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

$\Rightarrow$ samples sequentially revealed $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$, $\mathbf{x}_t \in \mathcal{X} \subset \mathbb{R}^d$

$\Rightarrow$ Applicable to both supervised/unsupervised learning

$\Rightarrow$ Focus: feasible set $\Rightarrow$ RKHS $\Rightarrow$ nonlinear interpolation

$\rightarrow$ Mathematically: fit predictive model $f \in \mathcal{H}_+ \subset \mathcal{H}$ ($\mathcal{H}$ is RKHS)

$\Rightarrow$ Expected risk $R(f) := \mathbb{E}[\ell(f(\mathbf{x}))]$, ℓ negative log-likelihood

$\rightarrow$ **Goal:** Find optimal *non-negative* function in RKHS

$$f^\star = \operatorname*{argmin}_{f\in\mathcal{H}_+} R(f)$$

$\Rightarrow$ Poisson process: $R(f) = \mathbb{E}\left[-\log(f(\mathbf{x}))\right] + \int_{\mathcal{X}} f(\mathbf{x})d\mathbf{x}$

Technological Context

Inhomogeneous Poisson Point Process (PPP) arise in:

$\Rightarrow$ Networking: Queuing theory

$\Rightarrow$ Communication: Base station placements

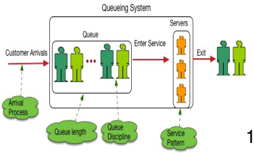
$\Rightarrow$ Crime: Determining crime density of a location

$\rightarrow$ Other instances where non-negativity is important:

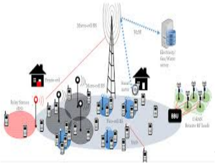
$\Rightarrow$ trajectory optimization

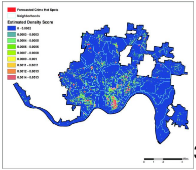
$\Rightarrow$ probabilistic supervised learning (logistic regression)

$\rightarrow$ We focus on PPP intensity estimation



1


2


3

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

[2] Azar Taufique, Mona Jaber, Ali Imran, Zaher Dawy, and Elias Yacoub, “Planning wireless cellular networks of future: Outlook, challenges and opportunities,” IEEE Access 5, pp. 4821-4845, 2017.

[3] Y. Lee, O. SooHyun and J.E. Eck, “A Theory-driven algorithm for real-time crime hot spot forecasting,” Police Quarterly, 23(2), pp.174-201, 2020.

Related Works

POLK [4] cannot preserve function positivity.

$\rightarrow$ Online PMD [5] $\Rightarrow$ learns fixed-subspace/grid approx

$\rightarrow$ No concept of data adaptive dictionary

$\rightarrow$ Offline BFGS[6] is not time/memory efficient

$\rightarrow$ Offline Quadratic Program solver [7]

$\rightarrow$ Points of contrast for this work:

$\Rightarrow$ learn data-driven representation $\Rightarrow$ subspace projections

$\Rightarrow$ theoretically trades off memory/accuracy

$\Rightarrow$ beats state of the art offline and online solvers

---

[4] 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

[5] 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.

[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] U. Marteau-Ferey, F. Bach, and A. Rudi, “Non-parametric models for non-negative functions,” in Neural Information Processing Systems, 2020.

RKHS Representations

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

$\Rightarrow$ (i) $\mathcal{H} := \overline{\text{span}(\kappa(\mathbf{x}, \cdot))}$; and (ii) $\langle f, \kappa(\mathbf{x}, \cdot)\rangle_{\mathcal{H}} = f(\mathbf{x})$

Representer Theorem for RKHS: $\hat{f}_N(\cdot) = \sum_{m=1}^{N} w_m \kappa(\mathbf{x}_m, \cdot)$

$\Rightarrow$ $\kappa(\mathbf{x}_m, \cdot)$ is the kernel

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

$\rightarrow$ amounts to search over $\mathbb{R}^N$

$\Rightarrow$ Define Gram matrix $\mathbf{K}_{\mathcal{D}\mathcal{D}} \in \mathbb{R}^{N \times N}$ with $\{\kappa(\mathbf{x}_m, \mathbf{x}_n)\}_{m,n}$

