Learning from Conditional Distributions via Dual Embeddings

Bo Dai[1], Niao He[2], Yunpeng Pan[1], Byron Boots[1], Le Song[1]

[1] Georgia Institute of Technology
{bodai, ypan37}@gatech.edu, {lsong, bboots}@cc.gatech.edu
[2] University of Illinois at Urbana-Champaign
niaohe@illinois.edu

Abstract

Many machine learning tasks, such as learning with invariance and policy evaluation in reinforcement learning, can be characterized as problems of *learning from conditional distributions*. In such problems, each sample x itself is associated with a conditional distribution $p(z|x)$ represented by samples $\{z_i\}_{i=1}^M$, and the goal is to learn a function f that links these conditional distributions to target values y. These learning problems become very challenging when we only have limited samples or in the extreme case only one sample from each conditional distribution.

To address these challenges, we propose a novel approach which employs a new *min-max reformulation* of the learning from conditional distribution problem. With such new reformulation, we only need to deal with the *joint distribution* $p(z, x)$. We also design an efficient learning algorithm, *Embedding-SGD*, and establish theoretical sample complexity for such problems. Empirical experiments demonstrate the advantages of our algorithm.

1 Introduction

We address the problem of *learning from conditional distributions* where the goal is to learn a function that links conditional distributions to target variables. Specifically, we are provided input samples $\{x_i\}_{i=1}^N \in \mathcal{X}^N$ and their corresponding responses $\{y_i\}_{i=1}^N \in \mathcal{Y}^N$. For each $x \in \mathcal{X}$, there is an associated conditional distribution $p(z|x) : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}$. However, we cannot access the entire conditional distributions $\{p(z|x_i)\}_{i=1}^N$ directly; rather, we only observe a limited number of samples or in the extreme case only *one sample* from each conditional distribution $p(z|x)$. The task is to learn a function f which links the conditional distribution $p(z|x)$ to target $y \in \mathcal{Y}$ by minimizing the expected loss:

$$\min_{f\in\mathcal{F}} L(f) = \mathbb{E}_{x,y}\left[\ell\left(y, \mathbb{E}_{z|x}\left[f(z,x)\right]\right)\right] \tag{1}$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a convex loss function. The function space $\mathcal{F}$ can be very general, but we focus on the case when $\mathcal{F}$ is a reproducing kernel Hilbert space (RKHS) in main text, namely, $\mathcal{F} = \{f : \mathcal{Z} \times \mathcal{X} \to \mathbb{R} \,|\, f(z,x) = \langle f, \psi(z,x)\rangle\}$ where $\psi(z,x)$ is a suitably chosen (nonlinear) feature map [1].

The problem of learning from conditional distributions appears in many different tasks. For example: **Learning with invariance.** The goal of invariance learning is to estimate a function which minimizes the expected risk while at the same time preserving consistency over a group of operations $g = \{g_j\}_{j=1}^\infty$. [10] shows that this can be accomplished by solving the following optimization problem

$$\min_{f\in\tilde{\mathcal{H}}} \mathbb{E}_{x,y}[\ell(y, \mathbb{E}_{z|x\sim\mu(g(x))}[\langle f, \psi(z)\rangle_{\tilde{\mathcal{H}}}])] + (\nu/2)\|f\|_{\tilde{\mathcal{H}}}^2 \tag{2}$$

where $\tilde{\mathcal{H}}$ is the RKHS corresponding to kernel $\tilde{k}$ with implicit feature map $\psi(\cdot)$, $\nu > 0$ is the regularization parameter, and $\mu(g(x))$ stands for some normalized Haar measure. Obviously, the above optimization (2) is a special case of (1).

[1] Please refer the full version [2] for the extension to arbitrary function approximators, *e.g.*, random features and neural networks.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Policy evaluation in reinforcement learning. Policy evaluation is a fundamental task in reinforcement learning. Given a policy $\pi(a|s)$ which is a distribution over action condition on state s, the goal is to estimate the value function $V^{\pi}(\cdot)$ over the state space. $V^{\pi}(s)$ is the fixed point of the Bellman equation $V^{\pi}(s) = \mathbb{E}_{s'|a,s}[R(s,a) + \gamma V^{\pi}(s')]$, where $R(s,a)$ is the reward and $\gamma \in (0,1)$ is the discount factor. Therefore, the value function can be estimated from data by minimizing the mean-square Bellman error [1, 16]:

$$\min_{V^{\pi}} \mathbb{E}_{s,a}\left[\left(R(s,a) - \mathbb{E}_{s'|a,s}\left[V^{\pi}(s) - \gamma V^{\pi}(s')\right]\right)^2\right]. \tag{3}$$

Restrict the policy to lie in some RKHS, this optimization is clearly a special case of (1) by viewing $((s,a), R(s,a), s')$ as (x,y,z) in (1). Due to the online nature of MDPs, we usually observe only one successor state s' sample from the conditional distribution given s, a.

