PROJECTED PSEUDO-MIRROR DESCENT IN REPRODUCING KERNEL HILBERT SPACE

Abhishek Chakraborty**, Ketan Rajawat*[§]*, and Alec Koppel*[†]

* NetApp India; [§] Dept. of EE, IIT Kanpur; [†] CISD, U.S. Army Research Laboratory

ABSTRACT

We consider expected risk minimization for strongly convex costs for the case that the decision variable belongs to a Reproducing Kernel Hilbert Space (RKHS) and its target domain is required to be nonnegative. This arises, e.g., in intensity estimation of inhomogeneous point processes. To solve it, we develop a variant of stochastic mirror descent that employs (i) *pseudo-gradients* and (ii) projections. Compressive projections are executed via kernel orthogonal matching pursuit (KOMP), which overcomes the fact that RKHS parameterizations grows unbounded with time. Moreover, pseudo-gradients are needed, e.g., when stochastic gradients themselves define integrals over quantities that must be evaluated numerically. We derive accuracy/complexity tradeoffs between convergence in mean and bounds on the model complexity of the learned functions under standard assumptions. Experiments then demonstrate favorable tradeoffs for inhomogeneous Poisson Process intensity estimation on real data.

1. INTRODUCTION

Nonnegative function fitting arises trajectory optimization [2], as well as unsupervised [3] and supervised learning [4] for costs associated with a negative log-likelihood. In this work, we consider the nonnegative function estimation problems where the cost is strongly convex and depends on sequentially observed samples from an unknown distribution, and the feasible set is a reproducing kernel Hilbert Space (RKHS) [5]. Our main motivation is efficient estimation of the intensity of an inhomogeneous Poisson Process [6].

For the moment, let's set aside non-negativity and suppose the function class is defined by a linear statistical model. Then, the problem reduces to a convex program, which, when analytical solutions are unavailable, may be solved with Newton or gradient methods to global optimality, assuming the gradient is computable [7] . However, doing so breaks down for costs that depend on sequentially observed samples [8], as in *expected risk minimization*. In this case, stochastic approximations are necessary [9], which use stochastic gradient updates, and whose performance is tethered to the properties of the unknown data distribution. Moreover, one may only guarantee their performance probabilistically [10]. Here we put forth a stochastic variant of proximal gradient [11], i.e., stochastic mirror descent, which employs Bregman divergences to refine the convergence of first-order stochastic methods for structured problems [12].

The strength of guarantees for linear models belies the fact that they are outperformed by deep neural networks (DNNs) [13] and kernel methods [14] across disparate domains [15]. We focus on RKHS since (i) under suitable choice of kernel, they may be equivalent to certain DNNs [16]; (ii) their training defines a convex program over a RKHS [17]; and (iii) one may impose structural hypotheses through choice of kernel or Bregman divergence [6, 18]. Unfortunately, via the Representer Theorem [19], their complexity scales with the sample size, which may be large-scale. Myriad approaches exist for RKHS approximations: matrix factorization [20], spectral methods [21, 22], random feature techniques [23–25], and subspace projections [26,27]. Notably, fixing the projection-induced error rather than model complexity is favorable both in theory and practice [28], and hence we adopt it to sparsify RKHS updates.

We focus on nonnegative functions, as is inherent to the intensity of an inhomogeneous point processes [6, 29], unsupervised/supervised learning when the cost is a negative log-likelihood [12], and trajectory optimization [2]. Two intertwined challenges then emerge: constraint satisfaction, and gradient evaluation. Suitable initialization and specifying the divergence as the I-Divergence (or KL [30]) ensures iterates exhibit nonnegativity. However, the stochastic gradient itself may require evaluation of an unknown integral [31]. In this case, *pseudo-gradients* [32], i.e., descent directions correlated with true gradients, may be employed to positive effect, that is, pseudo-mirror descent has recently achieved the state of the art in point processes intensity estimation [31]. However, doing so results in a function sequence exhibits intractable complexity growth with time. In this work, we overcome this issue by designing projected variants (Sec. 3), whose tunable tradeoffs between convergence accuracy and complexity we theoretically characterize (Sec. 4). Further, outperform existing approaches on a Poisson Process intensity estimation problem on a real NBA data set (Sec. 5).

