Projected Stochastic Primal-Dual Method for Constrained Online Learning with Kernels

Kaiqing Zhang, Hao Zhu, Tamer Başar, and Alec Koppel

***Abstract*— We consider the problem of stochastic optimization with nonlinear constraints, where the decision variable is not vector-valued but instead a function belonging to a reproducing Kernel Hilbert Space (RKHS). Currently, there exist solutions to only special cases of this problem. To solve this constrained problem with kernels, we first generalize the Representer Theorem to a class of saddle-point problems defined over RKHS. Furthermore, we develop a primal-dual method which executes alternating projected primal/dual stochastic descent/ascent on the dual-augmented Lagrangian of this problem. The primal projection sets are low-dimensional subspaces of the ambient function space which are greedily constructed using matching pursuit. By tuning the projection-induced error to the algorithm step-size, we are able to establish mean convergence both in primal objective sub-optimality and constraint violation, respectively to the $\mathcal{O}(\sqrt{T})$ and $\mathcal{O}(T^{3/4})$ neighborhoods, where T is the total number of iterations. We evaluate the proposed method through numerical tests for the application of risk-aware supervised learning.**

I. Introduction

In this work, we seek to solve nonlinearly constrained stochastic optimization problems, where the decision variable is a function rather than a parameter vector. Typically, the stochastic problem nature results in objective function as an expectation over some unknown data distribution. Moreover, the constraint is problem-setting dependent, though oftentimes it relates to the unknown data distribution. This situation could arise in several applications such as motion planning with obstacle avoidance [2], wireless communications with quality-of-service (QoS) guarantees [3], or nonlinear filtering with built-in outlier rejection capability [4].

The theory of function-valued constrained optimization dates back to calculus of variations [5] and Hamilton's Principle [6]. Nonetheless, engineered systems motivate more generic situations than those arising from physical laws. Meanwhile, variational inference methods have been developed to handle functional stochastic problems in statistical inference, especially hyper-parameter search [7]. Yet, unless special distributional structure is present, they typically do not admit efficient iterative solutions and require solving an intractable integral equation.

To tackle this intractability issue, one shall not only restrict the function of interest to yield a computationally tractable formulation, but also allow it to be rich enough for realistic scenarios. For instance, in learning theory and controls, we could restrict the function to take the form of a polynomial function [8], a Gaussian process [9], a neural network [10], or a nonparametric basis expansion in terms of data [11]. Motivated by the generalizability of the latter, we adopt the nonparametric approach based on the reproducing Kernel Hilbert Space (RKHS) This is motivated by a recent result that allows for memory-efficient parameterization for its infinite-dimensional function solution [12]. This approach subsumes polynomial interpolation [8], and provides a methodology that circumvents the memory explosion associated with large sample-size Gaussian process regression [13]. Moreover, it can preserve convexity, thus avoiding getting stuck at poor stationary points as in neural network training [14].

Our goal is to extend the kernelized functional stochastic programming approach of [12] to settings with nonlinear constraints. Preliminary efforts to address this problem include [15], [16] based on proximal projections or penalty methods. Unfortunately, their applicability is limited to specialized constraints that exclude obstacle avoidance [17], wireless QoS guarantees [3], or risk measures such as conditional value-at-risk (CVaR) [18] that may be crucial for addressing bias-variance issues in statistical learning. The main challenge in handling general nonlinear constraints in RKHS optimization is that the Representer Theorem [19], a key result for transforming unconstrained functional problems to the parametric form, does not apply directly. Thus, we develop an augmented Lagrangian relaxation of the constrained problem, and generalize the Representer Theorem for the resultant minmax problem. This generalization requires the constraints to be of certain structure that is reasonable in practice.

Furthermore, we develop a stochastic saddle-point algorithm which operates by executing alternating projected primal/dual stochastic gradient descent/ascent on the augmented Lagrangian. Owing to the recursive kernel addition to include new data points, the parameterization of the function iterate grows linearly with time. To address this issue, we project the primal function iterate onto low-dimensional subspaces which are greedily constructed using matching pursuit [20]. By tuning the projection-induced error to the algorithm step-size as in [12], we establish the mean convergence to a $\mathcal{O}(\sqrt{T})$ neighborhood of the objective sub-optimality and $\mathcal{O}(T^{3/4})$ constraint violation by choosing a constant step-size at $1/\sqrt{T}$ and approximation budget at $1/T$. These results are akin to those for primal-dual method for vector-valued stochastic programming with nonlinear constraints [21]. Our proposed algorithm is applied to develop for the first time an online risk-aware supervised learning with

K. Zhang and T. Başar are with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign ({kzhang66, basar1}@illinois.edu). H. Zhu is with Dept. of Electrical & Computer Engineering at University of Texas at Austin (haozhu@utexas.edu). A. Koppel is with U.S. Army Research Laboratory (alec.e.koppel.civ@mail.mil). Proofs are given in [1].

conditional value-at-risk constraints. Proofs are given in [1].

The rest of the paper is organized as follows. Section II formulates the constrained stochastic optimization problem in RKHS and generalizes the Representer Theorem to a class of saddle-point problems. The projected stochastic primal-dual method is presented in Section III and analyzed in Section IV. We evaluate the proposed algorithm numerically in Section V, with the paper wrapped up in Section VI.

II. Constrained Learning with Kernels

We consider the problem of constrained stochastic optimization in the reproducing kernel Hilbert spaces (RKHSs). Specifically, the objective is to minimize the average of a loss function given by $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, regularized by a complexity-reducing penalty term $\lambda/2\|f\|_{\mathcal{H}}^2$ for some $\lambda > 0$. Note that $\mathcal{H}$ represents a Hilbert space, and the sets $\mathcal{X} \in \mathbb{R}^p$ and $\mathcal{Y} \in \mathbb{R}^d$ for some $p, d > 0$. The standard interpretation of random pairs $(\boldsymbol{x}, \boldsymbol{y})$ is that $\boldsymbol{x}$ encodes feature vectors and $\boldsymbol{y}$ represents target variables, which follow some unknown joint distribution over $\mathcal{X} \times \mathcal{Y}$.The Hilbert space $\mathcal{H}$ here is a space of *functions*, $f : \mathcal{X} \to \mathcal{Y}$, that admit a representation in terms of elements of $\mathcal{X}$ when $\mathcal{H}$ has a special structure. In particular, we consider the RKHS, where $\mathcal{H}$ is equipped with a kernel function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that:

$$\begin{aligned} &\text{(i) } \langle f, \kappa(\boldsymbol{x}, \cdot)\rangle_{\mathcal{H}} = f(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x} \in \mathcal{X}, \\ &\text{(ii) } \mathcal{H} = \overline{\text{span}\{\kappa(\boldsymbol{x}, \cdot)\}} \quad \text{for all } \boldsymbol{x} \in \mathcal{X}. \end{aligned} \tag{1}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the Hilbert inner product for $\mathcal{H}$. We further assume that the kernel is positive semidefinite; i.e., $\kappa(\boldsymbol{x}, \boldsymbol{x}') \geq 0$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$. Throughout, we assume the loss function ℓ is convex with respect to (w.r.t.) $f(\boldsymbol{x})$.

Motivated by several practical applications, we further consider some hard nonlinear constraints on function f. Denoting these constraints by $\boldsymbol{G} = (G_1, \cdots, G_m)^\top$ with each $G_j : \mathcal{H} \to \mathbb{R}$ being a convex functional of f, the stochastic optimization problem can be formulated as

$$\begin{aligned} f^* = \underset{f \in \mathcal{H}}{\text{argmin}} \;\; & \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[\ell(f(\boldsymbol{x}), \boldsymbol{y})] + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 \\ \text{s.t.} \quad & \boldsymbol{G}(f) \leq \boldsymbol{0} \end{aligned} \tag{2}$$

where f^* denotes its optimum solution. Thanks to the strong convexity guaranteed by the regularization term for given positive λ, the solution f^* is unique.