Despite the prevalence of learning problems in the form of (1), solving such problem remains very challenging for two reasons: (*i*) we often have limited samples or in the extreme case only one sample from each conditional distribution $p(z|x)$, making it difficult to accurately estimate the conditional expectation. (*ii*) the conditional expectation is nested inside the loss function, making the problem quite different from the traditional stochastic optimization setting. As far as we known, very few results have been established in this domain.

To address the above challenges, we propose a novel approach called *dual kernel embedding*. The key idea is to reformulate (1) into a saddle point problem by utilizing the Fenchel duality of the loss function. We observe that with smooth loss function and continuous conditional distributions, the dual variables form a continuous function of x and y. Therefore, we can parameterize it as a function in some RKHS induced by any universal kernel, where the information about $p(x)$ and $p(z|x)$ can be aggregated via a kernel embedding of the joint distribution $p(x,z)$. Furthermore, we propose an efficient algorithm based on stochastic approximation to solve the resulted saddle point problem over RKHS spaces and establish finite-sample analysis. Compared to existing approaches, *e.g.*, stochastic average appproximation (SAA) and learning with kernel embedding [18], an advantage of the proposed method is that it requires only *one sample* from each conditional distribution. Under mild conditions, the overall sample complexity reduces to $\mathcal{O}(1/\epsilon^2)$ in contrast to the $\mathcal{O}(1/\epsilon^4)$ complexity required by SAA or kernel conditional embedding [15, 4, 5].

2 Dual Embedding Framework

In this section, we propose a novel and sample-efficient framework to solve problem (1).We start by introducing the interchangeability principle, which plays a fundamental role in our method. Due to space limit, please refer [2] for the complete proof.

Lemma 1 (interchangeability principle) *Let ξ be a random variable on Ξ and assume for any $\xi \in \Xi$, function $g(\cdot,\xi) : \mathbb{R} \to (-\infty, +\infty)$ is a proper and upper semicontinuous concave function. Then*

$$\mathbb{E}_{\xi}[\max_{u\in\mathbb{R}} g(u,\xi)] = \max_{u(\cdot)\in\mathcal{G}(\Xi)} \mathbb{E}_{\xi}[g(u(\xi),\xi)].$$

where $\mathcal{G}(\Xi) = \{u(\cdot) : \Xi \to \mathbb{R}\}$ is the entire space of functions defined on support Ξ.

The result implies that one can replace the expected value of point-wise optima by the optimum value over a function space. More general results of interchange between maximization and integration can be found in [13, Chapter 14] and [14, Chapter 7].

2.1 Saddle Point Reformulation

Let the loss function $\ell_y(\cdot) := \ell(y,\cdot)$ in (1) be a proper, convex and lower semicontinuous for any y. We denote $\ell_y^*(\cdot)$ as the convex conjugate; hence $\ell_y(v) = \max_u\{uv - \ell_y^*(u)\}$, which is also a proper, convex and lower semicontinuous function. Using the Fenchel duality, we can reformulate problem (1) as

$$\min_{f\in\mathcal{F}} \mathbb{E}_{xy}\left[\max_{u\in\mathbb{R}}\left[\mathbb{E}_{z|x}[f(z,x)]\cdot u - \ell_y^*(u)\right]\right], \tag{4}$$

Note that by the concavity and upper-semicontinuity of $-\ell_y^*(\cdot)$, for any given pair (x,y), the corresponding maximizer of the inner function always exists. Based on the interchangeability principle stated in Lemma 1, we can further rewrite (4) as

$$\min_{f\in\mathcal{F}} \max_{u(\cdot)\in\mathcal{G}(\Xi)} \Phi(f,u) := \mathbb{E}_{zxy}[f(z,x)\cdot u(x,y)] - \mathbb{E}_{xy}[\ell_y^*(u(x,y))], \tag{5}$$

where $\Xi = \mathcal{X} \times \mathcal{Y}$ and $\mathcal{G}(\Xi) = \{u(\cdot) : \Xi \to \mathbb{R}\}$ is the entire function space on Ξ. We emphasize that the $\max$-operator in (4) and (5) have different meanings: the one in (4) is taking over a single variable, while the other one in (5) is over all possible function $u(\cdot) \in \mathcal{G}(\Xi)$.

