

Manifold Regularization for Kernelized LSTD

Xinyan Yan
xinyanyan@google.com

Krzysztof Choromanski
kchoro@google.com

Byron Boots
bboots@google.com

Vikas Sindhwani
sindhwani@google.com

Abstract: Policy evaluation or value function or Q-function approximation is a key procedure in reinforcement learning (RL). It is a necessary component of policy iteration and can be used for variance reduction in policy gradient methods. Therefore its quality has a significant impact on most RL algorithms. Motivated by manifold regularized learning, we propose a novel kernelized policy evaluation method that takes advantage of the intrinsic geometry of the state space learned from data, in order to achieve better sample efficiency and higher accuracy in Q-function approximation. Applying the proposed method in the Least-Squares Policy Iteration (LSPI) framework, we observe superior performance compared to widely used parametric basis functions on two standard benchmarks in terms of policy quality.

Keywords: Reinforcement Learning, Manifold Learning, Policy Evaluation, Policy Iteration

1 Problem description

We consider discrete-time infinite horizon discounted Markov Decision Process (MDP), with states $s \in \mathcal{S}$, admissible actions $a \in \mathcal{A}$, discount factor $\gamma \in (0, 1)$, reward function $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbf{R}$, and transition probability density $p(s'|s, a)$. Solving the MDP means finding a policy that maximizes the accumulated discounted rewards with states s and actions a distribution induced by applying that policy:

$$\text{maximize}_{\pi} \mathbf{E}_{s, a \sim p_{\pi}} \sum_{i=0}^{\infty} \gamma^i r(s_i, a_i). \quad (1)$$

Policy evaluation is a key component of policy iteration and policy gradient approaches to (1), with the objective of finding the Q-function¹ $Q^{\pi} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbf{R}$ associated with a stationary deterministic policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$, which is the fixed point of the Bellman operator T^{π} or the solution of the Bellman equation:

$$Q(s, a) = T^{\pi} Q(s, a) := r(s, a) + \mathbf{E}_{s'|s, a} \gamma Q(s', \pi(s')), \quad \forall s, a. \quad (2)$$

When the state space is large or infinite, solving (2) exactly becomes intractable. Least-Squares Temporal Difference (LSTD) is a widely used simulation-based algorithm for approximately solving the projected Bellman equation with linear representation of value function [1]. In order to apply LSTD to policy iteration, Lagoudakis and Parr [2] proposed a Q-function extension, and showed that the resulting policy iteration algorithm can be successfully applied to control problems.

In this work, we study LSTD-based approximate policy evaluation methods that benefit from manifold regularized learning, with the intuition that the Q-function is smooth on the manifold where states lie and not necessary in the ambient Euclidean space. Such manifold structure naturally arises in many robotics tasks due to constraints in the state space. For example, contact, *e.g.*, between foot and the ground or a hand and an object [3], restricts feasible states to lie along a manifold or a union

¹Throughout this work, we consider Q-function, instead of state value function, in order to cover the model-free setting.

of manifolds. Other examples include the cases when state variables belong to Special Euclidean group SE(3) to encode 3D poses, and obstacle avoidance settings where geodesic paths between state vectors are naturally better motivated than geometrically infeasible straight-line paths in the ambient space.

We provide a brief introduction to manifold regularized learning and a kernelized LSTD approach in §2. Our proposed method is detailed in §3 and experimental results are presented in §4. Finally §5 and §6 conclude the paper with related and future work.

2 Background

2.1 Manifold regularized learning

Manifold regularization has previously been studied in semi-supervised learning [4]. This data-dependent regularization exploits the geometry of the input space, thus achieving better generalization error. Given a labeled dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, Laplacian Regularized Least-Squares method (LapRLS) [4], finds a function f in a reproducing kernel Hilbert space \mathcal{H} that fits the data with proper complexity for generalization:

$$\text{minimize}_{f \in \mathcal{H}} \quad \frac{1}{n} \|Y - f(X)\|_2^2 + \lambda_f \|f\|_{\mathcal{H}}^2 + \frac{1}{n^2} \lambda_{\mathcal{M}} \|f\|_{\mathcal{M}}^2, \quad (3)$$