2. PROBLEM FORMULATION

Consider the problem of expected risk minimization in the online setting: independent identically distributed (i.i.d.) training samples $\{\mathbf{x}_t\}_{t\geq 1}$ are observed in a sequential manner, where $\mathbf{x}_t \in \mathcal{X} \subset \mathbb{R}^d$. The goal is to fit a predictive model f that belongs to the hypothesis space $\mathcal{H}$. The target domain is defined by a likelihood model, which induces a loss function $\ell(f(\mathbf{x}))$, typically defined as the negative log-likelihood of a probabilistic model, which we seek to minimize in expectation over an unknown distribution $\mathbb{P}(\mathbf{x})$, i.e., $R(f) := \mathbb{E}[\ell(f(\mathbf{x}))]$, as may be specified by the stochastic program

$$f^{\star} = \underset{f\in\mathcal{H}_+}{\arg\min}\, R(f) \tag{1}$$

where $R(f)$ is λ-strongly convex. For ease of notation, from now onwards, we write the instantaneous costs as $r_t(f) := \ell(f(\mathbf{x}_t))$ corresponding to the data point $\mathbf{x}_t$. We further denote the class of functions with nonnegative range as $\mathcal{H}_+$, which we hypothesize is a subset of a Hilbert space $\mathcal{H}$ associated with symmetric positive definite kernel κ which satisfies (i) $\mathcal{H} := \overline{\text{span}(\kappa(\mathbf{x},\cdot))}$; and (ii) $\langle f, \kappa(\mathbf{x},\cdot)\rangle = f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. That is, every $f \in \mathcal{H}$ can be written as linear combination of the kernel evaluations and satisfy the reproducing property, making $\mathcal{H}$ a reproducing kernel Hilbert space (RKHS) [5]. Examples kernels include the Gaussian $\kappa(\mathbf{x},\mathbf{x}') =$

*Abhishek did this work as a part of his M.Tech thesis at IIT Kanpur.

Supported by the SMART Scholarship for Service and DCIST CRA. All proofs are available to reviewers at [1].

$\exp(-\left\|\mathbf{x}-\mathbf{x}'\right\|^2/2c)$ and polynomial $\kappa(\mathbf{x},\mathbf{x}') = (\mathbf{x}^\top\mathbf{x}+b)^c$ as well as sophisticated data-dependent convolutional kernels [33].

With the problem setting and the function class over which we search made clear, we expand upon implications of the restriction that the range of the functions f is nonnegative. As mentioned, nonnegativity inherently arises in probability density function (pdf) estimation and intensity estimation, which we introduce next.

Example 1. *Poisson Point Process intensity estimation:* Poisson Processes are a family of probabilistic models that counts the number of events $N(\mathcal{T})$ up to an interval $\mathcal{T} \subset \mathcal{S} \subset \mathbb{R}^d$, which are widely used in spatial statistics [29], time-series [3], and queueing [34]. A fundamental question that arises in its use is the intensity parameter $\lambda(\cdot)$, which determines the rate $\lambda(s)$ at which new events occur in an infinitesimal time-increment, i.e., $\lambda(s) = \lim_{\Delta s\to 0}\mathbb{E}[N(\Delta s)]/(\Delta s)$. In inhomogeneous cases, this parameter varies with its argument as a nonlinear function, where the likelihood of a collection of Poisson points $\{\mathbf{t}_n\}_{n=1}^N$ takes the form:

$$L(f) = \prod_{n=1}^{N} \lambda(\mathbf{t}_n)\exp\left\{-\int_{\mathcal{S}}\lambda(\mathbf{t})d\mathbf{t}\right\} \tag{2}$$

Then, one may construct an instantiation of (1) by considering the negative log-likelihood of (2), inspired by [6]:

$$R(f) = -\sum_{n=1}^{N}\log(\lambda(\mathbf{t}_n)) + \int_{\mathcal{S}}\lambda(\mathbf{t})d\mathbf{t} \tag{3}$$

where one may identify that the Poisson points $\mathbf{t}_n$ play the role of $\mathbf{x}_n$, and $\lambda(\cdot)$ is the known function $f(\cdot)$ we seek to estimate, which is required to be nonnegative. This problem has been studied in the offline and online settings, i.e., when $\{\mathbf{t}_n\}_{n=1}^N$ are available all at once or in an incremental fashion, respectively, in [6] and [31]. In this work, we develop online approaches to developing sparse solutions to (3). Further derivation details may be found in [3]

We focus on solving (1) via search directions that move in the interior of the (generalized) probability simplex. First, we close the section with a clarifying remark about the complexity of RKHS.