Now that we have eliminated the nested expectation in the problem of interest, and converted it into a stochastic saddle point problem with an additional dual function space to optimize over. By definition, $\Phi(f,u)$ is always concave in u for any fixed f. Since $f(z,x) = \langle f, \psi(z,x)\rangle$, $\Phi(f,u)$ is also convex in f for any fixed u. Our reformulation (5) is indeed a convex-concave saddle point problem.

2.2 Dual Continuation

Although the reformulation in (5) gives us more structure of the problem, it is not yet tractable in general. This is because the dual function $u(\cdot)$ can be an arbitrary function which we do not know how to represent. In the following, we will introduce a tractable representation for (5).

First, we will define the function $u^*(\cdot) : \Xi = \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ as the *optimal dual function* if for any pair $(x,y) \in \Xi$,

$$u^*(x,y) \in \operatorname{argmax}_{u\in\mathbb{R}} \left\{ u \cdot \mathbb{E}_{z|x}[f(z,x)] - \ell_y^*(u) \right\}.$$

Note the optimal dual function is well-defined since the optimal set is nonempty. Furthermore, $u^*(x,y)$ is related to the conditional distribution via $u^*(x,y) \in \partial \ell_y(\mathbb{E}_{z|x}[f(z,x)])$ [6]. Depending on the property of the loss function $\ell_y(v)$, we can further derive that:

Proposition 1 *Suppose both $f(z,x)$ and $p(z|x)$ are continuous in x for any z,*

(1) **(Discrete case)** *If the loss function $\ell_y(v)$ is continuously differentiable in v for any $y \in \mathcal{Y}$, then $u^*(x,y)$ is unique and continuous in x for any $y \in \mathcal{Y}$;*

(2) **(Continuous case)** *If the loss function $\ell_y(v)$ is continuously differentiable in (v,y), then $u^*(x,y)$ is unique and continuous in (x,y) on $\mathcal{X} \times \mathcal{Y}$.*

The fact that the optimal dual function is a continuous function has important consequences. As we mentioned earlier, the space of dual functions can be arbitrary and difficult to represent. Now we can simply restrict the parametrization to the space of continuous functions, which is tractable and still contains the global optimum of the optimization problem in (5).

2.3 Kernel Embedding

For the sake of simplicity, we focus on the case when $\mathcal{Y}$ is a continuous set, and thus, under Proposition 1, the optimal dual function is indeed continuous in $(x,y) \in \Xi = \mathcal{X} \times \mathcal{Y}$. Therefore, we lose nothing by restricting the dual function space $\mathcal{G}(\Xi)$ to be continuous function space on Ξ. Recall that with the universal kernel, we can approximate any continuous function with arbitrarily small error. Thus we approximate the dual space $\mathcal{G}(\Xi)$ by the bounded RKHS $\mathcal{H}^\delta$ induced by a universal kernel $k((x,y),(x',y')) = \langle \phi(x,y), \phi(x',y')\rangle_{\mathcal{H}}$, implying $u(x,y) = \langle u, \phi(x,y)\rangle_{\mathcal{H}}$. Note that $\mathcal{H}^\delta$ is a subspace of the continuous function space, and hence is a subspace of the dual space $\mathcal{G}(\Xi)$. We denote the inner product in $\mathcal{F}$ as $\langle\cdot,\cdot\rangle_{\mathcal{F}}$ to distinguish the dual RKHS $\mathcal{H}^\delta$.

We can rewrite the saddle point problem in (5) as

$$\min_{f\in\mathcal{F}} \max_{u\in\mathcal{H}^\delta} \Phi(f,u) = \mathbb{E}_{xyz}\left[\langle f, \psi(z,x)\rangle_{\mathcal{F}} \cdot \langle u, \phi(x,y)\rangle_{\mathcal{H}} - \ell_y^*(\langle u, \phi(x,y)\rangle_{\mathcal{H}}) \right]. \tag{6}$$

This new formulation based on dual kernel embedding allows us to efficient represent the dual function and get away from the fundamental difficulty with insufficient sampling from the conditional distribution. There is no need to access either the conditional distribution $p(z|x)$, the conditional expectation $\mathbb{E}_{z|x}[\cdot]$, or the conditional embedding operator $\mathcal{U}_{z|x}$ anymore, therefore, reducing both the statistical and computational complexity.