The constrained stochastic optimization problem in (2), with kernels, finds practical applications in many real-time learning and decision-making problems. Two such motivating examples are presented next.

Example 1. *(Risk-aware supervised learning using CVaR):* Consider the problem of supervised learning, for example, classification or regression, where a statistical model that maps data points to decisions is usually estimated through empirical risk minimization (ERM) [22]. In particular, an empirical approximation of the objective in (2), which quantifies the bias of the learning model, is minimized. However, a desired model f should be able to mitigate not only the bias, but also the error variance. One approach to strike this bias-variance balance is to account for the dispersion of an estimate in the problem formulation [22]. Most of the existing works consider the dispersion as an extra term included by the objective function, in the form of *coherent risk*, an example of which is the *conditional value-at-risk (CVaR)* [23]. This can be viewed as a penalty-based method to reduce the dispersion of the loss function. Instead, one could directly restrict the dispersion by imposing hard constraints on the CVaR. Toward this end, the function $G : \mathcal{H} \to \mathbb{R}$ can be expressed as

$$\begin{aligned} G(f) &= \text{CVaR}_\alpha(f) - \gamma \\ &= \min_{z \in \mathbb{R}} \left\{ z + \frac{1}{1-\alpha}\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\big\{[\ell(f(\boldsymbol{x}), \boldsymbol{y}) - z]_+\big\} \right\} - \gamma \end{aligned} \tag{3}$$

where CVaR_α denotes the α-CVaR as in [18], and $\gamma > 0$ is the tolerance level that CVaR should not exceed. Here, the value α denotes the significance level which is typically chosen between 0.9 and 0.95. It follows from [18, Prop. 5] that the CVaR operator preserves convexity, and thus $G(f) \leq 0$ is an instance of the constraint in (2).

Example 2. *(Chance-constrained motion planning):* Consider the problem of motion planning in RKHS, where the objective is to find the optimal trajectory for the controlled object like robots that are both smooth and collision-free; see e.g., [2]. Specifically, a trajectory $f : [0, 1] \to \mathcal{C} \subseteq \mathbb{R}^D$ is a mapping from time t to the object coordinate $f(t) \in \mathbb{R}^D$ for some $D = 2$ or 3. Instead of observing the entire trajectory in continuous time, one may only access discrete-time samples $\{t_i\}$ drawn randomly from $[0, 1]$.The goal here is to minimize some cost functional $\mathcal{U} : \mathcal{H} \to \mathbb{R}$, which is usually convex, that quantifies proximity of the trajectory $f \in \mathcal{H}$ to a reference one. Thus, the optimization objective can be written as

$$f^* = \underset{f \in \mathcal{H}}{\text{argmin}}\, \mathbb{E}_t[\mathcal{U}(f)] + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2, \tag{4}$$

where $\mathbb{E}_t$ is the expectation over samples of the time t. Moreover, we may want to impose the hard constraint on the probability that the object will stay in a certain safe area along the entire trajectory. To this end, let $g(f(t)) > 0$ represent the shape of the safe area in $\mathbb{R}^D$, and one can aim to upper bound the probability $\mathbb{P}(g(f(t)) > 0) \leq \gamma$ for a given threshold $\gamma > 0$. Note that the probability measure follows from the randomness of t. Nonetheless, the feasible set of a chance constraint is generally non-convex except for a few special cases. To convexify the constraint, one approach is to approximate the probabilistic constraint using a more conservative constraint based on expectations [24]. Specifically, the potential surrogate is given by

$$\inf_{\lambda > 0} \big[\Psi(f, \lambda) - \lambda\gamma\big] \leq 0, \tag{5}$$

where $\Psi(f, \lambda) = \lambda\mathbb{E}_t[\phi(\lambda^{-1}g(f(t)))]$ with $\phi(\cdot)$ being the generating function. It is proven in [24] that (5) forms a convex set, and thus is an instance of the constraint in (2).

Other applications include beamforming in communication systems under robustness constraints [25] and wireless

network utility maximization with QoS constraints [3]. To develop an algorithmic solution to (2), first there are technicalities regarding extending the Representer Theorem [19] to constrained problems that must be addressed, which are done in the following subsection.

A. Representer Theorem for Constrained Optimization

We turn to developing a Representer Theorem for nonlinearly constrained problems. We will see that for the Representer Theorem to be applicable, restrictions must be imposed on the structure of the constraint function $\mathbf{G}(f)$ in (2). To address the constraint in (2), we resort to the Lagrange duality theory. First, for simplicity, let

$$L(f)=\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[\ell(f(\boldsymbol{x}),\boldsymbol{y})],\quad \text{and}\quad R(f)=L(f)+\frac{\lambda}{2}\|f\|^2_{\mathcal{H}}.$$

Upon defining these quantities, the Lagrangian for problem (2) takes the form

$$\mathcal{L}^o(f,\boldsymbol{\mu})=L(f)+\boldsymbol{\mu}^\top\boldsymbol{G}(f)+\frac{\lambda}{2}\|f\|^2_{\mathcal{H}}, \tag{6}$$

where $\boldsymbol{\mu}=(\mu_1\cdots,\mu_m)^\top$ with each $\mu_j\in\mathbb{R}^+$ being the nonnegative Lagrange multiplier associated with G_j. With the regularization term, the Lagrangian is also strongly convex in f. Assuming that Slater's condition [26] holds in this paper, we have strong duality. Thus, the solution f^* to (2) is equivalent to the primal-dual pair $(f^*,\boldsymbol{\mu}^*)$ that solves the following saddle-point problem

$$(f^*,\boldsymbol{\mu}^*)=\arg\max_{\boldsymbol{\mu}\in\mathbb{R}^m_+}\min_{f\in\mathcal{H}}\quad \mathcal{L}^o(f,\boldsymbol{\mu}) \tag{7}$$

where $\mathbb{R}^m_+=\{\boldsymbol{\mu}:\boldsymbol{\mu}\in\mathbb{R}^m,\mu_j\geq 0,\forall j\in 1,\cdots,m\}$.

In stochastic optimization, however, the expectation over the random pair $(\boldsymbol{x},\boldsymbol{y})$ in $L(f)$ is not easily available. Instead, it is possible to evaluate the empirical estimate of $L(f)$ using a training set $\mathcal{S}=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),\cdots,(\boldsymbol{x}_T,\boldsymbol{y}_T)\}$ with T data samples. The solution to the unconstrained empirical objective is characterized by the well-known Representer Theorem; see e.g., [27], [19]. Specifically, the optimal $f(\boldsymbol{x})$ in $\mathcal{H}$ can be written as an expansion of kernel evaluations *only* at elements of the training set $\{\kappa(\boldsymbol{x}_t,\boldsymbol{x})\}_{t\in[T]}$[1].

