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

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

***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. Then, we develop a primal-dual method which executes alternating projected primal/dual stochastic gradient descent/ascent on the dual-augmented Lagrangian of the 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 in both primal objective sub-optimality and constraint violation, to respective $\mathcal{O}(\sqrt{T})$ and $\mathcal{O}(T^{3/4})$ neighborhoods. Here T is the final iteration index and the constant step-size is chosen as $1/\sqrt{T}$ with $1/T$ approximation budget. Finally, we demonstrate experimentally the effectiveness of the proposed method for risk-aware supervised learning.**

I. Introduction

Kernelized online learning arises in a variety of applications where the decision variable is a function rather than a vector. It is typically cast as an *unconstrained* stochastic optimization problem that aims to minimize the expectation of a certain loss functional over some data distribution. Nonetheless, constraints on the unknown function, oftentimes *nonlinear*, are necessary to meet the physical system modeling or to provide risk guarantees. This is increasingly the case in problems such as motion planning with obstacle avoidance [2], wireless communications with quality of service (QoS) guarantees [3], and nonlinear filtering with built-in outlier rejection [4].

Function-valued constrained optimization dates back to variational calculus [5] and Hamilton [6]. However, many engineering applications lead to a more generic problem formulation than those which arise from certain physical laws. Meanwhile, variational inference methods have been developed to handle functional stochastic programs that arise from statistical inference, especially in hyper-parameter search [7]. Unless special distributional structure is present, however, these methods typically do not admit efficient iterative solutions, but instead yields an intractable integral equation.

‡ Both authors contribute equally to this manuscript.

A. Koppel is with U.S. Army Research Laboratory (alec.e.koppel.civ@mail.mil). 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 Department of Electrical and Computer Engineering at the University of Texas at Austin (haozhu@utexas.edu). Research was supported in part by the Army Research Laboratory under the Cooperative Agreement W911NF-17-2-0196, National Science Foundation under the Award ECCS-1802319, and ASEE SMART. Part of this work without any proofs has been submitted as [1].

Generally speaking, the functional optimization problem is challenged by the trade-off between its computational tractability and its richness to address realistic scenarios, i.e. the universality of the function approximator. For instance, in learning theory [8] as well as control theory [9], we typically restrict the function to be in a polynomial form [10], or be a Gaussian process [11], a neural network [12], or a nonparametric basis expansion in terms of data [13]. In this work, we adopt the latter nonparametric approach, i.e., the function class is taken to be a reproducing Kernel Hilbert Space (RKHS), motivated by a recently developed memory-efficient parameterization of a function that is infinite dimensional [14]. This so-termed *POLK* method subsumes polynomial interpolation [10], and provides a methodology that circumvents the memory explosion associated with large sample-size Gaussian process regression [15]. It further preserves convexity, thus avoiding convergence to poor stationary points rampant in training neural networks [16].

In this work, we extend the kernelized functional stochastic programming approach of [14] to settings with nonlinear constraints. Constraints have been considered in some recent work on online learning in vector-spaces [17], [18]. In function spaces, preliminary efforts for constrained online learning include [19], [20], through proximal projections and penalty methods. However, their applicability is limited to specialized constraints that exclude obstacle avoidance [21], wireless QoS constraints [3], or risk measures such as conditional value-at-risk (CVaR) [22], [23] that may be used to overcome bias-variance issues in learning. One barrier to handling general nonlinear constraints in RKHS optimization is that the Representer Theorem [24], which is used to transform the functional problem in the unconstrained case to a parametric form, does not apply directly. Thus, we propose to transform the constrained optimization problem in RKHS to a minimax saddle-point problem, via Lagrange duality theory. Then, we extend the Representer Theorem to this saddle-point problem, under certain structural assumptions on the constraints.

With this tool in hand, we develop a stochastic saddle-point algorithm [25], which operates by executing alternating projected primal/dual stochastic gradient descent/ascent on the augmented Lagrangian function. Due to the structure of the RKHS and repeated application of the kernel trick, the complexity of parametrizing the function grows proportionately with the iteration index. To ameliorate this issue, we project the primal function iterates onto low-dimensional subspaces which are subspaces greedily constructed using matching pursuit [26]. By tuning the projection-induced error to the algorithm step-size [14], we establish mean convergence to

a $\mathcal{O}(\sqrt{T})$ neighborhood in terms of objective sub-optimality and $\mathcal{O}(T^{3/4})$ with respect to constraint violation, both of which depend on a chosen constant $1/\sqrt{T}$ step-size and $1/T$ approximation budget. These results are akin to existing results on primal-dual methods for vector-valued stochastic programming under nonlinear constraints [17], [27]. Following this, we use the resulting algorithmic framework to develop for the first time an online optimally compressed kernel support vector machine classifier and nonlinear filter using kernel Ridge regression, both with built-in outlier rejection through the use of CVaR constraints. We illustrate the utility of the designed nonlinear filter on LIDAR data [28].

The rest of the paper is organized as follows. In Section II, we formulate the constrained optimization problem in RKHS and extend the Representer Theorem to a class of saddle-point problems. The projected stochastic primal-dual method is introduced in Section III and analyzed in Section IV. We then evaluate the proposed method numerically and experimentally in Section V. Lastly, we conclude in Section VI.

II. Constrained Learning with Kernels

We consider the problem of constrained stochastic optimization in reproducing kernel Hilbert spaces. Specifically, the objective is to minimize the average of a loss function $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, regularized by a complexity-reducing penalty $(\lambda/2)\|f\|^2_{\mathcal{H}}$ for some $\lambda > 0$. Here $\mathcal{H}$ represents a Hilbert space, and we have $\mathcal{X} \in \mathbb{R}^p, \mathcal{Y} \in \mathbb{R}$ for some $p > 0$. 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 $\mathcal{H}$ is a space of *functions*, $f : \mathcal{X} \to \mathcal{Y}$, which admits representations in terms of elements of $\mathcal{X}$ when $\mathcal{H}$ has a special structure. We consider the RKHS, where $\mathcal{H}$ is equipped with a kernel function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that:

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

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 that 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

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

where f^* denotes its optimum solution. Thanks to the strong convexity guaranteed by the regularization term for a 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) [8]. 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 [8]. Most of the existing work 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)* [29]. 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

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

where CVaR_α denotes the α-CVaR as in [23], 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 [23, 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 an autonomous agent, e.g., a robot, that is 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

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

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 [30]. To convexify the constraint, one approach is to approximate the probabilistic constraint using a more conservative constraint based on expectations [31]. Specifically, the surrogate constraint is given by

$$\inf_{\lambda > 0} \left[\Psi(f, \lambda) - \lambda\gamma\right] \leq 0, \qquad (5)$$

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 [31] 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 [32] and wireless network utility maximization with QoS constraints [3]. To solve (2), the technicalities regarding extending the Representer Theorem [24] to constrained problems must be addressed, which we do in the following subsection.

A. Representer Theorem for Constrained Case

We now 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, define

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

With these definitions, we may formulate the Lagrangian relaxation of (2):

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

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 strongly convex in f. Assuming that Slater's condition [33] holds in this paper, we have strong duality. Thus, f^* [cf. (2)] is equivalent to the primal-dual pair $(f^*,\boldsymbol{\mu}^*)$ that solves the 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 $\mathbb{R}^m_+ = \{\boldsymbol{\mu}\in\mathbb{R}^m |\ \mu_j \geq 0,\ \forall j = 1,\cdots,m\} \subseteq \mathbb{R}^m$.

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., [34], [24]. Specifically, the optimal $f(\boldsymbol{x})$ in $\mathcal{H}$ can be written as a basis 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 the constrained counterpart of stochastic optimization problem in RKHS. To generalize this classical result to the constrained case, we study the problem setting with data-dependent constraints. In particular, we assume that the convex function $G_j(f)$ in the constraints of (2) is also an expectation of some $g_j : \mathcal{H}\times\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ over the joint distribution of the random $\boldsymbol{x}$ and $\boldsymbol{y}$; i.e., $\boldsymbol{G}(f) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[\boldsymbol{g}(f(\boldsymbol{x}),\boldsymbol{y})]$ with $\boldsymbol{g} = (g_1,\cdots,g_m)^\top$. This way, the empirical counterpart of (7) over the training set $\mathcal{S} = \{(\boldsymbol{x}_1,\boldsymbol{y}_1),\cdots,(\boldsymbol{x}_T,\boldsymbol{y}_T)\}$ becomes

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

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

with $\mathcal{L}^o(f,\boldsymbol{\mu};\mathcal{S})$ defined by

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

Next, we establish that the classical Representer Theorem extends to the sample average approximate saddle-point problem stated in (8).

Theorem 1. *Fix the kernel κ, with $\mathcal{H}$ being the corresponding RKHS. Let $\mathcal{S} = \{(\boldsymbol{x}_1,\boldsymbol{y}_1),\cdots,(\boldsymbol{x}_T,\boldsymbol{y}_T)\}$ be the training dataset. Suppose the empirical estimate of each constrained function G_j takes the form $G_j(f;\mathcal{S}) = \frac{1}{T}\sum_{t=1}^{T} g_j(f(\boldsymbol{x}_t),\boldsymbol{y}_t)$. Then, all primal-optimal solutions to* (8) *take the form*

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

where $w_t\in\mathbb{R}$ are some real-valued coefficients.