Remark 1. (Empirical Risk Minimization) An important special case of (1) is *empirical risk minimization* (ERM) where a fixed collection of data $\mathcal{D} := \{(\mathbf{x}_i)_{i=1}^N\}$ is available, and we seek to find the optimizer of the empirical loss $\hat{f}_N = \arg\min_{f\in\mathcal{H}}\frac{1}{N}\sum_{m=1}^N r_i(f)$ Observe that $\hat{f}_N$ in the search space is a Hilbert space, and hence is infinite-dimensional. The Representer Theorem [35, 36] implies, however, that $\hat{f}_N$ takes the form: $\hat{f}_N(\cdot) = \sum_{m=1}^N w_m\kappa(\mathbf{x}_m,\cdot)$ where $\{w_m\}_{m=1}^N$ are scalar weights, which via substitution into the empirical loss, simplifies search over $\mathcal{H}$ to $\mathbb{R}^N$:

$$\hat{\mathbf{w}}_N = \arg\min_{\mathbf{w}}\frac{1}{N}\sum_{i=1}^{N}\ell(\mathbf{w}^\top\mathbf{k}_{\mathcal{D}}(\mathbf{x}_i))\ , \tag{4}$$

where we have collected kernel evaluations $\{\kappa(\mathbf{x}_m,\mathbf{x}_n)\}_m$ into a vector called the empirical kernel map $\mathbf{k}_{\mathcal{D}}(\mathbf{x}_m) \in \mathbb{R}^N$ and $\{\kappa(\mathbf{x}_m,\mathbf{x}_n)\}_{m,n}$ into the Gram, or kernel, matrix $\mathbf{K}_{\mathcal{D}\mathcal{D}} \in \mathbb{R}^{N\times N}$. As the sample size $N\to\infty$ grows, the dimensionality of the search space $\mathbb{R}^N$ grows as well, an instance of the curse of kernelization. Therefore, it is not enough to solve (1) to optimality, but one must do so while also ensuring the complexity M of the function representation in terms of w_m and $\mathbf{x}_m$ is efficient as well, i.e., the $M \ll N$. We precisely define the model order M, in later sections.

3. ALGORITHM FORMULATION

Now we shift focus to deriving an iterative approach to solving (1) based upon functional extension of mirror descent [12]. We employ specific choice of Bregman divergences that ensure positivity of the range of the function f during optimization, as well as generalization of the gradients employed in the updates to "pseudo-gradients" put forth in [31], which are useful in point process intensity estimation.

Bregman Divergence We begin by presenting technicalities pertaining to the Bregman divergence and mirror map before deriving the proposed algorithm. Let $\psi : \mathcal{H}\to\mathbb{R}$ be a proper, closed, smooth, and strongly convex functional. The Frenchel conjugate of ψ is denoted as $\psi^* : \mathcal{H}^*\to\mathbb{R}$, where $\mathcal{H}^*$ is the dual space of $\mathcal{H}$. Define the shorthand for the objective evaluated at the gradient of the dual $R_\psi(z) = (R\circ\nabla\psi^*)(z) = R(\nabla\psi^*(z))$ where $z\in\mathcal{H}^*$. This composition allows one to write $\nabla R_\psi(\nabla\psi(f)) = \nabla R(f)$, $f\in\mathcal{H}$ since $\nabla\psi^* = (\nabla\psi)^{-1}$. This identity is used to establish convergence. The purpose of defining functional ψ is that it induces a distance-like functional Bregman divergence $B_\psi : \mathcal{H}\times\mathcal{H}\Rightarrow\mathbb{R}$ [37]:

$$B_\psi(f,\tilde{f}) := \psi(f) - \psi(\tilde{f}) - \langle\nabla\psi(\tilde{f}), f-\tilde{f}\rangle_{\mathcal{H}}. \tag{5}$$

The functional Bregman divergence satisfies most of the properties of its vector-valued counterpart: non-negativity, strong-convexity in the first argument, and a generalized Pythagorean theorem. A few common examples are *Squared difference*, *Squared Mahalanobis difference* and *KL-divergence or I-divergence* – see [37]. In this paper, we restrict focus to the KL-divergence or I-divergence, whose convex map $\psi(f) = \langle f, \log(f)-1\rangle_{\mathcal{H}}$ and $B_\psi(f,\tilde{f}) = \langle f, \log(f/\tilde{f})\rangle_{\mathcal{H}}$.

Pseudo-Gradient We shift to defining search directions for the objective (1) called pseudo-gradients: directions g_t that have positive (unnormalized) cosine similarity with gradient $\nabla_f R(f)$ in expectation [32] (any direction forming an acute angle with the original gradient $\nabla R(f_t)$ in the dual space):

$$\langle\nabla R(f_t), \mathbb{E}[g_t|\mathcal{F}_t]\rangle \geq 0 \tag{6}$$

where $\mathcal{F}_t$ denotes the past sigma algebra which contains all the past data points one iteration back, i.e. $\mathcal{F}_t = \sigma\{\mathbf{x}_i\}_{i=1}^{t-1}$, which may be employed when evaluating the exact gradient is possibly unavailable. *Stochastic Gradients*, *Kernel embeddings* and *Gradient sign* are examples, whose specific forms are omitted due to spatial constraints.