To the best of our knowledge, there is no Representer Theorem for constrained problems in RKHS. To generalize it to the constrained setting, we study problems with data-dependent constraints. We assume that the convex constraint function $G_j(f)$ in (2) is also the expectation of some $g_j:\mathcal{H}\times\mathcal{X}\to\mathbb{R}$ over the random vector $\boldsymbol{x}$; i.e., $\boldsymbol{G}(f)=\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{g}(f(\boldsymbol{x}))]$ with $\boldsymbol{g}=(g_1,\cdots,g_m)^\top$. Thus, the empirical counterpart of (7) over the training set $\mathcal{S}$ takes the form

$$\max_{\boldsymbol{\mu}\in\mathbb{R}^m_+}\min_{f\in\mathcal{H}}\quad \mathcal{L}^o(f,\boldsymbol{\mu};\mathcal{S}), \tag{8}$$

where $\mathcal{L}^o(f,\boldsymbol{\mu};\mathcal{S})$ is defined as

$$\mathcal{L}^o(f,\boldsymbol{\mu};\mathcal{S}) =: \frac{1}{T}\sum_{t=1}^{T}\left[\ell(f(\boldsymbol{x}_t),\boldsymbol{y}_t)+\sum_{j=1}^{m}\mu_j g_j(f(\boldsymbol{x}_t))\right]+\frac{\lambda}{2}\|f\|^2_{\mathcal{H}}. \tag{9}$$

[1]Here we use $[T]$ to denote the set of integers $\{1,2,\cdots,T\}$.

This way, the following Representer Theorem can be established for the saddle-point problem (8). Due to space limitation, we do not include the proof of the theorem here; it can be found in [1]. Similarly, detailed proofs for all the other technical results are available in [1].

Theorem 1. *Fix a kernel κ, and let $\mathcal{H}$ be the corresponding RKHS. Let $\mathcal{S}=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),\cdots,(\boldsymbol{x}_T,\boldsymbol{y}_T)\}$ be the training dataset. If the empirical estimate of the function G has the form $G(f;\mathcal{S})=(1/T)\sum_{t=1}^{T}g_j(f(\boldsymbol{x}_t))$, all solutions to the saddle-point problem* (8) *can be expressed as*

$$f^*=\sum_{t=1}^{T}w_t\kappa(\boldsymbol{x}_t,\cdot), \tag{10}$$

where $\{w_t\in\mathbb{R}\}_{t\in[T]}$ are the real-valued coefficients.

Theorem 1 shows that the solution f^* [cf. (2)] admits a representation as a countable linear combination of kernel evaluations of realizations of $(\mathbf{x},\mathbf{y})$. This generalizes the unconstrained stochastic optimization in RKHS [19]. With these technicalities clarified, we may now turn to developing an algorithmic solution to address constrained stochastic optimization in RKHS for the first time.

III. Stochastic Primal-dual Method in RKHS

Now we present an iterative numerical solution for achieving (2). To this end, consider its approximate Lagrangian relaxation:

$$\mathcal{L}(f,\boldsymbol{\mu})=L(f)+\boldsymbol{\mu}^\top\boldsymbol{G}(f)+\frac{\lambda}{2}\|f\|^2_{\mathcal{H}}-\frac{\delta\eta}{2}\|\boldsymbol{\mu}\|^2. \tag{11}$$

Note that (11) is the *augmented Lagrangian* of (2) with regularization coefficients $\delta,\eta>0$ for the dual variable $\boldsymbol{\mu}$. The last regularization term has been included by our algorithmic design in order to control the violation of nonnegative constraints on the dual variable over time t. Note that the augmented Lagrangian $\mathcal{L}(f,\boldsymbol{\mu})$ is strongly convex in f and strongly concave in $\boldsymbol{\mu}$ with positive δ and η. Thus, $\mathcal{L}(f,\boldsymbol{\mu})$ admits a saddle-point $(f^s,\boldsymbol{\mu}^s)$, where f^s can be viewed as an approximation of the optimal f^*. Therefore, the original saddle-point problem in (7) can be addressed by solving the following approximation

$$\max_{\boldsymbol{\mu}\in\mathbb{R}^m_+}\min_{f\in\mathcal{H}}\quad \mathcal{L}(f,\boldsymbol{\mu}). \tag{12}$$

Furthermore, under the assumption that $\boldsymbol{G}(f)=\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{g}(f(\boldsymbol{x}))]$ as in Section II, we obtain the following empirical version of (12) over the training set $\mathcal{S}=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),\cdots,(\boldsymbol{x}_T,\boldsymbol{y}_T)\}$:

$$\max_{\boldsymbol{\mu}\in\mathbb{R}^m_+}\min_{f\in\mathcal{H}}\mathcal{L}(f,\boldsymbol{\mu};\mathcal{S})=\frac{1}{T}\sum_{t=1}^{T}\left[\ell(f(\boldsymbol{x}_t),\boldsymbol{y}_t)+\sum_{j=1}^{m}\mu_j g_j(f(\boldsymbol{x}_t))\right]+\frac{\lambda}{2}\|f\|^2_{\mathcal{H}}-\frac{\delta\eta}{2}\|\boldsymbol{\mu}\|^2. \tag{13}$$

We also define the instantaneous augmented Lagrangian, i.e., one realization of $\mathcal{L}(f,\boldsymbol{\mu})$ based on data sample $(\boldsymbol{x}_t,\boldsymbol{y}_t)$, as

$$\widehat{\mathcal{L}}_t(f,\boldsymbol{\mu}) := \ell(f(\boldsymbol{x}_t),\boldsymbol{y}_t) + \sum_{j=1}^{m}\mu_j g_j(f(\boldsymbol{x}_t)) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 - \frac{\delta\eta}{2}\|\boldsymbol{\mu}\|^2. \quad (14)$$

Observe that (14) is (11) in expectation over $(\mathbf{x},\mathbf{y})$. Our algorithm, detailed soon, is developed on basis of the gradient update for solving the augmented saddle-point problem (12).

A. Functional Primal-dual Method

We are interested in the online setting where the data sample size T is not necessarily finite or samples $(\boldsymbol{x}_t,\boldsymbol{y}_t)$ are observed sequentially. To this end, we consider the case where $(\boldsymbol{x}_t,\boldsymbol{y}_t)$ are independent realizations from a stationary joint distribution of the random pair $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{X}\times\mathcal{Y}$ [28]. First, define the following function:

$$\widetilde{\ell}(f(\boldsymbol{x}),\boldsymbol{y},\boldsymbol{\mu}) := \ell(f(\boldsymbol{x}),\boldsymbol{y}) + \sum_{j=1}^{m}\mu_j g_j(f(\boldsymbol{x})).$$

By the reproducing property of RKHS in (1), we have

$$\frac{\partial f(\boldsymbol{x}_t)}{\partial f} = \frac{\partial\langle f,\kappa(\boldsymbol{x}_t,\cdot)\rangle_{\mathcal{H}}}{\partial f} = \kappa(\boldsymbol{x}_t,\cdot). \quad (15)$$

Thus, following the derivation in [29], we can compute the stochastic gradient of $\widetilde{\ell}$ w.r.t. f in RKHS by using the chain rule. For any given $\boldsymbol{\mu}\in\mathbb{R}_+^m$, we have

$$\begin{aligned}\nabla_f\widetilde{\ell}(f(\boldsymbol{x}_t),\boldsymbol{y}_t,\boldsymbol{\mu})(\cdot) &= \frac{\partial\widetilde{\ell}(f(\boldsymbol{x}_t),\boldsymbol{y}_t,\boldsymbol{\mu})}{\partial f(\boldsymbol{x}_t)}\frac{\partial f(\boldsymbol{x}_t)}{\partial f}(\cdot)\\ &= \widetilde{\ell}'(f(\boldsymbol{x}_t),\boldsymbol{y}_t,\boldsymbol{\mu})\kappa(\boldsymbol{x}_t,\cdot,) \quad (16)\end{aligned}$$

where we denote

$$\widetilde{\ell}'(f(\boldsymbol{x}_t),\boldsymbol{y}_t,\boldsymbol{\mu}) := \partial\widetilde{\ell}(f(\boldsymbol{x}_t),\boldsymbol{y}_t,\boldsymbol{\mu})/\partial f(\boldsymbol{x}_t)$$