Proof: The proof follows from that of the classical Representer Theorem. For any given $\boldsymbol{\mu}\in\mathbb{R}^m_+$, the inner minimization in (8) can be viewed as an instance of expected risk minimization problem with the empirical loss objective

$$Q(f;\mathcal{S},\boldsymbol{\mu}) = \frac{1}{T}\sum_{t=1}^{T}\left[\ell(f(\boldsymbol{x}_t),\boldsymbol{y}_t) + \sum_{j=1}^{m}\mu_j g_j(f(\boldsymbol{x}_t),\boldsymbol{y}_t)\right].$$

Note that $Q(f;\mathcal{S},\boldsymbol{\mu}) = Q(f(\boldsymbol{x}_1),\cdots,f(\boldsymbol{x}_T);\boldsymbol{\mu})$, only depending on the function values at elements of the training set. Let $\mathcal{F}_{\kappa,\mathcal{S}}$ be the subspace of functionals spanned by the kernel functions $\kappa(\boldsymbol{x}_t,\cdot),\ \forall t\in[T]$; i.e.,

$$\mathcal{F}_{\kappa,\mathcal{S}} = \text{span}\{\kappa(\boldsymbol{x}_t,\cdot) : \forall t\in[T]\}.$$

Also, denote the projection of f on $\mathcal{F}_{\kappa,\mathcal{S}}$ as $f_{\mathcal{S}}$, and the corresponding perpendicular component as $f_\perp = f - f_{\mathcal{S}}$. This way, we can show

$$\begin{aligned} f(\boldsymbol{x}_t) &= \langle f,\kappa(\boldsymbol{x}_t,\cdot)\rangle = \langle f_{\mathcal{S}},\kappa(\boldsymbol{x}_t,\cdot)\rangle + \langle f_\perp,\kappa(\boldsymbol{x}_t,\cdot)\rangle \\ &= \langle f_{\mathcal{S}},\kappa(\boldsymbol{x}_t,\cdot)\rangle = f_{\mathcal{S}}(\boldsymbol{x}_t). \end{aligned}$$

Accordingly, the empirical loss becomes

$$Q(f(\boldsymbol{x}_1),\cdots,f(\boldsymbol{x}_T);\boldsymbol{\mu}) = Q(f_{\mathcal{S}}(\boldsymbol{x}_1),\cdots,f_{\mathcal{S}}(\boldsymbol{x}_T);\boldsymbol{\mu}).$$

As projection is non-expansive, we have $\|f\|^2_{\mathcal{H}} \geq \|f_{\mathcal{S}}\|^2_{\mathcal{H}}$. Hence, given any $\lambda > 0$ and $\boldsymbol{\mu}\in\mathbb{R}^m_+$, $Q(f;\mathcal{S},\boldsymbol{\mu}) + (\lambda/2)\cdot\|f\|^2_{\mathcal{H}}$ is minimized at some $\check{f}^*(\boldsymbol{\mu})$ that lies in $\mathcal{F}_{\kappa,\mathcal{S}}$. In particular, this holds as well for $\check{\boldsymbol{\mu}}^*$ where $\check{f}^* = \check{f}^*(\check{\boldsymbol{\mu}}^*)$, which completes the proof. ∎

Theorem 1 shows that the solution of the empirical saddle-point problem (8) admits a basis expansion in terms of kernel evaluations over the training set. Now, [35] establishes that this result generalizes to expected-value problems, i.e., Theorem 1 holds for $T\to\infty$. Thus, f^* [cf. (2)] admits a basis representation of kernel evaluations at realizations of $(\mathbf{x},\mathbf{y})$, and hence generalizes the Representer Theorem for unconstrained settings in RKHS [24]. Upon this foundation, we now shift to developing an algorithmic solution to address constrained stochastic optimization in RKHS for the first time.

III. Stochastic Primal-dual Method in RKHS

Next, we present an iterative method for solving (2). To this end, we define the approximate Lagrangian relaxation as

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

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 in order to control the violation of non-negative constraints on the dual variable over time t, and incidentally further guarantees that it is strongly concave in the dual. Thus, the saddle point $(f^s, \boldsymbol{\mu}^s)$ of $\mathcal{L}(f,\boldsymbol{\mu})$ is such that f^s is an approximation of f^* [cf. (7)]. Therefore, the true saddle-point problem in (7) can be approximately solved by the following one

$$(f^s, \boldsymbol{\mu}^s) = \arg \max_{\boldsymbol{\mu}\in\mathbb{R}_+^m} \min_{f\in\mathcal{H}} \quad \mathcal{L}(f,\boldsymbol{\mu}). \quad (12)$$

Further define the instantaneous augmented Lagrangian based on one realization of $\mathcal{L}(f,\boldsymbol{\mu})$ using sample $(\boldsymbol{x}_t, \boldsymbol{y}_t)$, as

$$\widehat{\mathcal{L}}_t(f,\boldsymbol{\mu}) = \ell(f(\boldsymbol{x}_t),\boldsymbol{y}_t) + \sum_{j=1}^m \mu_j g_j(f(\boldsymbol{x}_t),\boldsymbol{y}_t) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 - \frac{\delta\eta}{2}\|\boldsymbol{\mu}\|^2. \quad (13)$$

Note that the expectation of (13) over data $(\mathbf{x},\mathbf{y})$ yields (11). Our algorithm, detailed soon, is developed on basis of the stochastic gradient updates using (13) for solving the augmented saddle-point problem (12) – see Sec. IV

A. Functional Primal-dual Method

We focus here on the online setting, i.e. the sample size T may grow unbounded or samples $(\boldsymbol{x}_t, \boldsymbol{y}_t)$ are sequentially observed. In particular, we consider the case where $(\boldsymbol{x}_t, \boldsymbol{y}_t)$ are independent realizations from a stationary joint distribution of the random pair $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{X}\times\mathcal{Y}$ [36]. For notational convenience, we define $\widetilde{\ell}(f(\boldsymbol{x}),\boldsymbol{y},\boldsymbol{\mu})$ as:

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

The reproducing property (1) (i) implies that for any $f\in\mathcal{H}$,

$$\frac{\partial f(\boldsymbol{x}_t)}{\partial f} = \frac{\partial\langle f, \kappa(\boldsymbol{x}_t,\cdot)\rangle_{\mathcal{H}}}{\partial f} = \kappa(\boldsymbol{x}_t,\cdot). \quad (14)$$

Thus, following the derivation in [37], 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

$$\nabla_f \widetilde{\ell}(f(\boldsymbol{x}_t),\boldsymbol{y}_t,\boldsymbol{\mu})(\cdot) = \frac{\partial\widetilde{\ell}(f(\boldsymbol{x}_t),\boldsymbol{y}_t,\boldsymbol{\mu})}{\partial f(\boldsymbol{x}_t)}\frac{\partial f(\boldsymbol{x}_t)}{\partial f}(\cdot) = \widetilde{\ell}'(f(\boldsymbol{x}_t),\boldsymbol{y}_t,\boldsymbol{\mu})\kappa(\boldsymbol{x}_t,\cdot) \quad (15)$$

where we define

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

as the derivative of $\widetilde{\ell}(f(\boldsymbol{x}_t),\boldsymbol{y}_t,\boldsymbol{\mu})$ w.r.t. its scalar 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

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

where ℓ' and g_j' denote the derivative w.r.t. the scalar $f(\boldsymbol{x}_t)$ evaluated at $\boldsymbol{x}_t$. With these definitions, we propose a stochastic variant of primal-dual method [25], [38] to address (12):

$$\begin{cases} f_{t+1} = (1-\eta\lambda)f_t - \eta\Big[\ell'(f_t(\boldsymbol{x}_t),\boldsymbol{y}_t) + \sum_{j=1}^m \mu_j g_j'(f_t(\boldsymbol{x}_t),\boldsymbol{y}_t)\Big]\kappa(\boldsymbol{x}_t,\cdot), & (16a) \\ \boldsymbol{\mu}_{t+1} = \big[(1-\eta^2\delta)\boldsymbol{\mu}_t + \eta\boldsymbol{g}(f_t(\boldsymbol{x}_t),\boldsymbol{y}_t)\big]_+, & (16b) \end{cases}$$

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 control the constraint violation of the dual variable using the learning rate, as we will show in Sec. IV.

For a given regularizer $\lambda > 0$ in (2), we require the step-size to satisfy $\eta < 1/\lambda$. The sequence of $(f_t, \boldsymbol{\mu}_t)$ is initialized by $f_1 = 0 \in \mathcal{H}$ and $\boldsymbol{\mu}_1 = \mathbf{0} \in \mathbb{R}_+^m$. Therefore, following the updates (16), the iterate f_t can be expressed as an expansion in terms of feature vectors $\boldsymbol{x}_t$ observed thus far; i.e.,

$$f_t(\boldsymbol{x}) = \sum_{t=1}^{t-1} w_t\kappa(\boldsymbol{x}_t,\boldsymbol{x}) = \boldsymbol{w}_t^\top \boldsymbol{\kappa}_{\boldsymbol{X}_t}(\boldsymbol{x}), \quad (17)$$

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

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

This way, f_t belongs to the functional subspace spanned by $\{\kappa(\boldsymbol{x}_1,\cdot),\cdots,\kappa(\boldsymbol{x}_{t-1},\cdot)\}$. Notice that performing the primal dual update of f_t as (16a) amounts to the following parametric updates on the kernel dictionary $\boldsymbol{X}$ and coefficient vector $\boldsymbol{w}$:

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

This update causes $\boldsymbol{X}_{t+1}$ to have one more column than $\boldsymbol{X}_t$. We define the *model order* as the number of data points M_t in the dictionary at time t. Hence, in the update (16a), the model order $M_t = t-1$ grows unbounded with iteration index t.

Proximal Projection: Motivated by the dimensionality reduction approach in [14], we propose to project the functional stochastic gradient update of f_t onto some subspace $\mathcal{H}_{\boldsymbol{D}} \subseteq \mathcal{H}$, which consists only of functions that can be represented using some dictionary $\boldsymbol{D} = [\boldsymbol{d}_1,\cdots,\boldsymbol{d}_M] \in \mathbb{R}^{p\times M}$ of fixed size M. In particular, $\mathcal{H}_{\boldsymbol{D}}$ has the form $\mathcal{H}_{\boldsymbol{D}} = \{f : f(\cdot) = \sum_{t=1}^M w_t\kappa(\boldsymbol{d}_t,\cdot) = \boldsymbol{w}^\top\boldsymbol{\kappa}_{\boldsymbol{D}}(\cdot)\}$, where we define $\boldsymbol{\kappa}_{\boldsymbol{D}}(\cdot) = [\kappa(\boldsymbol{d}_1,\cdot)\ldots\kappa(\boldsymbol{d}_M,\cdot)]$. The dictionary $\boldsymbol{D}$ is updated as $\boldsymbol{D}_{t+1}$ along iterations when a new sample $(\boldsymbol{x}_t,\boldsymbol{y}_t)$ becomes

available. Therefore, we replace the update (16a) with the following one that has a projection onto subspace $\mathcal{H}_{\boldsymbol{D}_{t+1}}$:

$$f_{t+1} = \operatorname*{argmin}_{f\in\mathcal{H}_{\boldsymbol{D}_{t+1}}} \Big\| f - \Big((1-\eta\lambda)f_t - \eta\nabla_f \widetilde{\ell}(f_t(\boldsymbol{x}_t), \boldsymbol{y}_t, \boldsymbol{\mu}_t)\Big)\Big\|_{\mathcal{H}}^2$$
$$:= \mathcal{P}_{\mathcal{H}_{\boldsymbol{D}_{t+1}}}\Big[(1-\eta\lambda)f_t - \eta\nabla_f \widetilde{\ell}(f_t(\boldsymbol{x}_t), \boldsymbol{y}_t, \boldsymbol{\mu}_t)\Big], \quad (18)$$

where we define the operator $\mathcal{P}_{\mathcal{H}_{\boldsymbol{D}}}$ as one that projects the input onto subspace $\mathcal{H}_{\boldsymbol{D}} \subseteq \mathcal{H}$.