where data matrices $X = [x_1 \ \dots \ x_n]^T$ and $Y = [y_1 \ \dots \ y_n]^T$ with $x_i \in \mathbf{R}^d$ and $y_i \in \mathbf{R}$, $\|\cdot\|_{\mathcal{H}}$ is the norm in \mathcal{H} and $\|\cdot\|_{\mathcal{M}}$ is a penalty that reflects the intrinsic structure of P_X (see §2.2 in [4] for choices of $\|\cdot\|_{\mathcal{M}}$), and scalars $\lambda_f, \lambda_{\mathcal{M}}$ are regularization parameters². A natural choice of $\|\cdot\|_{\mathcal{M}}$ is $\|f\|_{\mathcal{M}}^2 = \int_{x \in \mathcal{M}} \|\nabla_{\mathcal{M}} f\|^2 dP_X(x)$, where \mathcal{M} is the support of P_X which is assumed to be a compact manifold [4] and $\nabla_{\mathcal{M}}$ is the gradient along \mathcal{M} . When P_X is unknown as in most learning settings, $\|\cdot\|_{\mathcal{M}}$ can be estimated empirically, and the optimization problem (3) becomes

$$\text{minimize}_{f \in \mathcal{H}} \quad \frac{1}{n} \|Y - f(X)\|_2^2 + \lambda_f \|f\|_{\mathcal{H}}^2 + \frac{1}{n^2} \lambda_{\mathcal{M}} \|f(X)\|_L^2, \quad (4)$$

where $\|x\|_L^2 = x^T L x$ and matrix $L \in \mathbf{S}_+^n$ is a graph Laplacian³ (different ways of constructing graph Laplacian from data can be found in [5]). Note that problem (4) is still convex, since graph Laplacian L is positive semidefinite, and the multiplicity of eigenvalue 0 is the number of connected components in the graph. In fact, a broad family of graph regularizers can be used in this context [6, 7]. This includes the iterated graph Laplacian which is theoretically better than the standard Laplacian [7], or the diffusion kernel on graphs.

Based on Representer Theorem [8], the solution of (4) has the form $f^*(x) = \alpha^T k(X, x)$, where $\alpha \in \mathbf{R}^n$, and k is the kernel associated with \mathcal{H} . After substituting the form of f^* in (4), and solving the resulting least-squares problem, we can derive

$$\alpha = (K + \lambda_f n I + \lambda_{\mathcal{M}} \frac{1}{n} L K)^{-1} Y, \quad (5)$$

where matrix $K \in \mathbf{S}_+^n$ is the gram matrix, and matrix I is an identity matrix of matching size.

2.2 Kernelized LSTD with ℓ_2 regularization

Farahmand et al. [9] introduce an ℓ_2 regularization extension to kernelized LSTD [10], termed Regularized LSTD (REG-LSTD), featuring better control of the complexity of the function approximator through regularization, and mitigating the burden of selecting basis functions through kernel machinery. REG-LSTD is formulated by adding ℓ_2 regularization terms to both steps in a nested minimization problem that is equivalent to the original LSTD [11]:

$$h_Q = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \mathbf{E}_{s,a} (h(s, a) - T^\pi Q(s, a))^2 + \lambda_h \|h\|_{\mathcal{H}}^2, \quad (6)$$

$$Q^* = \underset{Q \in \mathcal{H}}{\operatorname{argmin}} \mathbf{E}_{s,a} (Q(s, a) - h_Q(s, a))^2 + \lambda_Q \|Q\|_{\mathcal{H}}^2, \quad (7)$$

² P_X denotes the marginal probability distribution of inputs, and $f(X) = [f(x_1) \ \dots \ f(x_n)]^T$.

³ \mathbf{S}_+^n denotes the set of symmetric $n \times n$ positive semidefinite matrices.

where informally speaking, h_Q in (6) is the projection of $T^\pi Q$ (2) on \mathcal{H} , and Q is enforced to be close to that projection in (7). The kernel k can be selected as a tensor product kernel $k(x, x') = k_1(s, s')k_2(a, a')$, where $x = (s, a)$ (note that the multiplication of two kernels is a kernel [12]). Furthermore, k_2 can be the Kronecker delta function, if the admissible action set \mathcal{A} is finite.