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

$\rightarrow$ Need memory affordable compression

$\Rightarrow$ e.g. KOMP[8] Nyström sampling[9], random feature approx. [10]

$\Rightarrow$ we adopt KOMP due trade off of memory/gradient bias

[8] 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

[9] Williams, C., & Seeger, M. (2001). Using the Nyström method to speed up kernel machines. In Proceedings of the 14th annual conference on neural information processing systems (No. CONF, pp. 682-688).

[10] Rahimi, A., & Recht, B. (2007, December). Random Features for Large-Scale Kernel Machines. In NIPS (Vol. 3, No. 4, p. 5).

Dai, B., Xie, B., He, N., Liang, Y., Raj, A., Balcan, M. F. F., & Song, L. (2014). Scalable Kernel Methods via Doubly Stochastic Gradients. Advances in Neural Information Processing Systems, 27, 3041-3049.

Non-Negativity $\Rightarrow$ Mirror Descent

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

$\Rightarrow$ Mirror descent in RKHS with Bregman divergence

$\Rightarrow$ **Kullback-Lieber** $B_\psi(f, \tilde{f}) = \langle f, \log(f/\tilde{f}) \rangle_{\mathcal{H}}$

Functional Bregman Divergence[11]:

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

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

$\Rightarrow$ Frenchel conjugate of ψ is $\psi^* : \mathcal{H}^* \rightarrow \mathbb{R}$ and $\nabla\psi^* = (\nabla\psi)^{-1}$
$\rightarrow$ $\mathcal{H}^*$ is the Frenchel dual space of $\mathcal{H}$

$\Rightarrow$ Define dual (auxiliary) variable $z \in \mathcal{H}^*$ as $z = \nabla\psi(f)$
$\rightarrow$ $f(\mathbf{x}) = \nabla\psi^*(z(\mathbf{x}))$
$\rightarrow$ For KL-divergence $z = \log(f)$ and $f(\mathbf{x}) = \exp(z(\mathbf{x}))$
$\rightarrow$ Exponential transformation preserves positivity

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

Mirror Descent in RKHS

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

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

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

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

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

Mirror Descent in RKHS

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

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

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

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

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

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

$\Rightarrow$ Pseudo-grad $g_t = g_t' \kappa(\mathbf{x}_t, \cdot)$ $\Rightarrow$ growing basis $z_t = \sum_u w_u g_u'$

$\Rightarrow$ via samples $\mathbf{X}_t = [\mathbf{x}_1; \cdots; \mathbf{x}_{t-1}]$, weights $\mathbf{w}_t$ via RKHS

Mirror Descent in RKHS

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

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

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

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

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

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

$\Rightarrow$ Pseudo-grad $g_t = g_t' \kappa(\mathbf{x}_t, \cdot)$ $\Rightarrow$ growing basis $z_t = \sum_u w_u g_u'$

$\Rightarrow$ via samples $\mathbf{X}_t = [\mathbf{x}_1; \cdots; \mathbf{x}_{t-1}]$, weights $\mathbf{w}_t$ via RKHS

$\rightarrow$ Employ KOMP fixed budget ϵ on $z_t \sim (\mathbf{X}_t, \mathbf{w}_t)$

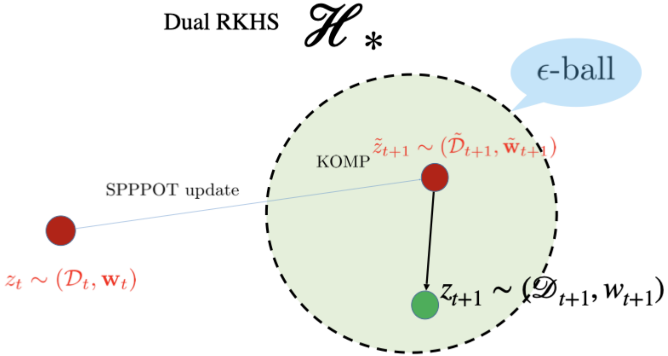
$\Rightarrow$ defines a subspace projection in $\mathcal{H}^*$ for z_t

Dictionary Compression via KOMP

$\tilde{\mathcal{D}}_{t+1}, \tilde{\mathbf{w}}_{t+1} \;\Rightarrow z_{t+1}$ params. w/o proj.