With Bregman divergence and pseudo-gradients defined, we shift to presenting our algorithmic solution to solving (1) via streaming samples $\{x_t\}$, which is built upon a functional generalization of stochastic mirror descent: $f_{t+1} = \arg\min_{f\in\mathcal{H}}\Big(\langle\nabla r_t(f_t), f\rangle_{\mathcal{H}} + \frac{1}{\eta}B_\psi(f,f_t)\Big)$, where η is a nonnegative constant step-size. Note that for squared difference, $B_\psi(f,\tilde{f}) = \frac{1}{2}\|f-\tilde{f}\|^2$, this reduces to functional stochastic gradient method. For this update to be tractable, we require evaluation of Bregman divergence in closed-form.

We propose the use of pseudo-gradients g_t in lieu of stochastic gradients $\nabla r_t(f_t)$ in mirror descent, which, for instantaneous loss is at function iterate f_t with data points $\mathbf{x}_t$, Functional Pseudo Mirror Descent (FPMD) takes the form:

$$\tilde{f}_{t+1} = \arg\min_{f\in\mathcal{H}}\Big(\langle g_t, f\rangle_{\mathcal{H}} + \frac{1}{\eta}B_\psi(f,f_t)\Big). \tag{7}$$

This update, which lives in a Hilbert space, may be executed parametrically in some instances. To see this, define $\tilde{z}_{t+1} = \nabla\psi(\tilde{f}_{t+1})$ and $z_t = \nabla\psi(f_t)$, which implies $f_t = \nabla\psi^*(z_t)$. For general pseudo-gradients, i.e. when $g_t \neq \nabla r_t(f_t)$, one may not easily find a parametric form for f_t by inverting the Bregman divergence. [1] However, for differentiable pseudo-gradients, we have $g_t = g_t'\kappa(\mathbf{x}_t, \cdot)$ (via the chain rule and reproducing property of the kernel). The specific g_t' depends on whether the pseudo-gradient is defined by, e.g., a kernel embedding, gradient sign, or stochastic gradient ($g_t' = \ell'(\nabla\psi^*(z_t(\mathbf{x}_t)))$). Thus, one may execute (7) as

$$\tilde{z}_{t+1} = z_t - g_t'\kappa(\mathbf{x}_t, \cdot) \tag{8}$$

with corresponding dictionary and weight updates:

$$\mathcal{D}_{z,t+1}=\mathcal{D}_{z,t}\cup\{\mathbf{x}_t\}\ ,\ [\mathbf{w}_{z,t+1}]_n=\begin{cases}[\mathbf{w}_{z,t}]_n & \mathbf{x}_n\in\mathcal{D}_t\\ -\eta g_t' & \mathbf{x}_n=\mathbf{x}_t\end{cases} \tag{9}$$

where $\mathcal{D}_{z,t+1}$ represents the set of dictionary points for function $\tilde{z}_{t+1}$ and $[\mathbf{w}_{z,t}]_n$ denotes the n-th coordinate of the vector $\mathbf{w}_{z,t}$. Due to the RKHS parameterization in terms of weights and feature vectors $\mathbf{x}_t$, the complexity of the $\tilde{z}_{t+1}$ grows unbounded with time t. We address this issue via subspace projections greedily constructed with kernel orthogonal matching pursuit (KOMP) [26, 38]. Specifically, given an input dictionary $\tilde{\mathcal{D}}_{z,t+1}$ and weight vector $\tilde{\mathbf{w}}_{z,t+1}$, returns lower-dimensional (compressed) dictionary and weights

$$\{z_{t+1},\mathcal{D}_{z,t+1},\mathbf{w}_{z,t+1}\}=\text{KOMP}(\tilde{z}_{t+1},\tilde{\mathcal{D}}_{z,t+1},\tilde{\mathbf{w}}_{z,t+1},\epsilon) \tag{10}$$

that are ϵ-away in the RKHS norm, where ϵ denotes the compression budget, which we call SPPPOT: **S**parse **P**ositive Functions via **P**rojected **P**seudo-Mirr**o**r Descen**t**. Here use of KOMP differs from [28]: the RKHS-norm approximation is in terms of the dual space via auxiliary sequence $\{z_t\}$ (8).