As in LSTD, the expectations in (6) and (7) are then approximated with finite samples $\mathcal{D} = \{(s_i, a_i, s'_i, r_i)\}_{i=1}^n$, leading to

$$h_Q = \operatorname{argmin}_{h \in \mathcal{H}} \frac{1}{n} \|h(X) - R - \gamma Q(X')\|_2^2 + \lambda_h \|h\|_{\mathcal{H}}^2 \quad (8)$$

$$Q^* = \operatorname{argmin}_{Q \in \mathcal{H}} \frac{1}{n} \|Q(X) - h_Q(X)\|_2^2 + \lambda_Q \|Q\|_{\mathcal{H}}^2, \quad (9)$$

where X is defined as in (3) with $x_i = (s_i, a_i)$, similarly X' consists of $x'_i = (s'_i, \pi(s'_i))$ with actions generated by the policy π to evaluate, and reward vector $R = r(X)$. Invoking Representer theorem [8], Farahmand et al. [9] showed that

$$h_Q(x) = \beta^T k(X, x), \quad Q^*(X) = \alpha^T k(\tilde{X}, x) \quad (10)$$

where weight vectors $\beta \in \mathbf{R}^n$ and $\alpha \in \mathbf{R}^{2n}$, and matrix \tilde{X} is X and X' stacked. Substituting (10) in (6) and (7), we obtain the formula for α :

$$\alpha = (F^T F K_Q + \lambda_Q n I)^{-1} F^T E R, \quad (11)$$

where $F = C_1 - \gamma E C_2$, $E = K_h (K_h + \lambda_h n I)^{-1}$, $K_Q = K(\tilde{X}, \tilde{X})$, $K_h = K(X, X)$, $C_1 = [I \ 0]$, and $C_2 = [0 \ I]$.

3 Our approach

Our approach combines the benefits of both manifold regularized learning (§2.1) that incorporates geometric structure, and kernelized LSTD with ℓ_2 regularization (§2.2) that offers better control over function complexity and eases feature selection.

More concretely, besides the regularization term in (9) that controls the complexity of Q in the ambient space, we augment the objective function with a manifold regularization term that enforces Q to be smooth on the manifold that supports P_X . In particular, if the manifold regularization term is chosen as in §2.1, large penalty is imposed if Q varying too fast along the manifold. With the empirically estimated manifold regularization term, optimization problem (9) becomes (cf., (4))

$$Q^* = \operatorname{argmin}_{Q \in \mathcal{H}} \frac{1}{n} \|Q(X) - h_Q(X)\|_2^2 + \lambda_Q \|Q\|_{\mathcal{H}}^2 + \lambda_{\mathcal{M}} \frac{1}{(2n)^2} \|Q(\tilde{X})\|_L^2, \quad (12)$$

which admits the optimal weight vector α (cf., (11)):

$$\alpha = (F^T F K_Q + \lambda_Q n I + \frac{1}{4n} \lambda_{\mathcal{M}} L K_Q)^{-1} F^T E R. \quad (13)$$

4 Experimental results

We present experimental results on two standard RL benchmarks: two-room navigation and cart-pole balancing. REG-LSTD with manifold regularization (MR-LSTD) (§3) is compared against REG-LSTD without manifold regularization (§2.2) and LSTD with three commonly used basis function construction mechanisms: polynomial [13, 2], radial basis functions (RBF) [2, 14], and Laplacian eigenmap [15, 16], in terms of the quality of the policy produced by Least-Squares Policy Iteration (LSPI) [2]. The kernel used in the experiments is $k(x, x') = \exp(-\frac{\|s-s'\|_2^2}{2\sigma^2})\delta(a-a')$ with hyperparameter σ . We use combinatorial Laplacian with adjacency matrix computed from ϵ -neighborhood with $\{0, 1\}$ weights [5].

