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  ‡ Supply Chain Optimization Technologies, Amazon

Asilomar Conference on Signals, Systems, and Computers
Oct. 31th - Nov 3rd, 2021
Introduction

Focus: function fitting when range is required to be non-negative
⇒ samples sequentially revealed \( \{x_t\}_{t \in \mathbb{N}}, x_t \in X \subset \mathbb{R}^d \)
⇒ Applicable to both supervised/unsupervised learning
⇒ Focus: feasible set ⇒ RKHS ⇒ nonlinear interpolation
⇒ Mathematically: fit predictive model \( f \in \mathcal{H}_+ \subset \mathcal{H} \) (\( \mathcal{H} \) is RKHS)
⇒ Expected risk \( R(f) := \mathbb{E}[\ell(f(x))] \), \( \ell \) negative log-likelihood
⇒ Goal: Find optimal *non-negative* function in RKHS

\[
f^* = \arg\min_{f \in \mathcal{H}_+} R(f)
\]
⇒ Poisson process: \( R(f) = \mathbb{E}[-\log(f(x))] + \int_X f(x) dx \)
Technological Context

Inhomogeneous Poisson Point Process (PPP) arise in:
⇒ Networking: Queuing theory
⇒ Communication: Base station placements
⇒ Crime: Determining crime density of a location
⇒ Other instances where non-negativity is important:
   ⇒ trajectory optimization
   ⇒ probabilistic supervised learning (logistic regression)
⇒ We focus on PPP intensity estimation

1. https://packetpushers.net/average-network-delay

Abhishek Chakraborty, Ketan Rajawat and Alec Koppel
Sparse Representations of Positive Functions
Related Works

POLK\textsuperscript{4} cannot preserve function positivity.
→ Online PMD\textsuperscript{5} ⇒ learns fixed-subspace/grid approx
  → No concept of data adaptive dictionary
→ Offline BFGS\textsuperscript{6} is not time/memory efficient
→ Offline Quadratic Program solver\textsuperscript{7}

→ Points of contrast for this work:
  ⇒ learn data-driven representation ⇒ subspace projections
  ⇒ theoretically trades off memory/accuracy
  ⇒ beats state of the art offline and online solvers

Properties of **Reproducing Kernel Hilbert Space (RKHS):**

\[
(i) \; \mathcal{H} := \text{span}(\kappa(x, \cdot)); \quad \text{and} \quad (ii) \; \langle f, \kappa(x, \cdot) \rangle_{\mathcal{H}} = f(x)
\]

**Representer Theorem for RKHS:**

\[
\hat{f}_N(\cdot) = \sum_{m=1}^{N} w_m \kappa(x_m, \cdot)
\]

\[
\Rightarrow \kappa(x_m, \cdot) \text{ is the kernel}
\]

\[
\Rightarrow \text{empirical loss minimizer } \hat{f}_N = \arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{m=1}^{N} r_m(f)
\]

\[
\rightarrow \text{amounts to search over } \mathbb{R}^N
\]

\[
\Rightarrow \text{Define Gram matrix } \mathbf{K}_{\mathcal{D}\mathcal{D}} \in \mathbb{R}^{N \times N} \text{ with } \{\kappa(x_m, x_n)\}_{m,n}
\]

\[
\rightarrow \text{As } N \rightarrow \infty, |\mathcal{D}| \rightarrow \infty, \text{ known as curse of kernelization}
\]

\[
\rightarrow \text{Need memory affordable compression}
\]

\[
\Rightarrow \text{e.g. KOMP}^8 \text{ Nyström sampling}^9, \text{ random feature approx.}^10
\]

\[
\Rightarrow \text{we adopt KOMP due trade off of memory/gradient bias}
\]

---


Want to *preserve positivity* of function estimate’s range?

⇒ Mirror descent in RKHS with Bregman divergence

⇒ **Kullback-Lieber** $B_\psi(f, \tilde{f}) = \langle f, \log(f/\tilde{f}) \rangle_H$

**Functional Bregman Divergence**\(^{11}):

\[
B_\psi(f, \tilde{f}) := \psi(f) - \psi(\tilde{f}) - \langle \nabla \psi(\tilde{f}), f - \tilde{f} \rangle_H
\]

⇒ $\psi : \mathcal{H} \to \mathbb{R}$ is proper, closed, smooth, and strongly convex

⇒ Frenchel conjugate of $\psi$ is $\psi^* : \mathcal{H}^* \to \mathbb{R}$ and $\nabla \psi^* = (\nabla \psi)^{-1}$