To project the function onto $\mathcal{H}_{\boldsymbol{D}_{t+1}}$, we first define the dictionary $\widetilde{\boldsymbol{D}}_{t+1}$ and weight $\widetilde{\boldsymbol{w}}_{t+1}$ defined by the updates (16) before projection as

$$\widetilde{\boldsymbol{D}}_{t+1} = [\boldsymbol{D}_t,\ \boldsymbol{x}_t], \quad (19)$$
$$\widetilde{\boldsymbol{w}}_{t+1} = [(1-\eta\lambda)\boldsymbol{w}_t,\ -\eta\widetilde{\ell}'(f_t(\boldsymbol{x}_t), \boldsymbol{y}_t, \boldsymbol{\mu}_t)]\ .$$

and denote the un-projected function sequence as $\widetilde{f}_{t+1} = (1-\eta\lambda)f_t - \eta\nabla_f \widetilde{\ell}(f_t(\boldsymbol{x}_t), \boldsymbol{y}_t, \boldsymbol{\mu}_t)$. Then, given any 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

$$\boldsymbol{w}_{t+1} = \boldsymbol{K}^{-1}_{\boldsymbol{D}_{t+1}\boldsymbol{D}_{t+1}} \boldsymbol{K}_{\boldsymbol{D}_{t+1}\widetilde{\boldsymbol{D}}_{t+1}} \widetilde{\boldsymbol{w}}_{t+1}\ , \quad (20)$$

where $\boldsymbol{K}^{-1}_{\boldsymbol{D}_{t+1}\boldsymbol{D}_{t+1}}$ and $\boldsymbol{K}_{\boldsymbol{D}_{t+1}\widetilde{\boldsymbol{D}}_{t+1}}$ are both kernel matrices between the dictionaries $\{\boldsymbol{D}_{t+1}, \boldsymbol{D}_{t+1}\}$ and $\{\boldsymbol{D}_{t+1}, \widetilde{\boldsymbol{D}}_{t+1}\}$, respectively. 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 a destructive variant of *kernel orthogonal matching pursuit* (KOMP) with pre-fitting [39][Sec. 2.3] as in [14]. KOMP operates by beginning with the full dictionary $\widetilde{\boldsymbol{D}}_{t+1}$ and sequentially removing its columns while the condition $\|\widetilde{f}_{t+1} - f_{t+1}\|_{\mathcal{H}} \leq \epsilon_t$ is true. This allows us to only keep kernel dictionary elements which preserve Lyapunov stability of the optimization sequence. Moreover, we also assume that the f_{t+1} output from KOMP has bounded Hilbert norm,[2] which is typical in analyses of primal-dual methods [17], [38], [27]. Hence, the following projection onto $\mathcal{H}_{\boldsymbol{D}_{t+1}}$ controls not only the model order but also the Hilbert norm of $\{f_t\}$,

$$(f_{t+1}, \boldsymbol{D}_{t+1}, \boldsymbol{w}_{t+1}) = \textbf{KOMP}(\widetilde{f}_{t+1}, \widetilde{\boldsymbol{D}}_{t+1}, \widetilde{\boldsymbol{w}}_{t+1}, \epsilon_t). \quad (21)$$

Here ϵ_t is the approximation budget which dictates how many model points are thrown away during compression. By design, we have $\|f_{t+1} - \widetilde{f}_{t+1}\|_{\mathcal{H}} \leq \epsilon_t$. Note that the dual variable $\boldsymbol{\mu}_t$ shows up in the weight vector $\widetilde{\boldsymbol{w}}_{t+1}$. To recap, the online primal-dual algorithm is updated as follows:

$$\begin{cases} f_{t+1} = \mathcal{P}_{\mathcal{H}_{\boldsymbol{D}_{t+1}}}\Big[(1-\eta\lambda)f_t \\ \qquad\qquad - \eta\nabla_f \widetilde{\ell}(f_t(\boldsymbol{x}_t), \boldsymbol{y}_t, \boldsymbol{\mu}_t)\Big] & \text{(22a)} \\ \boldsymbol{\mu}_{t+1} = \big[(1-\eta^2\delta)\boldsymbol{\mu}_t + \eta\boldsymbol{g}(f_t(\boldsymbol{x}_t), \boldsymbol{y}_t)\big]_+, & \text{(22b)} \end{cases}$$

Given sequentially observed data $(\mathbf{x}_t, \boldsymbol{y}_t)$, the algorithm alternates between primal stochastic descent steps (19) and dual stochastic ascent steps (22b). The primal iterates are projected onto sparse subspaces defined by the output of matching pursuit (21). The update rule of the projected primal-dual method is summarized as Algorithm 1.

[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 [39], which can be achieved by thresholding the coefficient sequence during compression.

Algorithm 1 Projected Primal-Dual Method in Kernel Space

Require: $\{\mathbf{x}_t, \boldsymbol{y}_t, \epsilon_t, \eta, \delta\}_{t=0,1,2,\ldots}$
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 $(\mathbf{x}_t, \boldsymbol{y}_t)$
Take stochastic descent step on Lagrangian [cf. (16a)]

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

Take stochastic ascent step on Lagrangian [cf. (16b)]

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

Update $\widetilde{\boldsymbol{D}}_{t+1} = [\boldsymbol{D}_t, \mathbf{x}_t]$ and $\widetilde{\boldsymbol{w}}_{t+1}$ [cf. (19)]
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

Before shifting to establishing that (2) may be successfully addressed by Algorithm 1, we present the specific update rules for the two motivating examples in Section II.

Example 1. *Risk-aware supervised learning using CVaR:* In this example, the objective is the regularized ERM as in (2). Moreover, the CVaR constraint in (3) is not exactly in the form of expectation over $(\boldsymbol{x}, \boldsymbol{y})$ as required in Theorem 1. Thus, we approximate the CVaR constraint $G(f) \leq 0$ with $\widetilde{G}(f) \leq 0$, by exchanging the minimization and expectation in $G(f)$, i.e.,

$$\widetilde{G}(f) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left\{\min_{z\in\mathbb{R}} z + \frac{1}{1-\alpha}[\ell(f(\boldsymbol{x}), \boldsymbol{y}) - z]_+\right\} - \gamma \quad (23)$$

Furthermore, due to the operators $\min$ and $[\cdot]_+$, the function $g(f(\boldsymbol{x}), \boldsymbol{y})$ defined in (23) is non-differentiable. Thus, we numerically approximate the positive projection by a softmax: $\max(a, 0) \approx \text{softmax}(a, 0) = \log(1 + e^a)$, whose gradient is $\nabla_a \text{softmax}(a, 0) = e^a/(1 + e^a)$. The minimization over z can be solved numerically since it is a scalar, which then yields a subgradient direction. In Section V, we evaluate the quality of this approximation in terms of the true CVaR (3), and show that the approximation preserves the feasibility.

Example 2. *Chance-constrained motion planning:* With only dependence on $\boldsymbol{x}$, the cost functional ℓ in (2) has the form of $\ell(f(\boldsymbol{x}), \boldsymbol{y}) = \ell(f(\boldsymbol{x})) = \mathcal{U}(f(\boldsymbol{x}))$. The gradient ℓ' can thus be defined accordingly. As in **Example 1**, in order to satisfy the assumptions in Theorem 1, we approximate the surrogate constraint (5) by exchanging the operations of taking infimum and expectation. Thus, the constraint $G(f)$ is approximated as

$$\widetilde{G}(f) = \mathbb{E}_{\boldsymbol{x}}\Big[\inf_{\lambda>0} \lambda\phi(\lambda^{-1}g(f(\boldsymbol{x}))) - \lambda\gamma\Big], \quad (24)$$

recalling that $\phi(\cdot)$ is the generating function. Moreover, due to the $\inf$ operator, the term inside the expectation $\widetilde{G}(f)$ in (24) may not be differentiable. However, since λ is a scalar, it can be numerically evaluated, as in the case of minimization over positive Lagrange multipliers in ADMM [40]. The gradient $\widetilde{\ell}'(f_t(\boldsymbol{x}_t), \mu_t)$ thus has the form

$$\widetilde{\ell}'(f_t(\boldsymbol{x}_t), \mu_t) = \ell'(f_t(\boldsymbol{x}_t)) + \mu_t \phi'(\lambda^{-1} g(f_t(\boldsymbol{x}_t))) g'(f_t(\boldsymbol{x}_t)), \tag{25}$$

where ϕ' and g' represent the gradients with respect to their corresponding arguments. Thus, the update rule (22) can be obtained by substituting (24) and (25).

Next, we validate (22)] theoretically and experimentally.

IV. Convergence Analysis

We 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. Before proceeding, we introduce some quantities for notational convenience, which simplify the proofs. First, define the stochastic functional gradient for the augmented Lagrangian $\mathcal{L}(f_t, \boldsymbol{\mu}_t)$ at f_t as

$$\nabla_f \widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t) = \nabla_f \widetilde{\ell}(f_t(\boldsymbol{x}_t), \boldsymbol{y}_t, \boldsymbol{\mu}_t) + \lambda f_t \tag{26}$$

Then, we define the projected stochastic functional gradient associated with the update in (18) as

$$\widetilde{\nabla}_f \widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t) = \frac{f_t - \mathcal{P}_{\mathcal{H}_{D_{t+1}}}\left[f_t - \eta_t \nabla_f \widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)\right]}{\eta}. \tag{27}$$

Thus, the update (18) can be expressed as

$$f_{t+1} = f_t - \eta \widetilde{\nabla}_f \widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t). \tag{28}$$

Let $\mathcal{F}_t$ denote the σ-algebra which measures the algorithm trajectory for times $u < t$; i.e., $\mathcal{F}_t = \sigma(\{\boldsymbol{x}_u, \boldsymbol{y}_u, f_u, \mu_u\}_{u=0}^{t-1})$. Note that $(\boldsymbol{x}_t, \boldsymbol{y}_t)$ are independent and identically distributed realizations of the random pair $(\boldsymbol{x}, \boldsymbol{y})$. Hence, $\nabla_f \widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)$ is an unbiased estimate of the gradient of the Lagrangian $\mathcal{L}(f_t, \boldsymbol{\mu}_t)$ w.r.t. f_t; i.e., for all $t \geq 0$,

$$\mathbb{E}\big[\nabla_f \widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t) \,|\, \mathcal{F}_t\big] = \nabla_f \mathcal{L}(f_t, \boldsymbol{\mu}_t). \tag{29}$$

Likewise, we define $\nabla_{\boldsymbol{\mu}} \widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)$ as

$$\nabla_{\boldsymbol{\mu}} \widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t) = \boldsymbol{g}(f_t(\boldsymbol{x}_t), \boldsymbol{y}_t) - \delta\eta \cdot \boldsymbol{\mu}_t, \tag{30}$$

and thus $\nabla_{\boldsymbol{\mu}} \widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)$ is an unbiased estimate of the gradient $\nabla_{\boldsymbol{\mu}} \mathcal{L}(f_t, \boldsymbol{\mu}_t)$ w.r.t. $\boldsymbol{\mu}$; i.e., $\mathbb{E}[\nabla_{\boldsymbol{\mu}} \widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t) \,|\, \mathcal{F}_t] = \nabla_{\boldsymbol{\mu}} \mathcal{L}(f_t, \boldsymbol{\mu}_t)$. Moreover, the dual update (22b) takes the form

$$\boldsymbol{\mu}_{t+1} = \big[\boldsymbol{\mu}_t + \eta \nabla_{\boldsymbol{\mu}} \widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)\big]_+. \tag{31}$$

We continue 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 $\mathcal{Y} \subset \mathbb{R}$ are compact, and the reproducing kernel map can be bounded by some constant $X > 0$ as*

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \sqrt{\kappa(\boldsymbol{x}, \boldsymbol{x})} = X < \infty \tag{32}$$

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*

$$|\ell(z, \boldsymbol{y}) - \ell(z', \boldsymbol{y})| \leq C_1 |z - z'|, \forall \boldsymbol{y} \in \mathcal{Y}, \tag{33}$$

$$|g_i(z) - g_i(z')| \leq C_2 |z - z'|, \forall i = 1, \cdots, m. \tag{34}$$

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 w.r.t. 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 $f \in \mathcal{H}$ that satisfies $\boldsymbol{G}(f) < \boldsymbol{0}$.*

Assumption 5. *The output f_{t+1} of the KOMP update (21) 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 [33], 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.