Moreover, function f at $\mathbf{x} \in \mathcal{X}$ at time $t+1$ is given as

$$f_{t+1}(\mathbf{x}) = \nabla\psi^*(z_{t+1}(\mathbf{x})) = \nabla\psi^*(\mathbf{w}_{z,t+1}^\top \mathbf{k}_{\mathcal{D}_{z,t+1}}(\mathbf{x})) \tag{11}$$

which takes the form $f_{t+1}(\cdot)=\exp(\mathbf{w}_{z,t+1}^\top\mathbf{k}_{\mathcal{D}_{z,t+1}}(\cdot))$ for KL Divergence. Crucial to our approach is the range of the exponential: if f_0 is initialized as positive, and with projections due to KOMP executed on the dual space, then positivity is preserved via (11). Next we discuss the convergence of (10).

4. CONVERGENCE AND COMPLEXITY ANALYSIS

We now present the conceptual aspects of (10) for solving (1) for a fixed compression budget ϵ and step-size η. The pseudo gradient is expressed as $g_t = \frac{1}{\eta}(z_t - \tilde{z}_{t+1})$, whereas the *projected* pseudo-gradient $\hat{g}_t := \frac{1}{\eta}(z_t - z_{t+1})$ is obtained by the application of KOMP on the auxiliary function $\tilde{z}_{t+1}$. Hence the projected auxiliary function iterates becomes $z_{t+1} = z_t - \eta\hat{g}_t$. To establish the convergence, we require the following technical conditions.

A1. The inner product between the gradient and the expectation of the pseudo-gradient given the filtration $\mathcal{F}_t = \sigma(\{\mathbf{x}_i\}_{i=1}^{t-1})$, is lower bounded by the second-moment of the gradient in the dual norm for positive constant D:

$$\mathbb{E}[\langle\nabla R(f_t), \mathbb{E}[g_t|\mathcal{F}_t]\rangle] \geq D\mathbb{E}[||\nabla R(f_t)||_*^2] \tag{12}$$

A2. The problem (1) has finite solution with $\|f^\star\|^2 \leq B$.

A3. The instantaneous and average costs R are λ-strongly convex.

A4. The function $R_\psi(\cdot)$ is L_1-Lipschitz smooth.

A5. For all $f \in \mathcal{H}$, $t \in \mathbb{N}$, and some $c \geq 1$, the instantaneous gradient $\nabla r_t(f)$ is unbiased and has variance bounded by

$$\mathbb{V}\left[\nabla r_t(f)\right] = \mathbb{E}\|\nabla r_t(f) - \nabla R(f)\|^2 \tag{13}$$

$$\leq \sigma^2 + (c-1)\left\|\nabla R(f)\right\|^2 \tag{14}$$

A1 ensures that this angle can be no worse than 90 degrees, where the constant D determines how correlated these gradients are in the worst-case. **A2** is employed to ensure that $R(f_0) - R(f^*)$ is finite. **A3** implies that P-L condition holds. **A4** is standard in the analysis of mirror descent, operator splitting, and proximal methods. Moreover, **A5** permits us to establish boundedness of the projected pseudo-gradient. We continue by first observing that Assumption **A5** is weaker than that of gradient boundedness, and implies $\mathbb{V}\left[\nabla r_t(f^\star)\right] \leq \sigma^2$ and $\mathbb{E}\|\nabla r_t(f)\|^2 \leq \sigma^2 + c\left\|\nabla R(f)\right\|^2$. For the sake of brevity, we also use the notation $\Gamma_t := \mathbb{E}\|\nabla R(f_t)\|_*^2$. With these conditions, we are ready to state our main result.

Theorem 1. *Under Assumptions **A1-A5**, upon running SPPPOT for $t+1$ iterations, the objective sub-optimality attenuates linearly up to a bounded neighborhood when run with constant step-size $\eta < \min(\frac{1}{q_1}, \frac{q_1}{q_2})$ and compression $\epsilon = \alpha\eta$,*

$$\mathbb{E}[R(f_{t+1}) - R(f^*)] \leq (1-\rho)^t\mathbb{E}[R(f_0) - R(f^*)] \tag{15}$$
$$+ \frac{1}{\rho}\left[L_1\eta^2\sigma^2 + \left(\frac{\eta\omega_1}{2} + L_1\eta^2\right)\alpha^2\right]$$

Further, the iterates under the same conditions satisfy:

$$\mathbb{E}[||f_{t+1} - f^*||^2] \leq \frac{2}{\lambda}(1-\rho)^t\mathbb{E}[R(f_0) - R(f^*)] \tag{16}$$
$$+ \frac{2}{\lambda\rho}\left[L_1\eta^2\sigma^2 + \left(\frac{\eta\omega_1}{2} + L_1\eta^2\right)\alpha^2\right]$$