as the derivative of $\widetilde{\ell}(f(\boldsymbol{x}_t),\boldsymbol{y}_t,\boldsymbol{\mu})$ w.r.t. its argument $f(\boldsymbol{x}_t)$ evaluated at $\boldsymbol{x}_t$. Note that by definition the derivative $\widetilde{\ell}'(f(\boldsymbol{x}_t),\boldsymbol{y}_t,\boldsymbol{\mu})$ has the form

$$\widetilde{\ell}'(f(\boldsymbol{x}_t),\boldsymbol{y}_t,\boldsymbol{\mu}) = \ell'(f(\boldsymbol{x}_t),\boldsymbol{y}_t) + \sum_{j=1}^{m}\mu_j g_j'(f(\boldsymbol{x}_t)),$$

where ℓ' and g_j' denote the derivative w.r.t. the scalar $f(\boldsymbol{x}_t)$ evaluated at $\boldsymbol{x}_t$, respectively. With these definitions, we propose the stochastic primal-dual method for solving (12):

$$\begin{cases} f_{t+1} = (1-\eta\lambda)f_t - \eta\Big[\ell'(f_t(\boldsymbol{x}_t),\boldsymbol{y}_t) + \sum_{j=1}^{m}\mu_j g_j'(f_t(\boldsymbol{x}_t))\Big]\kappa(\boldsymbol{x}_t,\cdot), & (17a)\\ \boldsymbol{\mu}_{t+1} = \big[(1-\eta^2\delta)\boldsymbol{\mu}_t + \eta\boldsymbol{g}(f_t(\boldsymbol{x}_t))\big]_+, & (17b)\end{cases}$$

where $\eta>0$ is a step-size parameter which can be selected as a small constant, and $[\cdot]_+ = \max(\cdot,0)$ denotes the vector-operator that projects its argument to $\mathbb{R}_+^m$. Recall that the step-size η is also used to define the augmented Lagrangian (11). This way, one can relate the control of the constraint violation of the dual variable with the algorithm update, as we will show later in the proof.

For given regularization parameter $\lambda>0$ in (2), we further require the step-size $\eta<1/\lambda$ and the sequence of $(f_t,\boldsymbol{\mu}_t)$ is initialized by $f_1=0\in\mathcal{H}$ and $\boldsymbol{\mu}_1=\mathbf{0}\in\mathbb{R}_+^m$. By induction, the updates in (17) lead to a linear expansion form for f_t in terms of all past observed feature vectors $\boldsymbol{x}_t$, as given by

$$f_t(\boldsymbol{x}) = \sum_{t=1}^{t-1} w_t\kappa(\boldsymbol{x}_t,\boldsymbol{x}) = \boldsymbol{w}_t^\top\boldsymbol{\kappa}_{\boldsymbol{X}_t}(\boldsymbol{x}), \quad (18)$$

where the coefficient $\boldsymbol{w}_t := [w_1,\cdots,w_t]^\top\in\mathbb{R}^{t-1}$, and

$$\boldsymbol{X}_t := [\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{t-1}]\in\mathbb{R}^{p\times(t-1)},$$
$$\boldsymbol{\kappa}_{\boldsymbol{X}_t}(\cdot) := [\kappa(\boldsymbol{x}_1,\cdot),\ \ldots\ ,\kappa(\boldsymbol{x}_{t-1},\cdot)]^\top.$$

Thus, function f_t belongs to the subspace spanned by $\boldsymbol{\kappa}_{\boldsymbol{X}_t}(\cdot)$. Performing the update of f_t as in (17a) amounts to the following parametric updates on both kernel dictionary $\boldsymbol{X}$ and coefficient vector $\boldsymbol{w}$:

$$\boldsymbol{X}_{t+1} = [\boldsymbol{X}_t,\ \ \boldsymbol{x}_t],$$
$$\boldsymbol{w}_{t+1} = \left[(1-\eta\lambda)\boldsymbol{w}_t, -\eta\ell'(f_t(\boldsymbol{x}_t),\boldsymbol{y}_t) - \eta\sum_{j=1}^{m}\mu_j g_j'(f_t(\boldsymbol{x}_t))\right].$$

Hence, the number of columns in $\boldsymbol{X}_t$ increases by one at every time, which inevitably leads to dimensionality concern. Specifically, define the *model order* as number of data points M_t in the dictionary at time t. For the stochastic gradient update in (17a), the order $M_t = t-1$ grows unbounded with iteration index t.

Proximal Projection Motivated by the dimensionality reduction approach in [12], we propose to project the functional stochastic gradient update of f_t onto a low-dimensional subspace $\mathcal{H}_{\boldsymbol{D}_t}\subseteq\mathcal{H}$. The latter consist only of functions that can be represented using some dictionary $\boldsymbol{D}_t = [\boldsymbol{d}_1,\cdots,\boldsymbol{d}_{M_t}]\in\mathbb{R}^{p\times M_t}$ of fixed size M_t. Recall that M_t is the model order of the dictionary at time t. In particular, $\mathcal{H}_{\boldsymbol{D}_t}$ has the form $\mathcal{H}_{\boldsymbol{D}_t} = \{f : f(\cdot) = \sum_{\tau=1}^{M_t} w_\tau\kappa(\boldsymbol{d}_\tau,\cdot) = \boldsymbol{w}_t^\top\boldsymbol{\kappa}_{\boldsymbol{D}_t}(\cdot)\}$, where we define $\boldsymbol{\kappa}_{\boldsymbol{D}_t}(\cdot) = [\kappa(\boldsymbol{d}_1,\cdot)\ldots\kappa(\boldsymbol{d}_{M_t},\cdot)]$. The dictionary $\boldsymbol{D}_t$ is updated to $\boldsymbol{D}_{t+1}$ when a new data sample $(\boldsymbol{x}_t,\boldsymbol{y}_t)$ becomes available. Therefore, we replace the update (17a) with the following one that has an additional projection operation onto the subspace $\mathcal{H}_{\boldsymbol{D}_{t+1}}$:

$$\begin{aligned} f_{t+1} &= \operatorname*{argmin}_{f\in\mathcal{H}_{\boldsymbol{D}_{t+1}}}\Big\|f - \Big((1-\eta\lambda)f_t - \eta\nabla_f\widetilde{\ell}(f_t(\boldsymbol{x}_t),\boldsymbol{y}_t,\boldsymbol{\mu}_t)\Big)\Big\|_{\mathcal{H}}^2 \quad (19)\\ &:= \mathcal{P}_{\mathcal{H}_{\boldsymbol{D}_{t+1}}}\Big[(1-\eta\lambda)f_t - \eta\nabla_f\widetilde{\ell}(f_t(\boldsymbol{x}_t),\boldsymbol{y}_t,\boldsymbol{\mu}_t)\Big],\end{aligned}$$

where $\mathcal{P}_{\mathcal{H}_{\boldsymbol{D}}}$ denotes the projection operator onto any subspace $\mathcal{H}_{\boldsymbol{D}}\subseteq\mathcal{H}$.