Under these assumptions, we are able to bound the primal and dual gradients of the stochastic augmented Lagrangian $\widehat{\mathcal{L}}_t(f, \boldsymbol{\mu})$. Different from the deterministic primal-dual (sub)gradient methods, the upper bounds for our stochastic framework depend on the norm of the dual variable, $\|\boldsymbol{\mu}\|^2$, and are not constant terms as in [38].

Lemma 1. *Under Assumptions 1-3, for any $(f, \boldsymbol{\mu}) \in \mathcal{B} \times \mathbb{R}_+^m$, the mean-squared-magnitude of the primal and dual gradients of the stochastic augmented Lagrangian $\widehat{\mathcal{L}}_t(f, \boldsymbol{\mu})$ as defined in* (13)*, can be bounded as follows*

$$\mathbb{E}[\|\nabla_f \widehat{\mathcal{L}}_t(f, \boldsymbol{\mu})\|_{\mathcal{H}}^2] \leq 4X^2 \cdot \big(C_1^2 + mC_2^2 \cdot \|\boldsymbol{\mu}\|^2\big) + 2\lambda^2 \cdot R_{\mathcal{B}}^2. \tag{35}$$

$$\mathbb{E}[\|\nabla_{\boldsymbol{\mu}} \widehat{\mathcal{L}}_t(f, \boldsymbol{\mu})\|^2] \leq 2\big(K_1 + mC_2^2 X^2 \cdot R_{\mathcal{B}}^2\big) + 2\delta^2 \eta^2 \cdot \|\boldsymbol{\mu}\|^2, \tag{36}$$

for some $0 < K_1 < \infty$.

Proof: This proof generalizes the analysis of the gradient direction's dependency on the dual variables in [27] to the functional setting. For any $(f, \boldsymbol{\mu}) \in \mathcal{B} \times \mathbb{R}_+^m$, $\mathbb{E}[\|\nabla_f \widehat{\mathcal{L}}_t(f, \boldsymbol{\mu})\|_{\mathcal{H}}^2]$ may be upper bounded as

$$\begin{aligned}
&\mathbb{E}[\|\nabla_f \widehat{\mathcal{L}}_t(f, \boldsymbol{\mu})\|_{\mathcal{H}}^2] \\
&\quad \leq 2\mathbb{E}[\|\nabla_f \widetilde{\ell}(f(\boldsymbol{x}_t), \boldsymbol{y}_t, \boldsymbol{\mu})\|_{\mathcal{H}}^2] + 2\lambda^2 \cdot \|f\|_{\mathcal{H}}^2 \\
&\quad \leq 4X^2 \cdot \big(C_1^2 + mC_2^2 \cdot \|\boldsymbol{\mu}\|^2\big) + 2\lambda^2 \cdot \|f\|_{\mathcal{H}}^2 \\
&\quad \leq 4X^2 \cdot \big(C_1^2 + mC_2^2 \cdot \|\boldsymbol{\mu}\|^2\big) + 2\lambda^2 \cdot R_{\mathcal{B}}^2.
\end{aligned} \tag{37}$$

To obtain (37), we apply twice the inequalities that $\|a+b\|_{\mathcal{H}}^2 \leq 2\cdot\big(\|a\|_{\mathcal{H}}^2+\|b\|_{\mathcal{H}}^2\big)$ for any $a,b\in\mathcal{H}$ and $\big(\sum_{j=1}^m|\mu_j|\big)^2 \leq m\cdot\big(\sum_{j=1}^m|\mu_j|^2\big)=m\cdot\|\boldsymbol{\mu}\|^2$, together with the fact that $\mathcal{B}$ has radius $R_{\mathcal{B}}$. Thus, the claim in (35) is valid.

Now, we shift focus to the dual. Its gradient's magnitude can be upper estimated as

$$\begin{aligned}\mathbb{E}[\|\nabla_{\boldsymbol{\mu}}\widehat{\mathcal{L}}_t(f,\boldsymbol{\mu})\|^2] &\leq 2\mathbb{E}[\|\boldsymbol{g}(f(\boldsymbol{x}_t),\boldsymbol{y}_t)\|^2]+2\delta^2\eta^2\|\boldsymbol{\mu}\|^2\\ &\leq 2\big[K_1+mC_2^2\mathbb{E}(|f(\boldsymbol{x}_t)|^2)\big]+2\delta^2\eta^2\|\boldsymbol{\mu}\|^2, \quad (38)\end{aligned}$$

for some absolute constant $K_1>0$. The first inequality is due to $\|a+b\|^2\leq 2\cdot\big(\|a\|^2+\|b\|^2\big)$ for any $a,b\in\mathbb{R}$, and the second one is due to boundedness of the norm of the constraint function $\|\boldsymbol{g}(f(\boldsymbol{x}_t),\boldsymbol{y}_t)\|$, which follows from the Lipschitz continuity of the function $\boldsymbol{g}(\cdot,\cdot)$. Now, we focus on the second term inside the first brackets on the right-hand side of (38), which depends on $f(\boldsymbol{x}_t)$. Apply Cauchy-Shwartz and Assumptions 1 and 5 regarding the compactness of the feature space and the fact that $\mathcal{B}$ has finite radius, to obtain

$$\begin{aligned}|f(\boldsymbol{x}_t)|^2=|\langle f,\kappa(\boldsymbol{x}_t,\cdot)\rangle|^2 &\leq \|f\|_{\mathcal{H}}^2\|\kappa(\boldsymbol{x}_t,\cdot)\|_{\mathcal{H}}^2\\ &\leq (X^2\cdot R_{\mathcal{B}}^2) \quad (39)\end{aligned}$$

Conclude (36) by combining (39) and (38). ∎

The two inequalities (35) and (36) play essential roles in the following analysis. By bounding the primal and dual gradients of the augmented Lagrangian in terms of the squared-norm of $\boldsymbol{\mu}$, we obviate the need of projecting $\boldsymbol{\mu}$ onto the compact subset of $\mathbb{R}_+^m$ as in the standard approach to analyze primal-dual methods; see e.g., [38]. In fact, the unbounded Lagrange multipliers here allow us to control the growth of constraint violation over time using regularization [cf. (11)].

A. Convergence Results

Now we turn to analyze the convergence of the proposed algorithm by establishing bounds on the two sequences, namely the objective function error $\{R(f_t)-R(f^*)\}$ and the accumulated constraint violation $\mathbf{G}(f_t)$, both in expectation. Before we present the main convergence results, a set of lemmas are stated for subsequent use. First, Lemma 2 asserts a bounded difference between the stochastic functional gradient $\nabla_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)$ and its projected counterpart $\widetilde{\nabla}_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)$ as defined in (26) and (27), respectively. Its proof follows from that of Proposition 7 in [14] and is omitted here for brevity. The key is that using Assumption 5, we can still relate the error caused by sparsification $\|f_{t+1}-\widetilde{f}_{t+1}\|_{\mathcal{H}}\leq\epsilon_t$ in (21) to the directional error in the stochastic gradient itself,

Lemma 2. [Proposition 7 in [14]] *Under Assumptions 1-5, given independent identical realizations $(\boldsymbol{x}_t,\boldsymbol{y}_t)$ of $(\boldsymbol{x},\boldsymbol{y})$, the difference between the stochastic primal functional gradient of the augmented Lagrangian* (26) *and its proximal projection* (27)*, is bounded for all $t>0$ as*

$$\big\|\nabla_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)-\widetilde{\nabla}_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)\big\|_{\mathcal{H}}\leq\frac{\epsilon_t}{\eta}, \quad (40)$$

*recalling that $\eta>0$ is the algorithm step-size, while $\epsilon_t>0$ is the approximation budget of **KOMP** update* (21)*.*

With the error associated with parsimonious projections [14], we shift to establishing that a decrement-like property holds for the instantaneous Lagrangian difference, namely $\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu})-\widehat{\mathcal{L}}_t(f,\boldsymbol{\mu}_t)$.