→ $\mathcal{H}^*$ is the Frenchel dual space of $\mathcal{H}$

⇒ Define dual (auxiliary) variable $z \in \mathcal{H}^*$ as $z = \nabla \psi(f)$

→ $f(x) = \nabla \psi^*(z(x))$

→ For KL-divergence $z = \log(f)$ and $f(x) = \exp(z(x))$

→ Exponential transformation preserves positivity

---

Mirror Descent in RKHS

Optimization problem in dual/mirror space (mirror descent in $\mathcal{H}$)

$$f_{t+1} = \arg \min_{f \in \mathcal{H}} \left( \langle g_t, f \rangle_{\mathcal{H}} + \frac{1}{\eta} B_\psi(f, f_t) \right)$$

→ Via auxiliary variable/mirror map $f_{t+1}(x) = \nabla\psi^*(z_{t+1}(x))$

$$f_{t+1} = f_t \exp(-\eta g_t) \quad \text{for KL divergence}$$

→ This update is not directly implementable in parameter space
Mirror Descent in RKHS

Optimization problem in dual/mirror space (mirror descent in $\mathcal{H}$)

$$f_{t+1} = \arg\min_{f \in \mathcal{H}} \left( \langle g_t, f \rangle_{\mathcal{H}} + \frac{1}{\eta} B_\psi(f, f_t) \right)$$

→ Via auxiliary variable/mirror map $f_{t+1}(x) = \nabla \psi^*(z_{t+1}(x))$

$$f_{t+1} = f_t \exp(-\eta g_t) \quad \text{for KL divergence}$$

→ This update is not directly implementable in parameter space

→ Aux. var. $z_t = \nabla \psi(f_t) = \log(f_t) \in \mathcal{H}$ yields $z_{t+1} = z_t - \eta g_t$

⇒ Pseudo-grad $g_t = g'_t \kappa(x_t, \cdot)$ ⇒ growing basis $z_t = \sum_u w_u g'_u$

⇒ via samples $X_t = [x_1; \cdots; x_{t-1}]$, weights $w_t$ via RKHS
Mirror Descent in RKHS

Optimization problem in dual/mirror space (mirror descent in $H$)

$$f_{t+1} = \arg \min_{f \in H} \left( \langle g_t, f \rangle_H + \frac{1}{\eta} B_\psi(f, f_t) \right)$$

$\rightarrow$ Via auxiliary variable/mirror map $f_{t+1}(x) = \nabla \psi^*(z_{t+1}(x))$

$\rightarrow$ This update is not directly implementable in parameter space

$\rightarrow$ Aux. var. $z_t = \nabla \psi(f_t) = \log(f_t) \in H$ yields $z_{t+1} = z_t - \eta g_t$

$\Rightarrow$ Pseudo-grad $g_t = g_t(x_t, \cdot) \Rightarrow$ growing basis $z_t = \sum_u w_u g_u'$

$\Rightarrow$ via samples $X_t = [x_1; \cdots; x_{t-1}]$, weights $w_t$ via RKHS

$\rightarrow$ Employ KOMP fixed budget $\epsilon$ on $z_t \sim (X_t, w_t)$

$\Rightarrow$ defines a subspace projection in $H^*$ for $z_t$
Dictionary Compression via KOMP

\[ \tilde{D}_{t+1}, \tilde{w}_{t+1} \Rightarrow z_{t+1} \text{ params. w/o proj.} \]
\[ \rightarrow \{D_{t+1}, w_{t+1}\} = \text{KOMP}(\tilde{D}_{t+1}, \tilde{w}_{t+1}, \epsilon) \]
\[ \Rightarrow \text{params } D_{t+1}, w_{t+1} \text{ after projection} \]
Pseudo-gradients

Stochastic grad. for PPP has integral \( \Rightarrow \) needs approximation
\( \Rightarrow \) **Pseudo-gradients** \( \Rightarrow \) direction correlated w/ true grad \( ^{12} \)

\[
\langle \nabla R(f_t), \mathbb{E}[g_t|F_t] \rangle \geq 0
\]

\( \Rightarrow \) e.g., *Stochastic grad, Kernel embeddings, Gradient sign*

\( \rightarrow \) Generic pseudo-gradient expression: \( g = g' \kappa(x, \cdot) \)

\( \Rightarrow \) Stochastic case: \( g' = \ell'(f_t(x)) = \ell'(\nabla \psi^*(z(x))) \)

\( \rightarrow \) Kernel embedding \( g_t = \langle \kappa(x_t, \cdot), \nabla R(f_t) \rangle \)