To project the function onto $\mathcal{H}_{\boldsymbol{D}_{t+1}}$, we first form the original dictionary $\widetilde{\boldsymbol{D}}_{t+1}$ and coefficient vector $\widetilde{\boldsymbol{w}}_{t+1}$ as defined by the update (17):

$$\widetilde{\boldsymbol{D}}_{t+1} = [\boldsymbol{D}_t,\ \ \boldsymbol{x}_t], \quad (20)$$
$$\widetilde{\boldsymbol{w}}_{t+1} = [(1-\eta\lambda)\boldsymbol{w}_t,\ \ -\eta\widetilde{\ell}'(f_t(\boldsymbol{x}_t),\boldsymbol{y}_t,\boldsymbol{\mu}_t)]\ .$$

Accordingly, we denote the pre-projected function sequence as $\widetilde{f}_{t+1} = (1-\eta\lambda)f_t - \eta\nabla_f \widetilde{\ell}(f_t(\boldsymbol{x}_t), \boldsymbol{y}_t, \boldsymbol{\mu}_t)$. For any given dictionary $\boldsymbol{D}_{t+1}$, the projection of $\widetilde{f}_{t+1}$ onto $\mathcal{H}_{\boldsymbol{D}_{t+1}}$ is equivalent to updating the coefficient vector $\boldsymbol{w}_{t+1}$ as

$$\boldsymbol{w}_{t+1} = \boldsymbol{K}^{-1}_{\boldsymbol{D}_{t+1}\boldsymbol{D}_{t+1}} \boldsymbol{K}_{\boldsymbol{D}_{t+1}\widetilde{\boldsymbol{D}}_{t+1}} \widetilde{\boldsymbol{w}}_{t+1} \ , \tag{21}$$

where $\boldsymbol{K}_{\boldsymbol{D}_{t+1}\boldsymbol{D}_{t+1}}$ and $\boldsymbol{K}_{\boldsymbol{D}_{t+1}\widetilde{\boldsymbol{D}}_{t+1}}$ are kernel matrices between the respective pair of dictionaries. One efficient way to obtain the dictionary $\boldsymbol{D}_{t+1}$ from $\widetilde{\boldsymbol{D}}_{t+1}$, as well as the coefficient $\boldsymbol{w}_{t+1}$, is to apply the approach of *kernel orthogonal matching pursuit* (KOMP) [30]. Moreover, we assume that the output f_{t+1} from the KOMP has bounded Hilbert norm,[2] which is typically required for establishing the convergence of primal-dual methods; see e.g., [31], [21], [32]. Hence, the following projection could control not only the model order but also the Hilbert norm of the output,

$$(f_{t+1}, \boldsymbol{D}_{t+1}, \boldsymbol{w}_{t+1}) = \textbf{KOMP}(\widetilde{f}_{t+1}, \widetilde{\boldsymbol{D}}_{t+1}, \widetilde{\boldsymbol{w}}_{t+1}, \epsilon_t), \tag{22}$$

where ϵ_t is the approximation budget such that $\|f_{t+1} - \widetilde{f}_{t+1}\|_{\mathcal{H}} \leq \epsilon_t$. Note that the dual variable $\boldsymbol{\mu}_t$ shows up in the coefficient vector $\widetilde{\boldsymbol{w}}_{t+1}$. To recap, the online primal-dual algorithm is updated as follows

$$\begin{cases} f_{t+1} = \mathcal{P}_{\mathcal{H}_{\boldsymbol{D}_{t+1}}}\Big[(1-\eta\lambda)f_t \\ \qquad\qquad - \eta\nabla_f \widetilde{\ell}(f_t(\boldsymbol{x}_t), \boldsymbol{y}_t, \boldsymbol{\mu}_t)\Big] & \text{(23a)} \\ \boldsymbol{\mu}_{t+1} = \big[(1-\eta^2\delta)\boldsymbol{\mu}_t + \eta \boldsymbol{g}(f_t(\boldsymbol{x}_t))\big]_+. & \text{(23b)} \end{cases}$$

Given sequentially observed realization $(\mathbf{x}_t, \mathbf{y}_t)$, the algorithm alternates between stochastic primal descent steps (20) and dual stochastic ascent steps (23b). The primal iterates are projected onto sparse stochastic subspaces defined by the output of matching pursuit (22). The update rule of the projected primal-dual method is summarized as Algorithm 1. The subsequent sections present the theoretical and experimental validation of update (23) for function-valued constrained stochastic programming.

IV. Convergence Analysis

We will establish that the proposed algorithm, a functional generalization of projected stochastic primal-dual method, achieves convergence in expectation in terms of both objective sub-optimality and constraint violation. We start by introducing several standard assumptions for the necessity of convergence analysis.

Assumption 1. *The feature space $\mathcal{X} \subset \mathbb{R}^p$ and target domain $\mathcal{Y} \subset \mathbb{R}^d$ are compact, and the reproducing kernel map can be bounded by some constant $X > 0$ as*

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \sqrt{\kappa(\boldsymbol{x},\boldsymbol{x})} = X < \infty \tag{24}$$

[2]Note that this assumption can be satisfied by imposing an additional bounded-norm constraint in the optimization problem for finding the best set of bases in the matching pursuit algorithm, e.g., in Eq. (7) in [30], which can be achieved by thresholding the coefficient sequence during compression.

Algorithm 1 Projected Primal-Dual Method in Kernel Space

Require: $\{\mathbf{x}_t, \mathbf{y}_t, \epsilon_t, \eta, \delta\}_{t=0,1,2,\ldots}$

initialize $f_0(\cdot) = 0, \mathbf{D}_0 = [], \mathbf{w}_0 = [], \boldsymbol{\lambda} = \mathbf{0}$; i.e., initial dictionary is null.

for $t = 0, 1, 2, \ldots$ **do**

Observe training example $(\mathbf{x}_t, y_t)$

Take stochastic descent step on Lagrangian [cf. (17a)]

$$\widetilde{f}_{t+1} = (1-\eta\lambda)f_t - \eta\Bigg[\ell'(f_t(\boldsymbol{x}_t), \boldsymbol{y}_t) + \sum_{j=1}^{m} \mu_j g_j'(f_t(\boldsymbol{x}_t))\Bigg]\kappa(\boldsymbol{x}_t, \cdot)$$

Take stochastic ascent step on Lagrangian [cf. (17b)]

$$\boldsymbol{\mu}_{t+1} = \big[(1-\eta^2\delta)\boldsymbol{\mu}_t + \eta \boldsymbol{g}(f_t(\boldsymbol{x}_t))\big]_+$$

Update $\widetilde{\boldsymbol{D}}_{t+1} = [\boldsymbol{D}_t, \mathbf{x}_t]$ and $\widetilde{\boldsymbol{w}}_{t+1}$ [cf. (20)]

Greedily compress function using KOMP

$(f_{t+1}, \boldsymbol{D}_{t+1}, \boldsymbol{w}_{t+1}) = \textbf{KOMP}(\widetilde{f}_{t+1}, \widetilde{\boldsymbol{D}}_{t+1}, \widetilde{\boldsymbol{w}}_{t+1}, \epsilon_t)$

end loop

end for

Assumption 2. *The instantaneous loss $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is uniformly C_1-Lipschitz continuous in its first (scalar) argument for any fixed $\boldsymbol{y} \in \mathcal{Y}$, and the constraint functions $g_i : \mathcal{H} \times \mathcal{X} \to \mathbb{R}$ for all $i = 1, \cdots, m$ are all uniformly C_2-Lipschitz continuous; i.e., for any $z, z' \in \mathbb{R}$, there exist constants $C_1, C_2 > 0$ such that*

$$|\ell(z, \boldsymbol{y}) - \ell(z', \boldsymbol{y})| \leq C_1|z - z'|, \forall \boldsymbol{y} \in \mathcal{Y}, \tag{25}$$

$$|g_i(z) - g_i(z')| \leq C_2|z - z'|, \forall i = 1, \cdots, m. \tag{26}$$

Assumption 3. *The loss $\ell(f(\boldsymbol{x}), \boldsymbol{y})$ and the constraints functions $g_i(f(\boldsymbol{x}))$ for $i = 1, \cdots, m$ are all convex with respect to the argument $f(\boldsymbol{x})$ on $\mathbb{R}$ for all $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{y} \in \mathcal{Y}$.*