Lemma 3. *Under Assumptions 1-5, the instantaneous Lagrangian difference for the sequence $(f_t,\boldsymbol{\mu}_t)$ from the update* (22) *satisfies the following decrement property for any $f\in\mathcal{B}$ and $\boldsymbol{\mu}\in\mathbb{R}_+^m$:*

$$\begin{aligned}&\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu})-\widehat{\mathcal{L}}_t(f,\boldsymbol{\mu}_t) \quad (41)\\ \leq&\frac{1}{2\eta}\big(\|f_t-f\|_{\mathcal{H}}^2-\|f_{t+1}-f\|_{\mathcal{H}}^2+\|\boldsymbol{\mu}_t-\boldsymbol{\mu}\|^2\\ &-\|\boldsymbol{\mu}_{t+1}-\boldsymbol{\mu}\|^2\big)+\frac{\eta}{2}\Big(2\cdot\big\|\nabla_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)\big\|_{\mathcal{H}}^2\\ &+\big\|\nabla_{\boldsymbol{\mu}}\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)\big\|^2\Big)+\frac{\epsilon_t}{\eta}\|f_t-f\|_{\mathcal{H}}+\frac{\epsilon_t^2}{\eta}.\end{aligned}$$

Proof: This lemma is the proximal RKHS generalization of [17, Lemma 2]. Consider the squared Hilbert norm of the difference between the iterate f_{t+1} and any feasible point f in the ball $\mathcal{B}$, and expand it using the update (28), to obtain

$$\begin{aligned}&\|f_{t+1}-f\|_{\mathcal{H}}^2=\|f_t-\eta\widetilde{\nabla}_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)-f\|_{\mathcal{H}}^2 \quad (42)\\ =&\|f_t-f\|_{\mathcal{H}}^2-2\eta\langle f_t-f,\nabla_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)\rangle\\ &-2\eta\langle f_t-f,\widetilde{\nabla}_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)-\nabla_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)\rangle\\ &+\eta^2\big\|\widetilde{\nabla}_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)\big\|_{\mathcal{H}}^2\end{aligned}$$

where the inner product with $\widetilde{\nabla}_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)$ has been separated into two terms on the right-hand side. Let's focus on the third term on the right-hand side of (42). Continue by using Cauchy-Schwartz inequality, together with Lemma 2, to bound the directional error associated with proximal stochastic gradients rather than the true one:

$$\begin{aligned}&\langle f_t-f,\widetilde{\nabla}_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)-\nabla_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)\rangle \quad (43)\\ &\leq\|f_t-f\|_{\mathcal{H}}\|\widetilde{\nabla}_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)-\nabla_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)\|_{\mathcal{H}}\\ &\leq\frac{\epsilon_t}{\eta}\|f_t-f\|_{\mathcal{H}}.\end{aligned}$$

As for the norm of $\widetilde{\nabla}_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)$, the last term on the right-hand side of (42), since f does not necessarily belong to the subspace $\widetilde{\mathcal{H}}_{\boldsymbol{D}_t}$, we cannot apply the non-expansiveness of the projection operator $\mathcal{P}_{\mathcal{H}_{\boldsymbol{D}_{t+1}}}$ to bound it in terms of $\|f_t-f\|_{\mathcal{H}}$. Instead, we add and subtract the primal stochastic gradient $\nabla_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)$ to bound $\big\|\widetilde{\nabla}_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)\big\|_{\mathcal{H}}^2$, i.e.,

$$\begin{aligned}&\big\|\widetilde{\nabla}_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)\big\|_{\mathcal{H}}^2\\ =&\big\|\widetilde{\nabla}_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)-\nabla_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)+\nabla_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)\big\|_{\mathcal{H}}^2\\ \leq&\,2\big\|\nabla_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)\big\|_{\mathcal{H}}^2+2\frac{\epsilon_t^2}{\eta^2}, \quad (44)\end{aligned}$$

where we have also applied $\|a+b\|_{\mathcal{H}}^2\leq 2\cdot\big(\|a\|_{\mathcal{H}}^2+\|b\|_{\mathcal{H}}^2\big)$. Now, substitute (43) and (44) into (42), to obtain

$$\begin{aligned}\|f_{t+1}-f\|_{\mathcal{H}}^2\leq\|f_t-f\|_{\mathcal{H}}^2-2\eta\langle f_t-f,\nabla_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)\rangle \quad (45)\\ +2\epsilon_t\|f_t-f\|_{\mathcal{H}}+2\eta^2\big\|\nabla_f\widehat{\mathcal{L}}_t(f_t,\boldsymbol{\mu}_t)\big\|_{\mathcal{H}}^2+2\epsilon_t^2.\end{aligned}$$

Reordering the terms in (45) leads to

$$\begin{aligned}\langle f_t - f, \nabla_f \widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)\rangle &\leq \frac{1}{2\eta}\big(\|f_t - f\|_{\mathcal{H}}^2 - \|f_{t+1} - f\|_{\mathcal{H}}^2\big) \\ &\quad + \eta\big\|\nabla_f \widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)\big\|_{\mathcal{H}}^2 + \frac{\epsilon_t}{\eta}\|f_t - f\|_{\mathcal{H}} + \frac{\epsilon_t^2}{\eta}.\end{aligned} \tag{46}$$

Since the instantaneous Lagrangian $\widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)$ is convex w.r.t. f_t, we may write

$$\widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t) - \widehat{\mathcal{L}}_t(f, \boldsymbol{\mu}_t) \leq \langle f_t - f, \nabla_f \widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)\rangle. \tag{47}$$

Substitute the left-hand side of (47) into (46), to obtain

$$\begin{aligned}&\widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t) - \widehat{\mathcal{L}}_t(f, \boldsymbol{\mu}_t) \\ &\leq \frac{1}{2\eta}\big(\|f_t - f\|_{\mathcal{H}}^2 - \|f_{t+1} - f\|_{\mathcal{H}}^2\big) + \eta\big\|\nabla_f \widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)\big\|_{\mathcal{H}}^2 \\ &\quad + \frac{\epsilon_t}{\eta}\|f_t - f\|_{\mathcal{H}} + \frac{\epsilon_t^2}{\eta}.\end{aligned} \tag{48}$$

We now mirror these analytical steps in the dual domain. Consider the squared difference between the Lagrange multiplier $\boldsymbol{\mu}_{t+1}$ and any $\boldsymbol{\mu}$, and bound it by using the update (31) as

$$\begin{aligned}\|\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}\|^2 &= \Big\|\big[\boldsymbol{\mu}_t + \eta\nabla_{\boldsymbol{\mu}}\widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)\big]_+ - \boldsymbol{\mu}\Big\|^2 \\ &\leq \big\|\boldsymbol{\mu}_t + \eta\nabla_{\boldsymbol{\mu}}\widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t) - \boldsymbol{\mu}\big\|^2,\end{aligned} \tag{49}$$

where the inequality follows from the non-expansiveness of projection. This bound can be further expanded as

$$\begin{aligned}\|\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}\|^2 \leq \|\boldsymbol{\mu}_t - \boldsymbol{\mu}\|^2 &+ 2\eta\nabla_{\boldsymbol{\mu}}\widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)^\top(\boldsymbol{\mu}_t - \boldsymbol{\mu}) \\ &+ \eta^2\big\|\nabla_{\boldsymbol{\mu}}\widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)\big\|^2.\end{aligned} \tag{50}$$

Re-ordering the terms in (50), we obtain

$$\begin{aligned}\nabla_{\boldsymbol{\mu}}\widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)^\top(\boldsymbol{\mu}_t - \boldsymbol{\mu}) &\geq \frac{1}{2\eta}\big(\|\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}\|^2 - \|\boldsymbol{\mu}_t - \boldsymbol{\mu}\|^2\big) \\ &\quad - \frac{\eta}{2}\big\|\nabla_{\boldsymbol{\mu}}\widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)\big\|^2.\end{aligned} \tag{51}$$

Since $\widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)$ is concave with respect to $\boldsymbol{\mu}_t$, we have

$$\widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t) - \widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}) \geq \nabla_{\boldsymbol{\mu}}\widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)^\top(\boldsymbol{\mu}_t - \boldsymbol{\mu}) \tag{52}$$

Combining the inequalities in (52) and (51), we may then write

$$\begin{aligned}\widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t) - \widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}) &\geq \frac{1}{2\eta}\big(\|\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}\|^2 - \|\boldsymbol{\mu}_t - \boldsymbol{\mu}\|^2\big) \\ &\quad - \frac{\eta}{2}\big\|\nabla_{\boldsymbol{\mu}}\widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)\big\|^2.\end{aligned} \tag{53}$$

Therefore, we obtain the relation (41) by subtracting inequality (53) from (48), which concludes the proof. ∎

Lemma 3 shows that the instantaneous Lagrangian difference 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 approximation budget ϵ_t. This stochastic decrement property is the basis for establishing convergence of Algorithm 1 when a certain constant step-size η is chosen, which is formally stated next.

Theorem 2. *Suppose the sequence $(f_t, \boldsymbol{\mu}_t)$ is generated from the update* (22)*, i.e., Algorithm 1, and Assumptions 1-5 hold. If the algorithm is run for T iterations with a constant step-size selected as $\eta = 1/\sqrt{T}$ and the approximation budget $\epsilon_t = \epsilon = P\eta^2$, where $P > 0$ is a fixed constant, then the time aggregation of the expected objective function error sequence $\mathbb{E}[R(f_t) - R(f^*)]$, with the optimum f^* defined as in* (2)*, grows sublinearly with the final iteration index T as*

$$\sum_{t=1}^{T}\mathbb{E}[R(f_t) - R(f^*)] \leq \mathcal{O}(\sqrt{T}). \tag{54}$$

Moreover, the time aggregation of the expected constraint violation of the algorithm grows sublinearly in T as

$$\sum_{j=1}^{m}\mathbb{E}\bigg[\sum_{t=1}^{T} G_j(f_t)\bigg]_+ \leq \mathcal{O}(T^{3/4}). \tag{55}$$

Proof: The proof relies on the result from Lemma 3. By expanding the expressions for $\widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu})$ and $\widehat{\mathcal{L}}_t(f, \boldsymbol{\mu}_t)$ as defined in (13) for any $f \in \mathcal{B}$ and $\boldsymbol{\mu} \in \mathbb{R}_+^m$, we have

$$\begin{aligned}&\ell(f_t(\boldsymbol{x}_t), \boldsymbol{y}_t) - \ell(f(\boldsymbol{x}_t), \boldsymbol{y}_t) + \frac{\lambda}{2}\big(\|f_t\|_{\mathcal{H}}^2 - \|f\|_{\mathcal{H}}^2\big) \\ &+ \frac{\delta\eta}{2}\big(\|\boldsymbol{\mu}_t\|^2 - \|\boldsymbol{\mu}\|^2\big) + \sum_{j=1}^{m}\big(\mu_j g_j(f_t(\boldsymbol{x}_t), \boldsymbol{y}_t) - \mu_{t,j} g_j(f(\boldsymbol{x}_t), \boldsymbol{y}_t)\big) \\ &\leq \frac{1}{2\eta}\big(\|f_t - f\|_{\mathcal{H}}^2 - \|f_{t+1} - f\|_{\mathcal{H}}^2 + \|\boldsymbol{\mu}_t - \boldsymbol{\mu}\|^2 \\ &\quad - \|\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}\|^2\big) + \frac{\epsilon_t}{\eta}\|f_t - f\|_{\mathcal{H}} + \frac{\epsilon_t^2}{\eta} \\ &\quad + \frac{\eta}{2}\big(2 \cdot \big\|\nabla_f \widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)\big\|_{\mathcal{H}}^2 + \big\|\nabla_{\boldsymbol{\mu}}\widehat{\mathcal{L}}_t(f_t, \boldsymbol{\mu}_t)\big\|^2\big),\end{aligned} \tag{56}$$

where we have $\boldsymbol{\mu}_t := (\mu_{t,1}, \cdots, \mu_{t,m})^\top$. Taking expectation over both sides of (56) and substituting in the bounds in (35) and (36) of Lemma 1, we obtain

$$\begin{aligned}&\mathbb{E}\Big[R(f_t) - R(f) + \frac{\delta\eta}{2}\big(\|\boldsymbol{\mu}_t\|^2 - \|\boldsymbol{\mu}\|^2\big) + \sum_{j=1}^{m}\big(\mu_j G_j(f_t) - \mu_{t,j} G_j(f)\big)\Big] \\ &\leq \mathbb{E}\Big[\frac{1}{2\eta}\big(\|f_t - f\|_{\mathcal{H}}^2 - \|f_{t+1} - f\|_{\mathcal{H}}^2 \\ &\quad + \|\boldsymbol{\mu}_t - \boldsymbol{\mu}\|^2 - \|\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}\|^2\big) + \frac{\epsilon_t}{\eta}\|f_t - f\|_{\mathcal{H}} + \frac{\epsilon_t^2}{\eta}\Big] \\ &\quad + \mathbb{E}\Big\{\frac{\eta}{2}\big[8X^2 \cdot \big(C_1^2 + mC_2^2 \cdot \|\boldsymbol{\mu}_t\|^2\big) + 4\lambda^2 \cdot R_{\mathcal{B}}^2 \\ &\quad + 2\big(K_1 + mC_2^2X^2 \cdot R_{\mathcal{B}}^2\big) + 2\delta^2\eta^2 \cdot \|\boldsymbol{\mu}_t\|^2\big]\Big\}.\end{aligned} \tag{57}$$

