}_{n+1}))\}\,.
\end{aligned}$$

It can be directly verified by this last recursion and induction that for any $K\in\mathbb{N}_*$, a.s.,

$$\begin{aligned}
&Q_\mu^\diamond(\mathbf{s}_{n_0},\mu(\mathbf{s}_{n_0}))\\
&=g(\mathbf{s}_{n_0},\mu(\mathbf{s}_{n_0}))+\sum\nolimits_{\nu=n_0+1}^{n_0+K-1}\alpha^{\nu-n_0}\,\mathbb{E}_{\mathbf{s}_\nu|\mathbf{s}_{n_0}}\{g(\mathbf{s}_\nu,\mu(\mathbf{s}_\nu))\}\\
&\quad+\alpha^K\,\mathbb{E}_{\mathbf{s}_{n_0+K}|\mathbf{s}_{n_0}}\{Q_\mu^\diamond(\mathbf{s}_{n_0+K},\mu(\mathbf{s}_{n_0+K}))\}\,,
\end{aligned}$$

and hence,

$$\begin{aligned}
\mathbb{E}\{Q_\mu^\diamond(\mathbf{s}_{n_0},\mu(\mathbf{s}_{n_0}))\}=&\sum\nolimits_{\nu=n_0}^{n_0+K-1}\alpha^{\nu-n_0}\,\mathbb{E}\{g(\mathbf{s}_\nu,\mu(\mathbf{s}_\nu))\}\\
&+\alpha^K\,\mathbb{E}\{Q_\mu^\diamond(\mathbf{s}_{n_0+K},\mu(\mathbf{s}_{n_0+K}))\}\,.
\end{aligned}\tag{67}$$

By (66a), $\forall K\in\mathbb{N}_*$,

$$\begin{aligned}
&\sum\nolimits_{\nu=n_0}^{n_0+K-1}\alpha^{\nu-n_0}\,\mathbb{E}\{g(\mathbf{s}_\nu,\mu(\mathbf{s}_\nu))\}\\
&\leq(1-c_3+\Delta_7^2)\sum\nolimits_{\nu=n_0}^{n_0+K-1}\alpha^{\nu-n_0}\\
&\leq(1-c_3+\Delta_7^2)\sum\nolimits_{\nu=n_0}^{+\infty}\alpha^{\nu-n_0}=\frac{1}{1-\alpha}(1-c_3+\Delta_7^2)\,,
\end{aligned}$$

so that $\sum_{\nu=n_0}^{+\infty}\alpha^{\nu-n_0}\,\mathbb{E}\{g(\mathbf{s}_\nu,\mu(\mathbf{s}_\nu))\}<+\infty$. Thus, by applying $\limsup_{K\to\infty}$ to (67) and by recalling Assumption 9(viii),

$$\mathbb{E}\{Q_\mu^\diamond(\mathbf{s}_{n_0},\mu(\mathbf{s}_{n_0}))\}=\sum\nolimits_{n=n_0}^{+\infty}\alpha^{n-n_0}\,\mathbb{E}\{g(\mathbf{s}_n,\mu(\mathbf{s}_n))\}\,.$$

Hence, by (66b),

$$\begin{aligned}
&\mathbb{E}\{Q_\mu^\diamond(\mathbf{s}_{n_0},\mu(\mathbf{s}_{n_0}))\}=\sum\nolimits_{n=n_0}^{+\infty}\alpha^{n-n_0}\,\mathbb{E}\{g(\mathbf{s}_n,\mu(\mathbf{s}_n))\}\\
&\geq c_1\lambda_{\min}(\Sigma_{xx})\sum\nolimits_{n=n_0}^{+\infty}\alpha^{n-n_0}\,\mathbb{E}\{\|\boldsymbol{\theta}_*-\boldsymbol{\theta}_{n+1}\|_2^2\}\\
&\quad+\sum\nolimits_{n=n_0}^{+\infty}\alpha^{n-n_0}[c_1\sigma_o^2+c_2-\log\operatorname{trace}(\Sigma_{xx})]\\
&=c_1\lambda_{\min}(\Sigma_{xx})\frac{1}{\alpha^{n_0+1}}\sum\nolimits_{n=n_0+1}^{+\infty}\alpha^n\,\mathbb{E}\{\|\boldsymbol{\theta}_*-\boldsymbol{\theta}_n\|_2^2\}\\
&\quad+\frac{1}{1-\alpha}[c_1\sigma_o^2+c_2-\log\operatorname{trace}(\Sigma_{xx})]\,,
\end{aligned}$$

which yields

$$\begin{aligned}
&\sum\nolimits_{n=n_0+1}^{\infty}\alpha^n\,\mathbb{E}\{\|\boldsymbol{\theta}_*-\boldsymbol{\theta}_n(\mu(\mathbf{s}_{n-1}))\|_2^2\}\\
&\leq\frac{\alpha^{n_0+1}}{c_1\lambda_{\min}(\Sigma_{xx})}\,\mathbb{E}\{Q_\mu^\diamond(\mathbf{s}_{n_0},\mu(\mathbf{s}_{n_0}))\}\\
&\quad+\frac{1}{c_1\lambda_{\min}(\Sigma_{xx})}\cdot\frac{\alpha^{n_0+1}}{1-\alpha}[\log\operatorname{trace}(\Sigma_{xx})-c_1\sigma_o^2-c_2]\,.
\end{aligned}$$

Observe that the way to update $\boldsymbol{\theta}_n$, $\forall n\geq n_0$, was not specified throughout the previous analysis. As such, set $\boldsymbol{\theta}_n\coloneqq\boldsymbol{\theta}_{n_0}$, $\forall n\geq n_0$, $\Delta_8\coloneqq c_1$, $\Delta_9\coloneqq c_2$ in the last inequality, and finally substitute n_0 by n to establish Theorem 10.