*where $\rho = q_1\eta - q_2\eta^2$ with $q_1 = 2\lambda\left(D - \frac{1}{2\omega_1}\right)$ and $q_2 = 2\lambda c L_1$, D is the correlation constant in **A1**, and ω_1 is a positive constant satisfying $\omega_1 > \frac{1}{2D}$.*

Theorem 1 characterizes the trade-off between the rate of the convergence and the asymptotic radius of convergence. First note that regardless of the choice of η and t, the mean distance from the optimum will always be $\mathcal{O}(\alpha^2)$ in the worst case. The bound in (16) is for $\epsilon > 0$, which causes the additional α^2 to appear. For $\epsilon = 0$, the α^2 term of (16) vanishes, and thus simplifies to $\mathcal{O}(\eta\sigma^2)$ asymptotically as ρ is approximately of order η overall given $\eta < 1$.

Relative to [31][Theorem 6], our convergence result holds under comparable conditions, but incorporates the additional aspect of the parameterization efficiency/rate of convergence tradeoffs associated with sparse projections. Specifically, for $\epsilon = 0$ our result simplifies to the aforementioned result, but requires an RKHS parameterization that grows unbounded with the time index t due to (9).

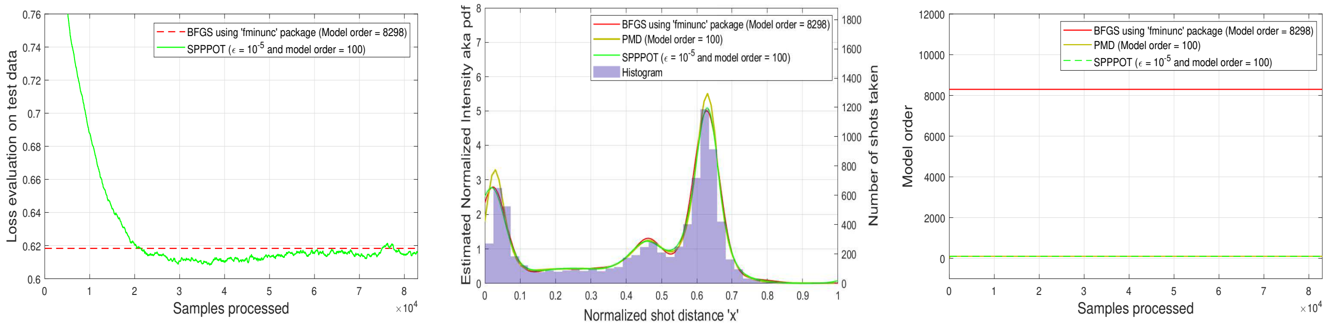
Parameterization Complexity Next we analyze the complexity of the function parameterization of SPPPOT for fixed compression $\epsilon > 0$. To do so, we need additional two assumptions.

A6. The pseudo-gradient expressed as $g_t = g_t'\kappa(\mathbf{x}_t, \cdot)$, always has scalar g_t' bounded by a positive constant: $|g_t'| \leq C$

A7. The feature space $\mathcal{X} \subset \mathbb{R}^d$ is compact.

With **A6** and **A7**, analogous logic of [28][Theorem 3] implies the following as a corollary.

[1] For future reference, we also comment that $\{z_t\} \subset \mathcal{H}_*$, i.e., z_t is an element of the dual space $\mathcal{H}_*$ of the RKHS $\mathcal{H}$.



(a) Test Loss vs. samples processed

(b) Fitted normalized Poisson Intensity

(c) Model Complexity vs. samples processed

Fig. 1: We compare a BFGS solver for an offline MLE problem [6] as well as existing incremental techniques which do not incorporate complexity-reducing projections, i.e., stochastic Pseudo-mirror descent (PMD) [31] on the NBA dataset of Stephen Curry. This data shot distances as data samples $x \in \mathbb{R}$. Shot distances less than 40 are taken and the data is normalized. SPPPOT yields a state of the art accuracy and complexity tradeoff for quickly fitting the intensity parameter of this inhomogeneous Poisson Process.

Corollary 2. *Denote as M_t the model order, or number of elements $\mathbf{x}_t$ in the dictionary associated with dual function z_t at time t. Then, we have that $M_t \leq M^\infty$, where M^∞ is the maximum model order possible. Moreover, M^∞ satisfies*

$$M^\infty \leq \mathcal{O}\left(\frac{1}{\epsilon}\right)^d \tag{17}$$

This establishes the complexity tradeoffs associated with different compression parameter selections. In the next section, we experiment with (10) on a real problem.