Note that $\|f_t - f\|_{\mathcal{H}}$ is bounded since both f_t and f in the ball $\mathcal{B}$ have finite Hilbert norm. It is also worth mentioning that the expectation is taken over not only the distribution of the random pair $(\boldsymbol{x}, \boldsymbol{y})$, but also the entire algorithm history $\mathcal{F}_t = \{f_u, \boldsymbol{\mu}_u\}_{u=0}^{t-1}$. Re-ordering the terms in (57) yields

$$\begin{aligned}&\mathbb{E}\Big[R(f_t) - R(f) - \frac{\delta\eta}{2}\|\boldsymbol{\mu}\|^2 + \sum_{j=1}^{m}\big(\mu_j G_j(f_t) - \mu_{t,j} G_j(f)\big)\Big] \\ &\leq \mathbb{E}\Big[\frac{1}{2\eta}\big(\|f_t - f\|_{\mathcal{H}}^2 - \|f_{t+1} - f\|_{\mathcal{H}}^2 + \|\boldsymbol{\mu}_t - \boldsymbol{\mu}\|^2 \\ &\quad - \|\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}\|^2\big) + \frac{2\epsilon_t}{\eta} \cdot R_{\mathcal{B}} + \frac{\epsilon_t^2}{\eta}\Big] \\ &\quad + \mathbb{E}\Big\{\frac{\eta}{2}\big[K + (8X^2mC_2^2 + 2\delta^2\eta^2 - \delta) \cdot \|\boldsymbol{\mu}_t\|^2\big]\Big\},\end{aligned} \tag{58}$$

where we define

$$K = 8X^2 \cdot C_1^2 + 4\lambda^2 \cdot R_{\mathcal{B}}^2 + 2\big(K_1 + mC_2^2 X^2 \cdot R_{\mathcal{B}}^2\big).$$

Next, we choose the constant parameter δ to satisfy $8X^2 m C_2^2 + 2\delta^2\eta^2 - \delta \leq 0$. This way, we can drop the term related to $\|\boldsymbol{\mu}_t\|^2$ from the right-hand side of (58). Moreover, we set the approximation budget as a constant $\epsilon_t = \epsilon$. Summing both sides of (58) over time, we obtain

$$\begin{aligned}\mathbb{E}\Big\{ \sum_{t=1}^{T} \big[R(f_t) - R(f)\big] - \frac{\delta\eta T}{2}\|\boldsymbol{\mu}\|^2 \\ + \sum_{t=1}^{T}\sum_{j=1}^{m} \left(\mu_j G_j(f_t) - \mu_{t,j} G_j(f)\right) \Big\} \qquad (59)\\ \leq \frac{1}{2\eta}\big(\|f_1 - f\|_{\mathcal{H}}^2 + \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}\|^2\big) + \frac{2\epsilon T}{\eta}\cdot R_{\mathcal{B}} + \frac{\epsilon^2 T}{\eta} + \frac{\eta K T}{2}.\end{aligned}$$

Note that the right-hand side of (59) is deterministic. We can set f in (59) to be the solution f^* of (2). Since f^* must satisfy the inequality constraint (2), the term $\mathbb{E}[\sum_{t=1}^{T}\sum_{j=1}^{m} -\mu_{t,j} G_j(f^*)] \geq 0$ holds true, and thus we can simply drop it from the left-hand side of (59). Under the initialization $f_1 = 0 \in \mathcal{H}$ and $\boldsymbol{\mu}_1 = \mathbf{0} \in \mathbb{R}_+^m$ and collecting the terms involving $\|\boldsymbol{\mu}\|^2$, we further obtain

$$\begin{aligned}\mathbb{E}\Big\{ \sum_{t=1}^{T}\big[R(f_t) - R(f^*)\big] - \Big(\frac{\delta\eta T}{2} + \frac{1}{2\eta}\Big)\cdot\|\boldsymbol{\mu}\|^2 \qquad (60)\\ + \quad \sum_{t=1}^{T}\sum_{j=1}^{m}\mu_j G_j(f_t)\Big\} \\ \leq \frac{1}{2\eta}\|f^*\|_{\mathcal{H}}^2 + \frac{\epsilon T}{\eta}\cdot(2R_{\mathcal{B}} + \epsilon) + \frac{\eta K T}{2}.\end{aligned}$$

There are three terms on the left-hand side of (60). The first is the time aggregation of the objective error; the last is inner product of an arbitrary dual variable $\boldsymbol{\mu} \in \mathbb{R}_+^m$ with the time-aggregation of constraint violation; and the second term relates to the squared norm of the dual variable $\boldsymbol{\mu}$. Thus, we can select $\boldsymbol{\mu}$ to maximize left-hand side of (60) as the *optimal* Lagrange multiplier that controls the growth of the long-term constraint violation. In particular, there exists a unique maximizer $\widetilde{\boldsymbol{\mu}} = (\widetilde{\mu}_1, \cdots, \widetilde{\mu}_m)^\top$ over the region $\mathbb{R}_+^m$, as given by

$$\widetilde{\mu}_j = \mathbb{E}\bigg\{ \frac{1}{\delta\eta T + 1/\eta} \cdot \bigg[\sum_{t=1}^{T} G_j(f_t)\bigg]_+ \bigg\},\ \forall j.$$

Hence, substituting $\boldsymbol{\mu} = \widetilde{\boldsymbol{\mu}}$ into (60), we obtain that

$$\begin{aligned}\mathbb{E}\bigg\{ \sum_{t=1}^{T}\big[R(f_t) - R(f^*)\big] + \sum_{j=1}^{m}\frac{\big[\sum_{t=1}^{T} G_j(f_t)\big]_+^2}{2(\delta\eta T + 1/\eta)} \bigg\} \\ \leq \frac{1}{2\eta}\|f^*\|_{\mathcal{H}}^2 + \frac{\epsilon T}{\eta}\cdot(2R_{\mathcal{B}} + \epsilon) + \frac{\eta K T}{2}. \qquad (61)\end{aligned}$$

Let the constant step-size be $\eta = 1/\sqrt{T}$ and approximation budget be $\epsilon = P\eta^2 = P/T$, with a fixed constant $P > 0$. Thus, we can simplify (61) as

$$\begin{aligned}\mathbb{E}\bigg\{ \sum_{t=1}^{T}\big[R(f_t) - R(f^*)\big] + \sum_{j=1}^{m}\frac{\big[\sum_{t=1}^{T} G_j(f_t)\big]_+^2}{2\sqrt{T}(\delta+1)} \bigg\} \\ \leq \frac{\sqrt{T}}{2}\big(\|f^*\|_{\mathcal{H}}^2 + 4PR_{\mathcal{B}} + 2P^2 + K\big). \qquad (62)\end{aligned}$$

The inequality (62) serves as the basis for establishing convergence in terms of both the objective function sub-optimality and the feasibility attainment for the proposed iterates.

First, consider the expected objective error sequence $\mathbb{E}[R(f_t) - R(f^*)]$. Since the second term on the left side of (62) is nonnegative, it can be subtracted and upper-estimated by null, to obtain

$$\sum_{t=1}^{T}\mathbb{E}[R(f_t) - R(f^*)] \leq \frac{\sqrt{T}}{2}\big(\|f^*\|_{\mathcal{H}}^2 + 4PR_{\mathcal{B}} + 2P^2 + K\big).$$

Clearly, this bound has the order of $\mathcal{O}(\sqrt{T})$, as stated in (54).

Second, to establish the sublinear growth of the constraint violation in T, we can bound the objective error by

$$\begin{aligned}&|R(f_t) - R(f^*)| \\ &\leq \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[|\ell(f_t(\boldsymbol{x}),\boldsymbol{y}) - \ell(f^*(\boldsymbol{x}),\boldsymbol{y})|] + \frac{\lambda}{2}\left|\|f_t\|_{\mathcal{H}}^2 - \|f^*\|_{\mathcal{H}}^2\right| \\ &\leq C_1\cdot\mathbb{E}[|f_t(\boldsymbol{x}) - f^*(\boldsymbol{x})|] + \frac{\lambda}{2}\cdot\left|\|f_t\|_{\mathcal{H}}^2 - \|f^*\|_{\mathcal{H}}^2\right|, \qquad (63)\end{aligned}$$

where the first inequality follows from triangle inequality, while the second one from the Lipschitz-continuity condition (33) of Assumption 2. Moreover, by the reproducing property of κ and Cauchy-Schwartz inequality, we obtain

$$\begin{aligned}|f^*(\boldsymbol{x}) - f_t(\boldsymbol{x})| &= |\langle f_t - f^*, \kappa(\boldsymbol{x},\cdot)\rangle| \\ &\leq \|f_t - f^*\|_{\mathcal{H}}\cdot\|\kappa(\boldsymbol{x},\cdot)\|_{\mathcal{H}}.\end{aligned}$$

Under Assumption 1, we have $\|\kappa(\boldsymbol{x},\cdot)\|_{\mathcal{H}} \leq X$, to assert that

$$\mathbb{E}[|f_t(\boldsymbol{x}) - f^*(\boldsymbol{x})|] \leq X\cdot\|f_t - f^*\|_{\mathcal{H}}. \qquad (64)$$