4.1 Two-room navigation

The two-room navigation problem is a classic benchmark for RL algorithms that cope with either discrete [16, 17, 15] or continuous state space [14, 18]. In the vanilla two-room problem, the state

# of samples	polynomial	RBF	eigenmap	REG-LSTD	MR-LSTD
250	7.00	16.31	6.44	6.69	4.40
500	5.46	11.7	3.9	2.43	0.56
1,000	4.55	2.4	1.56	0.24	0.02

Table 1: *Two-room navigation*. Quality of policies computed from LSPI-based methods with different basis functions, measured by the number of states whose corresponding actions are NOT optimal (the smaller the better), with varying number of samples and averaged over 100 runs. The maximum number of iterations of LSPI is 50. Best performance among LSPI iterations and coarsely tuned hyperparameters is reported. In particular, polynomial degree ranges from 1 to 8, and RBF discretization per-dimension ranges from 2 to 7.

# of samples	polynomial	RBF	REG-LSTD	MR-LSTD
250	121.10	116.59	186.82	195.78
500	150.50	112.22	181.44	195.91
1,000	154.38	130.74	188.82	198.12

Table 2: *Cart-pole*. Quality of policies computed from LSPI-based methods with different basis functions, measured by the average number of steps before termination over 100 trials (the larger the better). The maximum length of trials is set to 200. All the numbers are the average over 100 runs, and the LSPI iteration number and hyperparameter tuning are the same as in two-room navigation.

space is a 10×10 discrete grid world, and admissible actions are stepping in one of the four cardinal directions, *i.e.*, up, right, down, and left. The dynamics is stochastic: each action succeeds with probability 0.9 when movement is not blocked by obstacle or border, otherwise leaves the agent in the same location. The goal is to navigate to the diagonally opposite corner in the other room, with a single-cell doorway connecting the two rooms. Reward is 1 at the goal location otherwise 0, and the discount factor is set to 0.9. Data are collected beforehand by uniformly sampling states and actions, and used throughout LSPI iterations. Seen from Table 1, Laplacian eigenmap which exploits intrinsic state geometry outperforms parametric basis functions: polynomial and RBF. The MR-LSTD method we propose achieves the best performance.

4.2 Cart-pole Balancing

The cart-pole balancing task is to balance a pole upright by applying force to the cart to which it’s attached⁴ [19]. The agent constantly receives 0 reward until trial ends with -1 reward when it’s 12° away from the upright posture. On the contrary to the two-room navigation task, the state space is continuous, consisting of angle and angular velocity of the pole. Admissible actions are finite: pushing the cart to left or right. The discount factor $\gamma = 0.99$. Data are collected from random episodes, *i.e.*, starting from a perturbed upright posture and applying uniformly sampled actions. Results are reported in Table 2, which shows that REG-LSTD achieves significantly better performance than parametric basis functions, and performance is even improved further with manifold regularization.

5 Related work

The closest work to this paper was recently introduced by Li et al. [20], which utilized manifold regularization by learning state representation through unsupervised learning, and then adopting the learned representation in policy iteration. In contrast to this work, we naturally blend manifold regularization with policy evaluation with possibly provable performance guarantee (left for future work). There is also work on constructing basis functions directly from geometry, *e.g.*, Laplacian methods [15, 16], and geodesic Gaussian kernels [14]. Furthermore, different regularization mechanisms to LSTD have been proposed, including ℓ_1 regularization for feature selection [21], and nested ℓ_2 and ℓ_1 penalization to avoid overfitting [22].

⁴The cart-pole environment in OpenAI Gym package is used in our implementation.

6 Conclusion and future work

We propose manifold regularization for a kernelized LSTD approach in order to exploit the intrinsic geometry of the state space for better sample efficiency and Q-function approximation, and demonstrate superior performance on two standard RL benchmarks. Future work directions include 1) accelerating by structured random matrices for kernel machinery [23, 24], and graph sketching for graph regularizer construction to scale up to large datasets and rich observations, *e.g.*, images, 2) providing theoretical justification, and combining manifold regularization with deep neural nets [25, 26] and other policy evaluation, *e.g.*, [27, 10] and policy iteration algorithms, 3) learning with a data-dependent kernel that capturing the geometry (equivalent to the manifold regularized solution [28]) that makes it easier to derive new algorithms, and 4) extension to continuous action spaces by constructing kernels such that policy improvement (optimize over actions) is tractable [12].