5. EXPERIMENTS

We now shift to validating SPPPOT for estimating the intensity function $f(\cdot)$ of an inhomogeneous Poisson Process, which we compare with an offline batch BFGS method using quasi-Newton [6] and also with existing state of the art stochastic Pseudo-mirror descent (PMD) [31] (which executes no complexity reduction/dictionary point selection). We conduct this experiment on the NBA dataset of Stephen Curry as explained in the caption of Fig. 1. The implementation of the BFGS algorithm is done using 'fminunc' package in Matlab that takes care of the algorithm step size and it has been taken as our baseline since it is a batch algorithm. The function formulation for BFGS is taken same as [6, Eq. (3.1)], i.e. $f(\cdot) = af'(\cdot)^2$ where a is a positive scalar used for tuning and $f'(\cdot)$ is a RKHS function used for learning. BFGS and SPPPOT uses the loss (3) while PMD is implemented using the loss same as [31, Eq. (8)]. Note that for the PMD algorithm, the loss contains a term with optimal intensity f^* which is unknown and hence the loss evaluation for the real life data set of Stephen Curry is not possible for PMD.

Parameter Selection We split data into 8298 training examples and 1000 test samples. BFGS is ran on a single epoch of the data, where as 10 epochs are being run for PMD and SPPPOT. For the purpose of experiments, Gaussian kernel is taken with kernel bandwidth 0.0025. The constant a for BFGS is taken to be 1 by cross validation over randomly selected 1000 points from the train data. The mini-batch size is taken to be 30 for both SPPPOT and PMD. The step size η are taken to be 0.03 and 0.1 for SPPPOT and PMD respectively by trial and error. The KOMP budget for SPPPOT is fixed at $\epsilon = 10^{-5}$ so that model order matches with the number of uniform grid points for PMD. Note that the 100 points kept in the dictionary for SPPPOT are the 100 uniform grid points only. If required one can opt for a higher model order by reducing the budget where some Poisson point will also be kept.

Results The training loss, the estimated probability density function (pdf), i.e., intensity, and the model order are given in Figs. 1a, 1b, and 1c, respectively. Observe that the stochastic gradient belongs to the probability simplex due to gradient averaging. Hence the function learnt using PMD and SPPPOT are preserve feasibility, whereas the one obtained using BFGS has to be normalized with the number of training data points to obtain a density. Moreover, BFGS took about 5 hours to compute, as compared with online approaches: PMD and SPPPOT finished 10 epochs in about 7 minutes. This runtime difference is reflected in the model complexity difference in 1c. Moreover, SPPPOT outperforms the batch solver after a few training epochs (Fig. 1a), and yields a pdf much closer to the offline baseline BFGS compared to the previous online approach PMD that does not incorporate sparse projections for point selection.

6. CONCLUSION

We studied strongly convex expected risk minimization problems when the decision variable belongs to a Reproducing Kernel Hilbert Space (RKHS) and its target domain is required to be nonnegative, motivated by in intensity estimation of inhomogeneous point processes. We put forth a variant of stochastic mirror descent that employs (i) *pseudo-gradients* and (ii) projections. Compressive projections are executed via kernel orthogonal matching pursuit (KOMP), which overcomes the fact that RKHS parameterizations grows unbounded with time. We established accuracy/complexity tradeoffs between convergence in mean and bounds on the model complexity of the learned functions under standard assumptions. Experiments outperformed state of the art techniques for inhomogeneous Poisson Process intensity estimation on real data. Future directions include scaling these approaches to higher dimensions through convolutional kernels, as well as the use of event-triggers and censoring techniques for communication-efficient networking/actuation mechanisms.

7. REFERENCES

[1] A. Chakraborty, K. Rajawat, and A. Koppel, *U.S. Army Research Laboratory/India Institute of Technology Kanpur Technical Report, 2021.*, https://koppel.netlify.app/assets/papers/2021_report_abhishek_etal.pdf.

[2] G. Francis, L. Ott, and F. Ramos, "Stochastic functional gradient for motion planning in continuous occupancy maps," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 3778–3785.

[3] L. C. Drazek, "Intensity estimation for poisson processes," *The University of Leeds, School of Mathematics*, 2013.

[4] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[5] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[6] S. Flaxman, Y. W. Teh, D. Sejdinovic *et al.*, "Poisson intensity estimation with reproducing kernels," *Electronic Journal of Statistics*, vol. 11, no. 2, pp. 5081–5104, 2017.