Under Assumption 5, we also have

$$\begin{aligned}\left|\|f_t\|_{\mathcal{H}}^2 - \|f^*\|_{\mathcal{H}}^2\right| &\leq \|f_t - f^*\|_{\mathcal{H}}\cdot\|f_t + f^*\|_{\mathcal{H}} \\ &\leq 2R_{\mathcal{B}}\cdot\|f_t - f^*\|_{\mathcal{H}} \qquad (65)\end{aligned}$$

Combining (64) and (65), we can rewrite (63) as

$$|R(f_t) - R(f^*)| \leq (C_1 X + \lambda R_{\mathcal{B}})\cdot\|f_t - f^*\|_{\mathcal{H}}. \qquad (66)$$

Using Assumption 5 again, we have $\|f_t - f^*\|_{\mathcal{H}} \leq 2R_{\mathcal{B}}$. Thus, the inequality (66) boils down to

$$\begin{aligned}R(f_t) - R(f^*) &\geq -(C_1 X + \lambda R_{\mathcal{B}})\cdot\|f_t - f^*\|_{\mathcal{H}} \\ &\geq -2R_{\mathcal{B}}(C_1 X + \lambda R_{\mathcal{B}}).\end{aligned}$$

Substituting this bound into (62), and letting $K_1 = \|f^*\|_{\mathcal{H}}^2 + 4R_{\mathcal{B}} + 2 + K$ and $K_2 = 2R_{\mathcal{B}}(C_1 X + \lambda R_{\mathcal{B}})$, we obtain

$$\mathbb{E}\bigg\{ \sum_{j=1}^{m}\frac{\big[\sum_{t=1}^{T} G_j(f_t)\big]_+^2}{2\sqrt{T}(\delta+1)} \bigg\} \leq \frac{\sqrt{T}}{2}K_1 + K_2 T. \qquad (67)$$

Re-ordering the terms, we further obtain

$$\sum_{j=1}^{m}\mathbb{E}\bigg[\sum_{t=1}^{T}G_j(f_t)\bigg]_+^2\leq 2\sqrt{T}(\delta+1)\left(\frac{\sqrt{T}}{2}K_1+TK_2\right). \quad (68)$$

Noticing that $\sum_{j=1}^{m}Z_j^2\cdot m^{-1}\geq\big(\sum_{j=1}^{m}Z_j\cdot m^{-1}\big)^2$ for any $\{Z_j\}_{j=1,\cdots,m}$, we have

$$m\cdot\sum_{j=1}^{m}\mathbb{E}\bigg[\sum_{t=1}^{T}G_j(f_t)\bigg]_+^2\geq\mathbb{E}\bigg\{\sum_{j=1}^{m}\bigg[\sum_{t=1}^{T}G_j(f_t)\bigg]_+\bigg\}^2.$$

Then, by Jensen's inequality, we further have

$$\begin{aligned}\sum_{j=1}^{m}\mathbb{E}\bigg[\sum_{t=1}^{T}G_j(f_t)\bigg]_+^2&\geq\frac{1}{m}\cdot\mathbb{E}\bigg\{\sum_{j=1}^{m}\bigg[\sum_{t=1}^{T}G_j(f_t)\bigg]_+\bigg\}^2\\&\geq\frac{1}{m}\cdot\bigg\{\mathbb{E}\sum_{j=1}^{m}\bigg[\sum_{t=1}^{T}G_j(f_t)\bigg]_+\bigg\}^2. \quad (69)\end{aligned}$$

Combining (68) and (69), and taking the square root of both sides, we obtain the sublinear rate in (55). ∎

Theorem 2 establishes the result 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. These results are akin to originally established mean convergence to $\mathcal{O}(\eta T)$ and $\mathcal{O}(\eta T^{5/4})$ error neighborhoods in terms of primal sub-optimalty and constraint violation [41], where the radius of these neighborhoods may be minimized by an appropriately chosen step-size $\eta=1/\sqrt{T}$. We present results in this fashion to make clear the conceptual link between mean convergence behavior of stochastic algorithms and regret analysis of online learning [17]. 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. *Suppose that Assumptions 1-5 hold, and Algorithm 1 is run for T iterations with a constant step-size selected as $\eta=1/\sqrt{T}$ and the approximation budget $\epsilon_t=\epsilon=P/T$ for a fixed constant $P>0$. For $\overline{f}_T=\sum_{t=1}^{T}f_t/T$ as the functional formed by averaging the primal iterates f_t over time $t=1,\cdots,T$, its objective function satisfies*

$$\mathbb{E}[R(\overline{f}_T)-R(f^*)]\leq\mathcal{O}(1/\sqrt{T}). \quad (70)$$

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

$$\sum_{j=1}^{m}\mathbb{E}\Big[\big(G_j(\overline{f}_T)\big)\Big]_+\leq\mathcal{O}(T^{-1/4}). \quad (71)$$

Proof: This proof builds on Theorem 2. Specifically, (70) may be obtained by dividing by T on both sides of (70), and applying the convexity of $R(f)$ based on Assumption 3:

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

Similarly, by convexity of $G(f)$, we have that

$$\begin{aligned}\sum_{j=1}^{m}\mathbb{E}\Big[G_j(\overline{f}_T)\Big]_+&\leq\sum_{j=1}^{m}\mathbb{E}\bigg[\sum_{t=1}^{T}G_j(f_t)\cdot T^{-1}\bigg]_+\\&=\sum_{j=1}^{m}\mathbb{E}\bigg[\sum_{t=1}^{T}G_j(f_t)\bigg]_+\cdot T^{-1}\leq\mathcal{O}(T^{-1/4}),\end{aligned}$$

which completes the proof. ∎

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 for the constraint violation bound. 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 [38] or the stochastic setting for vector-space constrained convex optimization [27]. One departing feature of the RKHS setting is that by averaging f_t over time, its model order may be unbounded; thus, Corollary 1 is a theoretical result solely for interpreting the convergence bounds of Theorem 2, as such time-averaging 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 be able to 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 [14] 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.

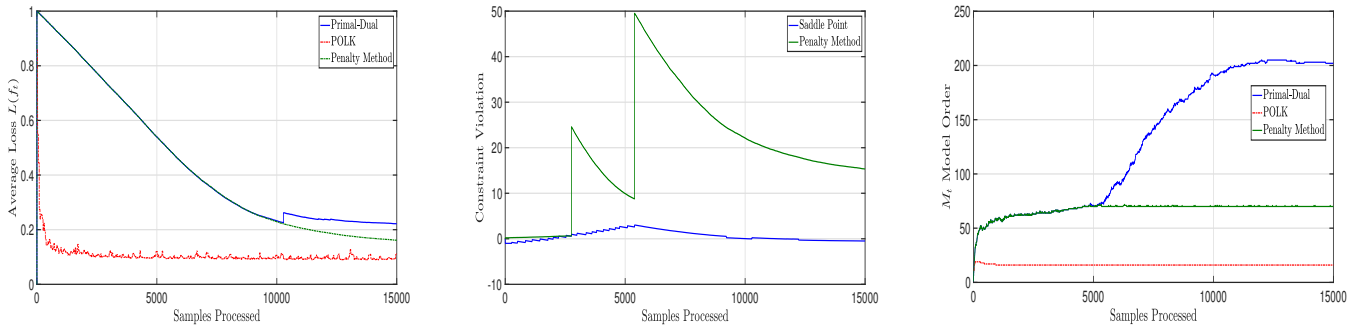
Corollary 2. *Suppose the sequence $(f_t,\boldsymbol{\mu}_t)$ is generated by Algorithm 1 under constant step-size $\eta=1/\sqrt{T}$ and approximation budget $\epsilon=P\eta^2$ where $P>0$ is a fixed constant. For the model order M_t of function f_t, there exists a finite upper bound M^∞ such that $M_t\leq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

Thus, Algorithm 1 solves 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.

V. Experiments

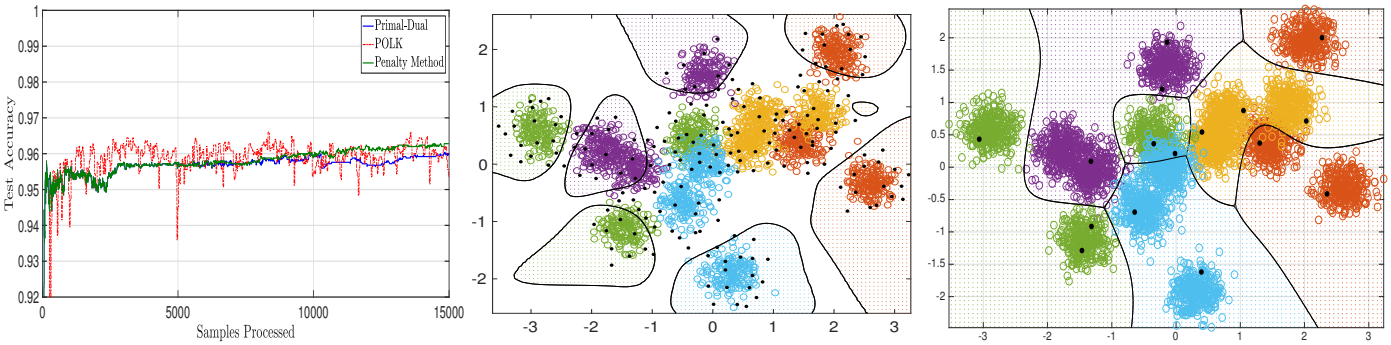
We now turn to numerically evaluating our proposed method for solving constrained stochastic optimization problems in RKHS. We focus on the the risk-aware supervised learning with CVaR constraints as stated in Example 1. This constraint is used to mitigate the unknown variance of the modeling hypothesis that $f\in\mathcal{H}$, also known as the approximation error in statistical learning [8]. We consider two different instantiations under this problem formulation.

Multi-class Kernel Support Vector Machines (SVM): In Kernel SVM (KSVM), the merit of a particular function is



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

Fig. 1: Algorithm 1 for kernel SVM [cf. (72)] with CVaR constraints (3) (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 function and the constraint. Smaller step-sizes are required for constrained versus unconstrained problems. The objective and constraint violation converge to null and the model order stabilizes. We compare with an unconstrained projected FSGD based algorithm POLK [14] and a penalty method [20] where the penalty coefficient doubles every 200 iterations. The comparators converge to lower model complexity, albeit *infeasible*, solutions.



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

Fig. 2: Algorithm 1 for KSVM with objective in (72) 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.