\( \Rightarrow \) smoothing to approximate integral in Poisson process

---

Sparse Representations of Positive Functions via Projected Pseudo-Mirror Descent

**Require:** kernel $\kappa$, step-size $\eta$, compression parameter $\epsilon$

**Initialize** Arbitrary small $z_0$

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

1. **Read:** data $x_t$
2. **Evaluate:** Pseudo Gradient $g_t = g_t' \kappa(x_t, \cdot)$
3. **Update:** $\tilde{z}_{t+1} = z_t - \eta g_t$
   - **Update Dictionary:** $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{x_t\}$
   - **Update weights:** $[w_{t+1}]_n = \begin{cases} [w_t]_n & x_n \in \mathcal{D}_t \\ -\eta g_t' & x_n = x_t \end{cases}$
4. **Compress:** $\{\mathcal{D}_{t+1}, w_{t+1}\} = \text{KOMP}(\tilde{D}_{t+1}, \tilde{w}_{t+1}, \epsilon)$
5. **Broadcast:** $z_{t+1}$

end for

**Evaluation of actual function** $f_{t+1}(x) = \nabla \psi^*(w_{t+1}^T k_{\mathcal{D}_{t+1}}(x))$
Technical Conditions

Assumption 1 \( g_t \) satisfies pseudo-gradient inequality:

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

and its expectation bounded below by 2nd-moment of dual norm:

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

Assumption 2 The optimizer of \( R(f) \) is finite and satisfies the Polyak-Łojasiewicz (P-Ł) condition

\[
\|\nabla R(f)\|^2_* \geq 2\lambda[R(f) - R(f^*)],
\]

Assumption 3 The function \( R_\psi(\cdot) \) which takes as inputs the dual functions \( z = \nabla \psi(f) \) is \( L_1 \)-smooth.

Assumption 4 Pseudo-gradient \( g_t \) satisfies variance growth condition

\[
\mathbb{E}[\|g_t\|^2_*] \leq b^2 + c^2 \mathbb{E}[\langle \nabla R(f_t), \mathbb{E}[g_t | \mathcal{F}_t] \rangle],
\]
Theorem
For constant step-size $\eta < \min\left(\frac{1}{q_1}, \frac{q_1}{q_2}\right)$ and compression $\epsilon = \alpha \eta$, the risk sub-optimality attenuates linearly up to a bounded neighborhood

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

where $\rho = q_1 \eta - q_2 \eta^2$, with constants $q_1 = 2\lambda \left(D - \frac{1}{2\omega_1}\right)$ and $q_2 = 2\lambda DL_1 c^2$. 
Assumption 5 Pseudo-gradient admits the form $g_t = g'_t \kappa(x_t, \cdot)$ with $|g'_t| \leq C$.

Assumption 6 The feature space $\mathcal{X}$ is compact.

Corollary

Denote as $M_t$ the model order, or number of elements $x_t$ in the dictionary associated with dual function $z_t$ at time $t$. Then, we have that $M_t \leq M^\infty$, where $M^\infty$ is the maximum model order possible. Moreover, $M^\infty$ satisfies

$$M^\infty \leq O \left( \frac{1}{\epsilon} \right)^d$$
Experimental setup

Poisson process intensity of NBA dataset of Stephen Curry

→ Contains shot distances from basket as data $x \in \mathbb{R}$
→ Compared SPPPOT with offline BFGS\textsuperscript{13} and online PMD\textsuperscript{14}

Performance merits:

⇒ Test Loss between SPPPOT and BFGS
   → PMD loss cannot be calculated for real world data
⇒ Learnt ”normalized intensity” aka pdf for all
⇒ Computational time and complexity


Simulation results

Loss evaluation on test data

- Offline BFGS (Model order = 8298)
- SPPPOT ($\epsilon = 10^{-5}$ and model order = 100)

Model order

- Offline BFGS (Model order = 8298)
- PMD (Model order = 100)
- SPPPOT ($\epsilon = 10^{-5}$ and model order = 100)

Estimated Normalized Intensity aka pdf

- Offline BFGS (Model order = 8298)
- PMD (Model order = 100)
- SPPPOT ($\epsilon = 10^{-5}$ and model order = 100)
- Histogram with 41 bins
Conclusion

SPPPOT beats the state of the art
⇒ Offline BFGS has high computational time/complexity
⇒ PMD employs fixed grid points, cannot extrapolate
⇒ SPPPOT has comparable complexity as PMD
   → superior performance
⇒ SPPPOT ⇒ guarantees w/ compressed dictionary
⇒ PMD does not characterize error of fixed subspace approx.
→ additional experiments, Quasi-Newton variant in the journal