2.4 Sample-Efficient Algorithm

The algorithm is summarized in Algorithm 1. At each iteration, the algorithm performs a projected gradient step both for the primal variable f and dual variable u based on the unbiased stochastic gradient. The proposed algorithm avoids the need for overwhelmingly large sample sizes from the conditional distributions when estimating the gradient. At each iteration, only one sample from the conditional distribution is required in our algorithm!

Theorem 1 *If $f(z,x)$ is uniformly bounded and $\ell_y^*(v)$ is uniformly Lipschitz continuous in v for any y, and the kernel function and $\mathbb{E}_{z,x}[\|f(z,x)\|_2^2]$, $\mathbb{E}_{z,x}[\|\psi(z,x)\|_{\mathcal{F}}^2]$, $\mathbb{E}_y[\|\nabla \ell_y^*(u)\|_2^2]$ are bounded, denote f_* be the optimal solution to (1), we have*

$$\mathbb{E}[L(\bar{f}_t) - L(f_*)] \leqslant \mathcal{O}\left(\frac{\delta^{3/2}}{\sqrt{t}} + \mathcal{E}(\delta) \right). \tag{7}$$

There is clearly a delicate trade-off between the optimization error and approximation error. Using large δ will increase the optimization error but decrease the approximation error. When δ is moderately large (which is expected in the situation when the optimal dual function has small magnitude), our dual kernel embedding algorithm can achieve an overall $\mathcal{O}(1/\epsilon^2)$ sample complexity when solving learning problems in the form of (1).

Algorithm 1 Embedding-SGD for Optimization (6)

```
Input: $p(x, y)$, $p(z|x)$, $\psi(z, x)$, $\phi(x, y)$, $\{\gamma_i \geqslant 0\}_{i=1}^t$
for $i = 1, \ldots, t$ do
    Sample $(x_i, y_i) \sim p(x, y)$ and $z_i \sim p(z|x)$.
    $f_{i+1} = \Pi_{\mathcal{F}}(f_i - \gamma_i \psi(z_i, x_i) u_i(x_i, y_i))$.
    $u_{i+1} = \Pi_{\mathcal{H}^\delta}(u_i + \gamma_i [f_i(z_i, x_i) - \nabla \ell_{y_i}^*(u_i(x_i, y_i))] \phi(x_i, y_i))$
end for
Output: $\bar{f}_t = \frac{\sum_{i=1}^t \gamma_i f_i}{\sum_{i=1}^t \gamma_i}$, $\bar{u}_t = \frac{\sum_{i=1}^t \gamma_i u_i}{\sum_{i=1}^t \gamma_i}$
```

3 Experiments

We test the proposed algorithm for two applications, *i.e.*, learning with invariant representation and policy evaluation.

3.1 Experiments on Invariance Learning

We test the proposed algorithm for learning with invariance task on QuantumMachine 5-fold dataset for atomization energy prediction. We compare the proposed algorithm with SGD with virtual samples technique [11, 8] and SGD with finite sample average for inner expectation (SGD-SAA). We use Gaussian kernel in all tasks. We follow [9] that the data points are represented by Coulomb matrices, and the virtual samples are generated by random permutation. To demonstrate the sample-efficiency of our algorithm, 10 virtual samples are generated for each datum in training phase. The average results are shown in Figure 1(b). The proposed algorithm achieves a significant better solution, while SGD-SAA and SGD with virtual samples stuck in inferior solutions due to the inaccurate inner expectation estimation and optimizing indirect objective, respectively.

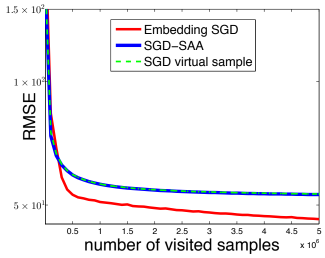
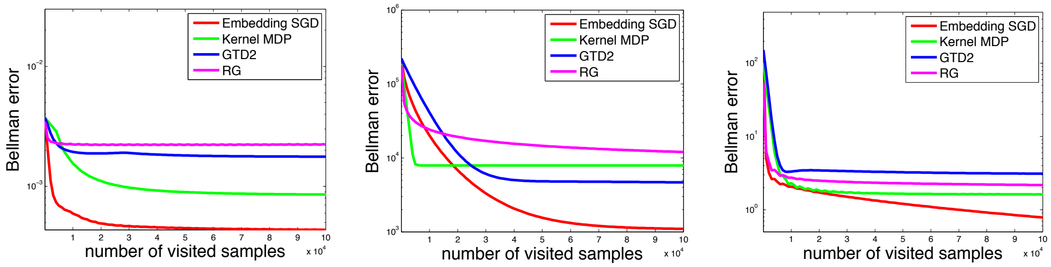


Figure 1: Invariance learning.