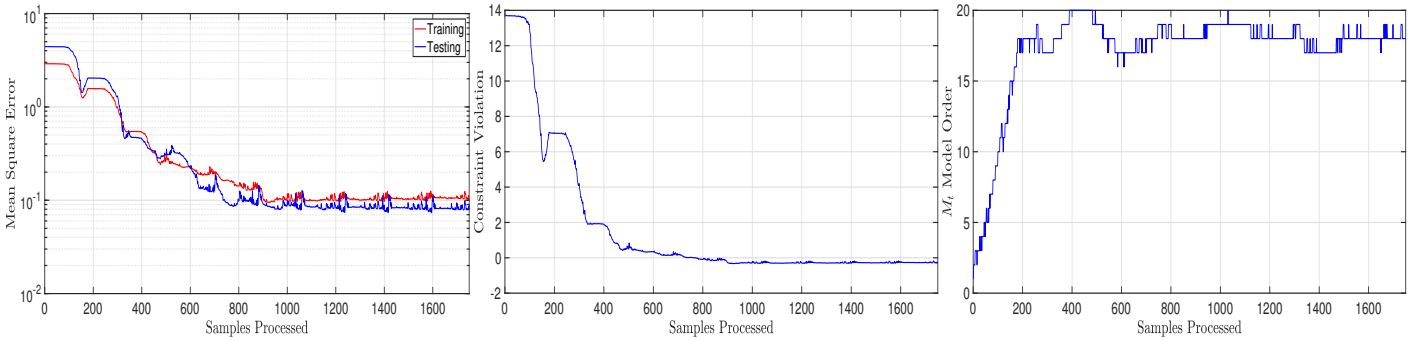
defined by its ability to maximize its classification margin. Define a set of class-specific activation functions $f_c : \mathcal{X} \to \mathbb{R}$, jointly denoted as $\mathbf{f} \in \mathcal{H}^C$. In multi-KSVM, points are assigned to the class label of the activation function that yields the maximum response. KSVM is trained by specifying the loss to be the multi-class hinge function which defines the margin separating hyperplanes in the kernelized feature space:

$$\ell(\mathbf{f}(\mathbf{x}_n), y_n) = \max\big(0, 1 + f_r(\mathbf{x}_n) - f_{y_n}(\mathbf{x}_n)\big) + \lambda \sum_{c'=1}^{C} \|f_{c'}\|_{\mathcal{H}}^2 , \qquad (72)$$

where $r = \operatorname{argmax}_{c' \neq y_n} f_{c'}(\mathbf{x})$. Further details can be found in [42]. We test Algorithm 1 for this setting on a synthetic data set, where data vectors are $p = 2$ dimensional, drawn from a set of Gaussian mixture models similar to [43]. Each label y_n is first drawn randomly and uniformly from the label set. The corresponding data point $\mathbf{x}_n \in \mathbb{R}^p$ is then drawn from an equitably-weighted Gaussian mixture model, i.e., $\mathbf{x} \,|\, y \sim (1/3) \sum_{j=1}^{3} \mathcal{N}(\boldsymbol{\mu}_{y,j}, \sigma_{y,j}^2 \mathbf{I})$ where $\sigma_{y,j}^2 = 0.2$ for all values of y and j. Additionally, $\boldsymbol{\mu}_{y,j}$ are realizations of 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 $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_C\}$ are equitably spaced around the unit circle, one for each class label, with unit variance $\sigma_y^2 = 1.0$. The number of classes is fixed at $C = 5$ and thus the feature distribution has 15 distinct modes. The data set consists of $N = 5000$ feature-label pairs for training and additional 2500 pairs for testing.

We run the algorithm for three training epochs, i.e., $T = 15000$, with a Gaussian kernel, whose bandwidth is $\sigma = 0.3$. Moreover, the algorithm step-size is $\eta = 0.009$, with approximation budget $\epsilon = P\eta^2$ using a parsimony constant at $P = 3.7$ and a mini-batch size of 4. The primal regularizer has $\lambda = 10^{-4}$ and the dual regularizer $\delta = 10^{-4}$. The significance level of CVaR_α is $\alpha = 0.9$ and the tolerance is set to $\gamma = 2$. This enforces more conservative learning and avoiding moving the regression function in directions that could cause the loss function to spike with prob. less than $1 - \alpha = 0.1$.

The results of this experiment are given in Fig. 1, comparing the proposed primal-dual based Algorithm 1 with the unconstrained FSGD based counterpart (POLK) [14] and its penalized variant [27]. The statistical average loss converges



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

Fig. 3: Kernel ridge regression/nonlinear filtering of LIDAR DATA: $\mathbf{x}_n$ is a scalar LIDAR-based range scan and $\mathbf{y}_n$ is the ground truth position of a robot near a wall [28]. Again, we use the CVaR constraint to mitigate learning volatility. We run Algorithm 1 with a Gaussian kernel at bandwidth $\sigma = 0.04$ with step-size $\eta = 0.1$, and parsimony constant $P = 0.008$. The training and test mean square errors converge to small constants in Fig. 3a; the constraint violation settles to null in Fig. 3b; and the model complexity remains small in Fig. 3c.

to a small constant as the number of samples increases (Fig. 1a), while the infeasibility initially spikes and then settles to feasibility (Fig. 1b). Meanwhile,the model complexity remains under control (Fig. 1c). Jumps in objective function 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 the kernel dictionary elements; curved lines denote high-confidence decision boundaries, which are far from areas of class overlap due to their likelihood of causing loss spikes. Decisions made in areas of overlap are still correct, but the learning agent recognizes the risk. On the contrary, POLK and its penalized variants converge to more accurate solutions but cannot handle constraints, yielding infeasible solutions, and thus riskier decisions.

Kernel Ridge Regression (Nonlinear Filtering): We further consider the problem of kernel ridge regression, in which case the loss is the squared mismatch error:

$$\ell(f(\mathbf{x}_n), y_n) = (f(\mathbf{x}_n) - y_n)^2 \tag{73}$$

where $\mathbf{x}_n$ is the data vector and $\mathbf{y}_n \in \mathbb{R}$ is the target variable. We use a standard LIDAR data set for this application [28]. Both $\mathbf{x}_n$ and $\mathbf{y}_n$ are scalars: $\mathbf{x}_n$ denotes LIDAR-based range scans, while $\mathbf{y}_n$ the ground truth position of a robot near a wall. The training sample size is 179 and the test set consists of 22 hold out data points. Again, we use the CVaR constraint to mitigate the volatility of the learning process, and control the intrinsic error variance of our modeling hypothesis $f \in \mathcal{H}$.

We run the algorithm for ten training epochs, i.e., $T = 1790$, with a Gaussian kernel, whose bandwidth is $\sigma = 0.04$. No mini-batching is used here. Moreover, the algorithm step-size is $\eta = 0.1$, with approximation budget $\epsilon = P\eta^2$ using a parsimony constant at $P = 0.008$. The primal regularizer is set as $\lambda = 10^{-5}$ and the dual regularizer as $\delta = 10^{-5}$. The significance level of CVaR_α is $\alpha = 0.99$ and the tolerance is set to $\gamma = 0.8$, meaning we avoid possible loss spikes with probability less that $1 - \alpha = 0.01$.

The results of this implementation can be seen in Fig. 3: the mean square error values for both training and test stages converge to similarly small level as the number of samples increases (Fig. 3a), while the constraint violation settles to null (Fig. 3b). Meanwhile, the model complexity remains under control (Fig. 3c). Volatility intrinsic to online training has been effectively mitigated using the CVaR constraint.

VI. Conclusion

In this work, we have considered the function-valued stochastic optimization problem with nonlinear constraints, motivated by applications to risk-aware supervised learning, navigation with obstacle-avoidance constraints, and wireless communications. We considered the case where functions belong to a reproducing Kernel Hilbert space and thus admit a basis expansion in terms of the observed data through the Representer Theorem. First, we extended the Representer Theorem to saddle-point problems over RKHS through the definition of a modified empirical loss that incorporates constraints. We then developed a saddle-point algorithm that operates by alternating primal/dual projected stochastic gradient descent/ascent steps on the augmented Lagrangian of the optimization problem. The primal projection sets are function subspaces that are greedily constructed from a subset of past observed data using matching pursuit.

By tuning the approximation budget to the algorithm step-size, and by selecting both as fixed small constants, we established convergence in expectation of both the objective function error sequence and the constraint violation to fixed error neighborhoods. The size of the neighborhood depends on the chosen step-size and the final algorithm index. This result generalizes existing guarantees of primal-dual method in constrained stochastic programs with vector-valued variables to function spaces. We experimentally validated this method for the task of supervised learning with risk constraints, both for kernel support vector machines and ridge regression. As future work, we hope to investigate how the methods developed here may be used for new approaches to trajectory optimization based on sensory observations.

References

[1] K. Zhang, H. Zhu, T. Başar, and A. Koppel, "Projected stochastic primal-dual method for constrained online learning with kernels," in *2018 IEEE*

57th Annual Conference on Decision and Control (CDC) (submitted). IEEE, 2018.

[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, University of Michigan, Ann Arbor, Michigan, USA, June 18 - June 22, 2016*, 2016. [Online]. Available: http://www.roboticsproceedings.org/rss12/p46.html

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

[9] C. Richter, A. Bry, and N. Roy, "Polynomial trajectory planning for aggressive quadrotor flight in dense indoor environments," in *Robotics Research*. Springer, 2016, pp. 649–666.

[10] 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.

[11] C. E. Rasmussen, "Gaussian Processes in Machine Learning," in *Advanced Lectures on Machine Learning*. Springer, 2004, pp. 63–71.

[12] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, 1994.

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

[14] 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.

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

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

[17] 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.

[18] R. Jenatton, J. Huang, D. Csiba, and C. Archambeau, "Online optimization and regret guarantees for non-additive long-term constraints," *arXiv preprint arXiv:1602.05394*, 2016.

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

[20] 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.

[21] 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.

[22] R. T. Rockafellar, S. Uryasev *et al.*, "Optimization of conditional value-at-risk," *Journal of Risk*, vol. 2, pp. 21–42, 2000.

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

[24] 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.

[25] K. Arrow, L. Hurwicz, and H. Uzawa, *Studies in Linear and Non-Linear Programming*, ser. Stanford Mathematical Studies in the Social Sciences. Stanford University Press, Stanford, Dec. 1958, vol. II.

[26] 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.

[27] 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.

[28] D. Ruppert, M. P. Wand, and R. J. Carroll, "Semiparametric regression during 2003–2007," *Electronic Journal of Statistics*, vol. 3, p. 1193, 2009.

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

[30] S. Paternain and A. Ribeiro, "Safe online navigation of convex potentials in spaces with convex obstacles," in *Decision and Control (CDC), 2017 IEEE 56th Annual Conference on*. IEEE, 2017, pp. 2473–2478.

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

[32] 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.

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

[34] 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.

[35] V. Norkin and M. Keyzer, "On stochastic optimization and statistical learning in reproducing kernel Hilbert spaces by support vector machines (SVM)," *Informatica*, vol. 20, no. 2, pp. 273–292, 2009.

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

[37] 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.

[38] 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.

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

[40] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[41] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

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

[43] 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.