[7] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[8] E. Dall'Anese, A. Simonetto, S. Becker, and L. Madden, "Optimization and learning with information streams: Time-varying algorithms and applications," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 71–83, 2020.

[9] K. Slavakis, G. B. Giannakis, and G. Mateos, "Modeling and optimization for big data analytics:(statistical) learning tools for our era of data deluge," *IEEE Signal Processing Magazine*, vol. 31, no. 5, pp. 18–31, 2014.

[10] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

[11] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Springer Science & Business Media, 2009, vol. 317.

[12] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Operations Research Letters*, vol. 31, no. 3, pp. 167–175, 2003.

[13] M. Anthony and P. L. Bartlett, *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

[14] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *The annals of statistics*, pp. 1171–1220, 2008.

[15] D. Yu, G. Hinton, N. Morgan, J.-T. Chien, and S. Sagayama, "Introduction to the special section on deep learning for speech and language processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 4–6, 2011.

[16] J. Mairal, "End-to-end kernel learning with supervised convolutional kernel networks," in *Advances in neural information processing systems*, 2016, pp. 1399–1407.

[17] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE transactions on signal processing*, vol. 52, no. 8, pp. 2165–2176, 2004.

[18] Y. Lei and D.-X. Zhou, "Convergence of online mirror descent," *Applied and Computational Harmonic Analysis*, vol. 48, no. 1, pp. 343–373, 2020.

[19] 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.

[20] C. K. Williams and M. Seeger, "Using the nyström method to speed up kernel machines," in *Advances in neural information processing systems*, 2001, pp. 682–688.

[21] C. Richard, J. Carlos, M. Bermudez, and P. Honeine, "Online prediction of time series data with kernels," vol. 57, no. 3, pp. 1058–1067, 2009.

[22] W. Liu, P. P. Pokharel, and J. C. Principe, "The kernel least-mean-square algorithm," vol. 56, no. 2, pp. 543–554, 2008.

[23] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, 2008, pp. 1177–1184.

[24] Y. Yang, M. Pilanci, M. J. Wainwright *et al.*, "Randomized sketches for kernels: Fast and optimal nonparametric regression," *The Annals of Statistics*, vol. 45, no. 3, pp. 991–1023, 2017.

[25] B. Dai, B. Xie, N. He, Y. Liang, A. Raj, M.-F. F. Balcan, and L. Song, "Scalable kernel methods via doubly stochastic gradients," in *Advances in Neural Information Processing Systems*, 2014, pp. 3041–3049.

[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] Z. Wang, K. Crammer, and S. Vucetic, "Breaking the curse of kernelization: Budgeted stochastic gradient descent for large-scale svm training," *Journal of Machine Learning Research*, vol. 13, no. Oct, pp. 3103–3131, 2012.

[28] A. Koppel, G. Warnell, E. Stump, and A. Ribeiro, "Parsimonious online learning with kernels via sparse projections in function space," *Journal of Machine Learning Research*, vol. 20, no. 1, pp. 83–126, 2019.

[29] P. J. Diggle, P. Moraga, B. Rowlingson, B. M. Taylor *et al.*, "Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm," *Statistical Science*, vol. 28, no. 4, pp. 542–563, 2013.

[30] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *The annals of probability*, pp. 146–158, 1975.

[31] Y. Yang, H. Wang, N. Kiyavash, and N. He, "Learning positive functions with pseudo mirror descent," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 144–14 154.

[32] B. Poljak and Y. Z. Tsypkin, "Pseudogradient adaptation and training algorithms," *Automation and Remote Control*, vol. 34, pp. 45–67, 1973.

[33] J. Mairal, P. Koniusz, Z. Harchaoui, and C. Schmid, "Convolutional kernel networks," in *Advances in neural information processing systems*, 2014, pp. 2627–2635.

[34] W. Fischer and K. Meier-Hellstern, "The markov-modulated poisson process (mmpp) cookbook," *Performance evaluation*, vol. 18, no. 2, pp. 149–171, 1993.

[35] R. Wheeden and A. Zygmund, *Measure and Integral: An Introduction to Real Analysis*. Taylor and Francis, 1977.

[36] V. Norkin and M. Keyzer, "On stochastic optimization and statistical learning in reproducing kernel hilbert spaces by support vector machines (svm)," *Informatica*, vol. 20, pp. 273–292, 2009.

[37] B. A. Frigyik, S. Srivastava, and M. R. Gupta, "Functional bregman divergence and bayesian estimation of distributions," vol. 54, no. 11, pp. 5130–5139, 2008.

[38] P. Vincent and Y. Bengio, "Kernel matching pursuit," *Machine Learning*, vol. 48, pp. 165–187, 2002.