3.2 Experiments on Policy Evaluation



(a) Navigation (b) Cart-Pole (c) PUMA-560

Figure 2: Policy evaluation.

We compare the proposed algorithm to several prevailing algorithms for policy evaluation, including gradient-TD2 (GTD2) [17, 7], residual gradient (RG) [1] and kernel MDP [5] in terms of mean square Bellman error [3]. It should point out that kernel MDP is not an online algorithm, since it requires to visit the entire dataset when estimating the embedding and inner expectation in each iteration. We conduct experiments for policy evaluation on several benchmark datasets, including navigation, cart-pole swing up and PUMA-560 manipulation. We use Gaussian kernel in the nonparametric algorithms, *i.e.*, kernel MDP and Embedding SGD, while we test random Fourier features [12] for the parametric competitors, *i.e.*, GTD2 and RG. Results are averaged over 10 independent trials.

In all experiments, the proposed algorithm performs consistently better than the competitors. The advantages of proposed algorithm mainly come from three aspects: **i**), it utilizes more flexible dual function space, rather than the constrained space in GTD2; **ii**), it directly optimizes the MSBE, rather than its surrogate as in GTD2 and RG; **iii**), it directly targets on value function estimation and forms an one-shot algorithm, rather than a two-stage procedure in kernel MDP including estimating conditional kernel embedding as an intermediate step.

References

[1] Leemon Baird. Residual algorithms: reinforcement learning with function approximation. In *Proc. Intl. Conf. Machine Learning*, pages 30–37. Morgan Kaufmann, 1995.

[2] Bo Dai, Niao He, Yunpeng Pan, Byron Boots, and Le Song. Learning from conditional distributions via dual kernel embeddings. *CoRR*, abs/1607.04579, 2016.

[3] Christoph Dann, Gerhard Neumann, and Jan Peters. Policy evaluation with temporal differences: a survey and comparison. *Journal of Machine Learning Research*, 15(1):809–883, 2014.

[4] S. Grunewalder, G. Lever, L. Baldassarre, S. Patterson, A. Gretton, and M. Pontil. Conditional mean embeddings as regressors. In *ICML*, 2012.

[5] S. Grunewalder, G. Lever, L. Baldassarre, M. Pontil, and A. Gretton. Modeling transition dynamics in MDPs with RKHS embeddings. In *ICML*, 2012.

[6] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2012.

[7] Bo Liu, Ji Liu, Mohammad Ghavamzadeh, Sridhar Mahadevan, and Marek Petrik. Finite-sample analysis of proximal gradient td algorithms. In *Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2015.

[8] G. Loosli, S. Canu, and L. Bottou. Training invariant support vector machines with selective sampling. In L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, editors, *Large Scale Kernel Machines*, pages 301–320. MIT Press, 2007.

[9] Grégoire Montavon, Katja Hansen, Siamac Fazli, Matthias Rupp, Franziska Biegler, Andreas Ziehe, Alexandre Tkatchenko, Anatole von Lilienfeld, and Klaus-Robert Müller. Learning invariant representations of molecules for atomization energy prediction. In *Neural Information Processing Systems*, pages 449–457, 2012.

[10] Youssef Mroueh, Stephen Voinea, and Tomaso A Poggio. Learning with group invariant features: A kernel perspective. In C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1558–1566. Curran Associates, Inc., 2015.

[11] P. Niyogi, F. Girosi, and T. Poggio. Incorporating prior knowledge in machine learning by creating virtual examples. *Proceedings of IEEE*, 86(11):2196–2209, November 1998.

[12] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, MA, 2008.

[13] R. T. Rockafellar and R. J-B. Wets. *Variational Analysis*. Springer Verlag, 1998.

[14] Alexander Shapiro, Darinka Dentcheva, et al. *Lectures on stochastic programming: modeling and theory*, volume 16. SIAM, 2014.

[15] L. Song, A. Gretton, and K. Fukumizu. Kernel embeddings of conditional distributions. *IEEE Signal Processing Magazine*, 2013.

[16] Richard S. Sutton, Hamid R. Maei, and Csaba Szepesvári. A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1609–1616. 2008.

[17] Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 993–1000. ACM, 2009.

[18] Zoltán Szabó, Bharath Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40, 2016.