References

- [1] S. J. Bradtke and A. G. Barto. Linear least-squares algorithms for temporal difference learning. In *Recent Advances in Reinforcement Learning*, pages 33–57. Springer, 1996.
- [2] M. G. Lagoudakis and R. Parr. Least-squares policy iteration. *Journal of machine learning research*, 4(Dec):1107–1149, 2003.
- [3] M. Posa, S. Kuindersma, and R. Tedrake. Optimization and stabilization of trajectories for constrained dynamical systems. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 1366–1373. IEEE, 2016.
- [4] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov): 2399–2434, 2006.
- [5] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.
- [6] A. J. Smola and R. Kondor. Kernels and regularization on graphs. In *COLT*, volume 2777, pages 144–158. Springer, 2003.
- [7] X. Zhou and M. Belkin. Semi-supervised learning by higher order regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 892–900, 2011.
- [8] B. Schölkopf, R. Herbrich, and A. Smola. A generalized representer theorem. In *Computational learning theory*, pages 416–426. Springer, 2001.
- [9] A. M. Farahmand, M. Ghavamzadeh, S. Mannor, and C. Szepesvári. Regularized policy iteration. In *Advances in Neural Information Processing Systems*, pages 441–448, 2009.
- [10] X. Xu, D. Hu, and X. Lu. Kernel-based least squares policy iteration for reinforcement learning. *IEEE Transactions on Neural Networks*, 18(4):973–992, 2007.
- [11] A. Antos, C. Szepesvári, and R. Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1): 89–129, 2008.
- [12] M. G. Genton. Classes of kernels for machine learning: a statistics perspective. *Journal of machine learning research*, 2(Dec):299–312, 2001.
- [13] P. J. Schweitzer and A. Seidmann. Generalized polynomial approximations in markovian decision processes. *Journal of mathematical analysis and applications*, 110(2):568–582, 1985.
- [14] M. Sugiyama, H. Hachiya, C. Towell, and S. Vijayakumar. Geodesic gaussian kernels for value function approximation. *Autonomous Robots*, 25(3):287–304, 2008.
- [15] S. Mahadevan. Representation policy iteration. *arXiv preprint arXiv:1207.1408*, 2012.

- [16] M. Petrik. An analysis of laplacian methods for value function approximation in mdps. In *IJCAI*, pages 2574–2579, 2007.
- [17] R. Parr, L. Li, G. Taylor, C. Painter-Wakefield, and M. L. Littman. An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pages 752–759. ACM, 2008.
- [18] G. Taylor and R. Parr. Kernelized value function approximation for reinforcement learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1017–1024. ACM, 2009.
- [19] A. G. Barto, R. S. Sutton, and C. W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, (5): 834–846, 1983.
- [20] H. Li, D. Liu, and D. Wang. Manifold regularized reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2017.
- [21] J. Z. Kolter and A. Y. Ng. Regularization and feature selection in least-squares temporal difference learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 521–528. ACM, 2009.
- [22] M. Hoffman, A. Lazaric, M. Ghavamzadeh, and R. Munos. Regularized least squares temporal difference learning with nested l2 and l1 penalization. *Recent Advances in Reinforcement Learning*, 7188:102–114, 2012.
- [23] X. Y. Felix, A. T. Suresh, K. M. Choromanski, D. N. Holtmann-Rice, and S. Kumar. Orthogonal random features. In *Advances in Neural Information Processing Systems*, pages 1975–1983, 2016.
- [24] K. Choromanski, M. Rowland, and A. Weller. The unreasonable effectiveness of random orthogonal embeddings. *arXiv preprint arXiv:1703.00864*, 2017.
- [25] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.
- [26] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller. Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 387–395, 2014.
- [27] B. Dai, N. He, Y. Pan, B. Boots, and L. Song. Learning from conditional distributions via dual kernel embeddings. *arXiv preprint arXiv:1607.04579*, 2016.
- [28] V. Sindhwani, P. Niyogi, and M. Belkin. Beyond the point cloud: from transductive to semi-supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 824–831. ACM, 2005.