Assumption 4. *There exists a strictly feasible point, i.e., some $f \in \mathcal{H}$ satisfies $\boldsymbol{G}(f) < \mathbf{0}$.*

Assumption 5. *The output f_{t+1} of the KOMP update (22) has Hilbert norm bounded by $R_{\mathcal{B}} < \infty$, and the optimal f^* lies in the ball $\mathcal{B}$ with radius $R_{\mathcal{B}}$.*

Assumptions 1 and 2 hold in most practical settings by the data domain itself. Assumption 3 ensures that the constrained stochastic optimization problem (2) is convex. Assumption 4, namely the Slater's Condition [26], ensures the satisfiability of the constraints and thus the feasible set of (2) is non-empty. Moreover, it guarantees that the strong duality holds for (2). Assumption 5 formally states that the KOMP output has bounded Hilbert norm, as mentioned in Section III. In addition, it assumes that the optimal f^* belongs to the ball $\mathcal{B}$ with radius $R_{\mathcal{B}}$ such that the algorithm output and the set of optimizers have non-empty intersection.

With the technical setting clarified, we can establish

bounds on the primal-sub-optimality $\{R(f_t) - R(f^*)\}$ and the accumulated constraint violation $\mathbf{G}(f_t)$, both in expectation. The following lemma is needed to bound the decrement of the instantaneous Lagrangian difference, namely $\widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}) - \widehat{\mathcal{L}}_t(f, \boldsymbol{\mu}_t)$.

Lemma 1. *Suppose the sequence $(f_t, \boldsymbol{\mu}_t)$ is generated from the update* (23)*, and Assumptions 1-5 hold. Then, the instantaneous Lagrangian difference satisfies the following decrement property for any $f \in \mathcal{B}$ and $\boldsymbol{\mu} \in \mathbb{R}^m_+$:*

$$\begin{aligned}
&\widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}) - \widehat{\mathcal{L}}_t(f, \boldsymbol{\mu}_t) \qquad (27)\\
\leq &\frac{1}{2\eta}\big(\|f_t - f\|^2_{\mathcal{H}} - \|f_{t+1} - f\|^2_{\mathcal{H}} + \|\boldsymbol{\mu}_t - \boldsymbol{\mu}\|^2\\
&- \|\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}\|^2\big) + \frac{\eta}{2}\big(2 \cdot \big\|\nabla_f \widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)\big\|^2_{\mathcal{H}}\\
&+ \big\|\nabla_{\boldsymbol{\mu}} \widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)\big\|^2\big) + \frac{\epsilon_t}{\eta}\|f_t - f\|_{\mathcal{H}} + \frac{\epsilon_t^2}{\eta},
\end{aligned}$$

where ϵ_t is the approximation budget specified in (22)*.*

Lemma 1 shows that the difference of instantaneous Lagrangians can be upper bounded in terms of the difference between the primal and dual iterates to a fixed primal-dual pair $(f, \boldsymbol{\mu}) \in \mathcal{B} \times \mathbb{R}^m_+$ over two consecutive iterations, the squared norm of primal and dual gradients, and the terms related to the approximation budget ϵ_t. This is the basis for the ensuing convergence analysis with certain choice of constant step-size η.

Now we are ready to present the convergence result for the proposed reduced-order online algorithm.

Theorem 2. *Let $(f_t, \boldsymbol{\mu}_t)$ be the sequence generated from Algorithm 1. Suppose Assumptions 1-5 hold with constant step-size $\eta = 1/\sqrt{T}$ and approximation budget $\epsilon_t = \epsilon = P\eta^2$, where $P > 0$ is a scalar termed as the parsimony constant. Then, the time-aggregation of the expected objective function error sequence $\mathbb{E}[R(f_t)] - R(f^*)$, with f^* defined as in* (2)*, grows sub-linearly with termination index T:*

$$\sum_{t=1}^{T} \mathbb{E}[R(f_t) - R(f^*)] \leq \mathcal{O}(\sqrt{T}). \qquad (28)$$

Moreover, the time-aggregation of the expected violation of all algorithm constraints grows sub-linearly in T as

$$\sum_{j=1}^{m} \mathbb{E}\bigg[\sum_{t=1}^{T} G_j(f_t)\bigg]_+ \leq \mathcal{O}(T^{3/4}). \qquad (29)$$

Theorem 2 establishes that given a fixed step-size $\eta = 1/\sqrt{T}$, the objective function error accumulates at a sub-linear rate of $\mathcal{O}(\sqrt{T})$ over time as does the constraint violation at a rate of $\mathcal{O}(T^{3/4})$. Thus, for large enough T, both the objective function error and the constraint violation vanish to zero on average. Theorem 2 also allows us to establish the convergence of the time-average iterates to a certain accuracy depending on the total number of iterations T, as stated formally in the following corollary.

Corollary 1. *Let $\overline{f}_T =: \sum_{t=1}^{T} f_t/T$ be the functional formed by averaging the primal iterates f_t over time $t = 1, \cdots, T$. Suppose Assumptions 1-5 hold. With T iterations for Algorithm 1 under a constant step-size $\eta = 1/\sqrt{T}$ and approximation budget $\epsilon_t = \epsilon = P/T$, where $P > 0$ is a fixed constant, the objective function evaluated at $\overline{f}_T$ satisfies*

$$\mathbb{E}[R(\overline{f}_T) - R(f^*)] \leq \mathcal{O}(1/\sqrt{T}). \qquad (30)$$

In addition, the constraint violation evaluated at $\overline{f}_T$ satisfies

$$\sum_{j=1}^{m} \mathbb{E}\Big[\big(G_j(\overline{f}_T)\big)\Big]_+ \leq \mathcal{O}(T^{-1/4}). \qquad (31)$$

Corollary 1 shows that the time-average iterate $\overline{f}_T$ achieves a convergence rate at $\mathcal{O}(1/\sqrt{T})$ for the objective function value, and an $\mathcal{O}(T^{-1/4})$ rate to bound the constraint violation. We note that for any fixed T, this result essentially shows the convergence to a neighborhood of the actual solution on the average. The size of this neighborhood depends on the parameters of the problem, including the radius of the ball $R_{\mathcal{B}}$, the coefficient δ, the Lipschitz constants for ℓ and g_i, and the upper bound for the reproducing kernel map X. We also note that the results in Theorem 2 and Corollary 1 are comparable to those under the deterministic setting [31] or the stochastic setting in the vector-valued constrained convex optimization [32]. One departing feature of the RKHS setting is that by averaging f_t over time, its model order may blow up; thus, Corollary 1 is a theoretical result only for interpreting the convergence bounds of Theorem 2, although such averaging over time may violate the sparsity of the instantaneous function iterate.

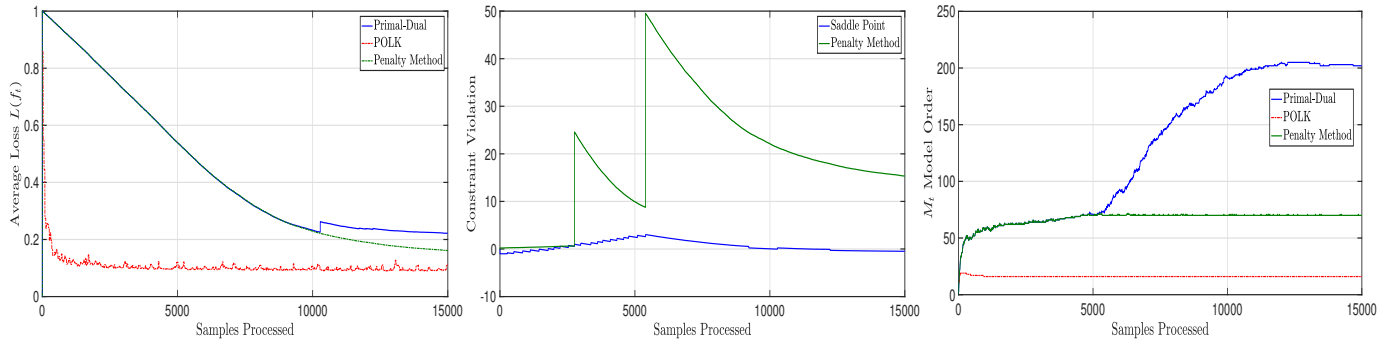
An additional benefit of using constant step-sizes for a fixed $T < \infty$ is that we may limit the complexity of the primal function sequence and establish that it is at-worst finite. Specifically, with constant step-size and approximation budget, we could apply Theorem 3 in [12] using a slight modification that $\epsilon = \mathcal{O}(\eta^2)$ rather than $\mathcal{O}(\eta^{3/2})$. This result guarantees that the model order of the function sequence remains finite and is related to the covering number of the data domain, which is formally stated here as a corollary.

Corollary 2. *Suppose the sequence $(f_t, \boldsymbol{\mu}_t)$ is generated by Algorithm 1 under constant step-size $\eta = 1/\sqrt{T}$ and approximation budget $\epsilon = P\eta^2$ where $P > 0$ is a positive scalar. For the model order M_t of function f_t, there exists a finite upper bound M^∞ such that $M_t \leq M^\infty$ for all $t \geq 0$.*

Henceforth, the update (23) solves the problem (2) to a bounded error neighborhood that is dependent on final iteration and step-size, and ensures that the function complexity is under control. Subsequently, we substantiate these theoretical findings on a canonical statistical inference task with risk-constrained kernel classifiers.

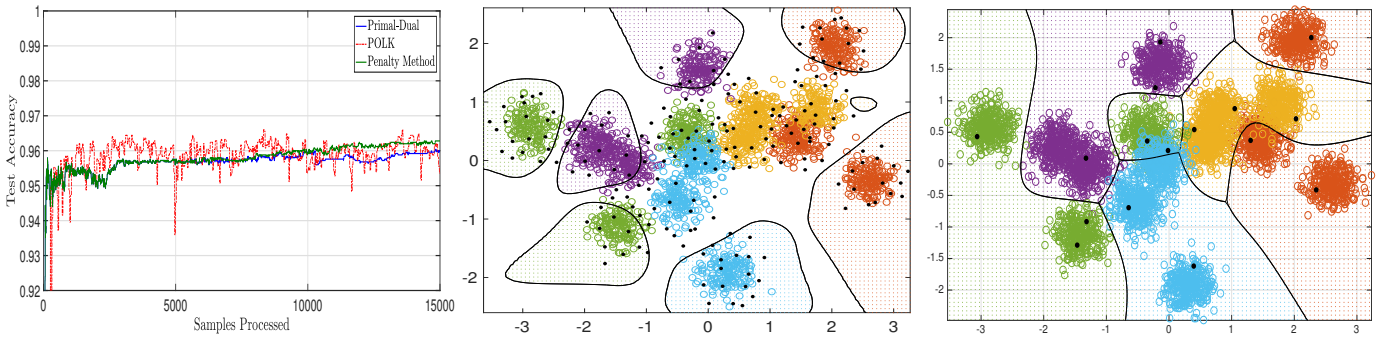
V. Experiments

This section presents numerical evaluation results for our proposed method for solving constrained stochastic optimization problems in RKHS. In particular, we consider the risk-aware supervised learning problem with conditional value-at-risk constraints as stated in Example 1. This constraint is



(a) Objective v.s. Samples (b) Infeasibility v.s. Samples (c) Model Order M_t v.s. Samples

Fig. 1: Algorithm 1 for KSVM with objective in (32) and CVaR constraint in (3) (cf. Example 1) for three training epochs over a multi-class problem with synthetic Gaussian mixture data. We use a Gaussian kernel with bandwidth $\sigma = 0.3$, constant step-size $\eta = 0.009$, with parsimony constant $P = 3.7$, and a mini-batch size of 4. Spikes are due to non-differentiability of the objective and constraint. Smaller step-sizes are required for constrained versus unconstrained problems. The objective and constraint violation converge to null and the model order remains stable. We compare with an unconstrained projected functional stochastic gradient descent (FSGD) based method, namely, POLK [12], and a penalty method [16] with the penalty coefficient doubled every 200 iterations. These two alternatives converge to lower model-order, yet *infeasible*, solutions.



(a) Test Set Accuracy (b) Risk-Aware Decision Surface (c) Unconstrained Decision Surface

Fig. 2: Algorithm 1 for KSVM with objective in (32) and CVaR constraint in (3) (cf. Example 1). Fig. 2(a) shows that the test set accuracy stabilizes to near a 4% error rate; Fig. 2(b) displays the decision surface, where bold black dots denote kernel dictionary elements, grid colors denote classifier decisions. Each class label is assigned with a distinct color, and curved lines delineate confident decision boundaries. As shown in Fig. 2(c), high-confidence decision boundaries are only drawn far from class overlap, which is the expected effect of minimizing CVaR of a classifier. This is despite the fact that points in the overlap region are still classified correctly. For comparison, we also display the surface learned by POLK, which does not incorporate risk into decision making and thus is closer to the mean data density function.

imposed to mitigate the unknown variance of the modeling hypothesis that $f \in \mathcal{H}$, known as the approximation error in statistical learning [22]. The supervised learning problem we consider here is *Multi-class Kernel Support Vector Machines (KSVM)*. In KSVM, the merit of a function is defined by its ability to maximize the so-termed classification margin. In particular, define a set of C class-specific activation functions $f_c : \mathcal{X} \to \mathbb{R}$, and denote them jointly as $\mathbf{f} \in \mathcal{H}^C$. In Multi-KSVM, points are assigned the class label of the activation function that yields the maximum response. KSVM is trained by taking the instantaneous loss ℓ to be the multi-class hinge function which defines the margin separating hyperplane in the kernelized feature space, as given by

$$\ell(\mathbf{f}(\mathbf{x}_n), y_n) = [1 + f_r(\mathbf{x}_n) - f_{y_n}(\mathbf{x}_n)]_+ + \lambda \sum_{c'=1}^{C} \|f_{c'}\|_{\mathcal{H}}^2, \quad (32)$$

where $r = \operatorname{argmax}_{c' \neq y_n} f_{c'}(\mathbf{x})$. Please refer to [33] for more details on the problem formulation. We test our algorithm on a synthetic data set, where data vectors are 2-dimensional drawn from a set of Gaussian mixture models as in [34]. Specifically, first draw the label y_n randomly and uniformly from the label set. Based on y_n, the data point $\mathbf{x}_n$ is then drawn from an equitably-weighted Gaussian mixture model as $\mathbf{x} \,|\, y \sim (1/3) \sum_{j=1}^{3} \mathcal{N}(\boldsymbol{\mu}_{y,j}, \sigma_{y,j}^2 \mathbf{I})$ where $\sigma_{y,j}^2 = 0.2$ for all values of y and j. This way, $\boldsymbol{\mu}_{y,j}$ are realizations a distinct Gaussian distribution with class-dependent parameters, i.e., $\boldsymbol{\mu}_{y,j} \sim \mathcal{N}(\boldsymbol{\theta}_y, \sigma_y^2 \mathbf{I})$, where $\sigma_y^2 = 1.0$ and $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_C\}$ are equitably spaced around the unit circle, one for each class label. The number of classes is fixed at $C = 5$ so the feature distribution has 15 distinct modes. The data set includes $N = 5000$ samples for training and 2500 for testing.

We run the algorithm for three training epochs, i.e., $T = 15000$, with a Gaussian kernel whose bandwidth is $\sigma = 0.3$. The algorithm step-size is $\eta = 0.009$, under the compression budget $\epsilon = P\eta^2$ with parsimony constant $P = 3.7$ and a mini-batch size of 4. The primal regularizer $\lambda = 10^{-4}$ and the dual regularizer $\delta = 10^{-4}$. The significance level for CVaR_α is $\alpha = 0.9$ and the tolerance is $\gamma = 2$. This enforces the learning process to be more conservative and avoid

moving the regression function in ways that could cause the loss function to spike with probability under $1 - \alpha = 0.1$.

The results of this experiment are given in Fig. 1: the statistical average loss converges to a small constant as the number of samples increases (Fig. 1a), while the infeasibility initially spikes and gradually settles to feasibility (Fig. 1b). In the meanwhile, the model complexity remains under control (Fig. 1c). Jumps in objective and constraints are caused by the non-differentiability of the hinge loss. The resulting classifier attains test accuracy near 96% by the end of the second training epoch (Fig. 2a), and the resulting risk-aware decision surface is given in Fig. 2b. Bold black dots denote kernel dictionary elements; grid colors denote classifier decisions and curved lines denote decision boundaries, which are far from areas of class overlap due to their increased likelihood of causing loss spikes. Alternative methods based on projected FSGD and a penalty method converge to comparably accurate solutions but cannot handle constraints. They eventually lead to infeasible solutions, and thus define riskier decisions.

VI. Conclusions

This paper considered the function-valued stochastic optimization problem in reproducing Kernel Hilbert space (RKHS) under nonlinear constraints. We extended the Representer Theorem to a certain form of saddle-point problems over RKHS and developed a projected stochastic primal-dual method. We then established convergence in expectation for both the objective function and constraint violation level with bounded error neighborhoods. The convergence results have been numerically validated using a risk-aware supervised learning task. As for future work, we plan to investigate the effectiveness of our algorithm for trajectory optimization based on sensory observations, among other constrained learning applications.

References

[1] K. Zhang, H. Zhu, T. Başar, and A. Koppel, "Projected stochastic primal-dual method for constrained online learning with kernels," *Technical Report*, 2018. [Online]. Available: http://koppel.bitballoon.com/assets/papers/2018_kernel_primal_dual_report.pdf

[2] Z. Marinho, B. Boots, A. D. Dragan, A. Byravan, G. J. Gordon, and S. Srinivasa, "Functional gradient motion planning in reproducing kernel Hilbert spaces," in *Robotics: Science and Systems XII*.

[3] A. Ribeiro, "Ergodic stochastic optimization algorithms for wireless communication and networking," *IEEE Transactions on Signal Processing*, vol. 58, no. 12, pp. 6369–6386, 2010.

[4] H. Tanizaki, *Nonlinear Filters: Estimation and Applications*. Springer Science & Business Media, 2013.

[5] I. M. Gelfand, R. A. Silverman *et al.*, *Calculus of Variations*. Courier Corporation, 2000.

[6] C. Bailey, "Hamilton's principle and the calculus of variations," *Acta Mechanica*, vol. 44, no. 1-2, pp. 49–57, 1982.

[7] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[8] Z. Jarvis-Wloszek, R. Feeley, W. Tan, K. Sun, and A. Packard, "Control applications of sum of squares programming," in *Positive Polynomials in Control*. Springer, pp. 3–22.

[9] C. E. Rasmussen, "Gaussian processes in machine learning," in *Advanced Lectures on Machine Learning*. Springer, 2004, pp. 63–71.

[10] S. Haykin, "Neural networks: A comprehensive foundation," 1994.

[11] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[12] A. Koppel, G. Warnell, E. Stump, and A. Ribeiro, "Parsimonious online learning with kernels via sparse projections in function space," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4671–4675.

[13] J. Hensman, N. Fusi, and N. D. Lawrence, "Gaussian processes for big data," in *Uncertainty in Artificial Intelligence*. Citeseer, 2013, p. 282.

[14] I. Safran and O. Shamir, "Spurious local minima are common in two-layer ReLU neural networks," *arXiv preprint arXiv:1712.08968*, 2017.

[15] J. A. Bagnell and A.-m. Farahmand, "Learning positive functions in a Hilbert space," in *NIPS Workshop on Optimization (OPT2015)*, 2015.

[16] A. Koppel, S. Paternain, C. Richard, and A. Ribeiro, "Decentralized efficient nonparametric stochastic optimization," in *Signal and Information Processing (GlobalSIP), 2017 IEEE Global Conference on (to appear)*. IEEE, 2017.

[17] S. Paternain, D. E. Koditschek, and A. Ribeiro, "Navigation functions for convex potentials in a space with convex obstacles," *IEEE Transactions on Automatic Control*, 2017.

[18] S. Ahmed, "Convexity and decomposition of mean-risk stochastic programs," *Mathematical Programming*, vol. 106, no. 3, pp. 433–446, 2006.

[19] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," *Subseries of Lecture Notes in Computer Science Edited by JG Carbonell and J. Siekmann*, p. 416, 2001.

[20] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[21] M. Mahdavi, R. Jin, and T. Yang, "Trading regret for efficiency: Online convex optimization with long term constraints," *Journal of Machine Learning Research*, vol. 13, no. Sep, pp. 2503–2528, 2012.

[22] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*. Springer Series in Statistics New York, 2001, vol. 1.

[23] R. K. Arora, *Optimization: Algorithms and Applications*. CRC Press, 2015.

[24] A. Nemirovski and A. Shapiro, "Convex approximations of chance constrained programs," *SIAM Journal on Optimization*, vol. 17, no. 4, pp. 969–996, 2006.

[25] K. Slavakis, S. Theodoridis, and I. Yamada, "Adaptive constrained learning in reproducing kernel Hilbert spaces: the robust beamforming case," *IEEE Transactions on Signal Processing*, vol. 57, no. 12, pp. 4744–4764, 2009.

[26] S. Boyd and L. Vanderberghe, *Convex Programming*. New York, NY: Wiley, 2004.

[27] G. Kimeldorf and G. Wahba, "Some results on tchebycheffian spline functions," *Journal of Mathematical Analysis and Applications*, vol. 33, no. 1, pp. 82–95, 1971.

[28] K. Slavakis, P. Bouboulis, and S. Theodoridis, "Online learning in reproducing kernel Hilbert spaces," *Signal Processing Theory and Machine Learning*, pp. 883–987, 2013.

[29] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE Transactions on Signal Processing*, vol. 52, pp. 2165–2176, August 2004.

[30] P. Vincent and Y. Bengio, "Kernel matching pursuit," *Machine Learning*, vol. 48, no. 1, pp. 165–187, 2002.

[31] A. Nedić and A. Ozdaglar, "Subgradient methods for saddle-point problems," *Journal of Optimization Theory and Applications*, vol. 142, no. 1, pp. 205–228, 2009.

[32] A. Koppel, B. M. Sadler, and A. Ribeiro, "Proximity without consensus in online multi-agent optimization," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 6089–6104, 2016.

[33] K. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT press, 2012.

[34] J. Zhu and T. Hastie, "Kernel logistic regression and the import vector machine," *Journal of Computational and Graphical Statistics*, vol. 14, no. 1, pp. 185–205, 2005.