$\rightarrow \; \{\mathcal{D}_{t+1}, \mathbf{w}_{t+1}\} = \text{KOMP}(\tilde{\mathcal{D}}_{t+1}, \tilde{\mathbf{w}}_{t+1}, \epsilon)$

$\Rightarrow$ params $\mathcal{D}_{t+1}$, w_{t+1} after projection



Pseudo-gradients

Stochastic grad. for PPP has integral $\Rightarrow$ needs approximation

$\Rightarrow$ **Pseudo-gradients** $\Rightarrow$ direction correlated w/ true grad [12]

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

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

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

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

$\rightarrow$ Kernel embedding $g_t = \langle \kappa(\mathbf{x}_t, \cdot), \nabla R(f_t) \rangle$

$\Rightarrow$ smoothing to approximate integral in Poisson process

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

Sparse **R**epresentations of **P**ositive Functions via **P**rojected **P**seudo-Mirr**o**r Descent

Require: kernel κ, step-size η, compression parameter ϵ

Initialize Arbitrary small z_0

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

Read: data $\mathbf{x}_t$

Evaluate: Pseudo Gradient $g_t = g'_t \kappa(x_t, \cdot)$

Update: $\tilde{z}_{t+1} = z_t - \eta g_t$

Update Dictionary: $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{\mathbf{x}_t\}$

Update weights: $[\mathbf{w}_{t+1}]_n = \begin{cases} [\mathbf{w}_t]_n & \mathbf{x}_n \in \mathcal{D}_t \\ -\eta g'_t & \mathbf{x}_n = \mathbf{x}_t \end{cases}$

Compress: $\{\mathcal{D}_{t+1}, \mathbf{w}_{t+1}\} = \text{KOMP}(\tilde{\mathcal{D}}_{t+1}, \tilde{\mathbf{w}}_{t+1}, \epsilon)$

Broadcast: z_{t+1}

end for

Evaluation of actual function $f_{t+1}(\mathbf{x}) = \nabla\psi^*(\mathbf{w}_{t+1}^\top \mathbf{k}_{\mathcal{D}_{t+1}}(\mathbf{x}))$

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

$$\mathbb{E}[\|g_t\|_*^2] \leq b^2 + c^2\mathbb{E}[\langle \nabla R(f_t), \mathbb{E}[g_t | \mathcal{F}_t] \rangle] \ ,$$

SPPPOT Convergence

Theorem

For constant step-size $\eta < \min(\frac{1}{q_1}, \frac{q_1}{q_2})$ 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 $q_2 = 2\lambda D L_1 c^2$.

SPPPOT Complexity

Assumption 5 Pseudo-gradient admits the form $g_t = g'_t \kappa(\mathbf{x}_t, \cdot)$ 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 $\mathbf{x}_t$ 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

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

Experimental setup

Poisson process intensity of NBA dataset of Stephen Curry

$\rightarrow$ Contains shot distances from basket as data $x \in \mathbb{R}$

$\rightarrow$ Compared SPPPOT with offline BFGS[13] and online PMD[14]

Performance merits:

$\Rightarrow$ Test Loss between SPPPOT and BFGS

  $\rightarrow$ PMD loss cannot be calculated for real world data

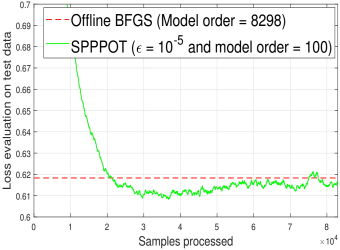
$\Rightarrow$ Learnt "normalized intensity" aka pdf for all

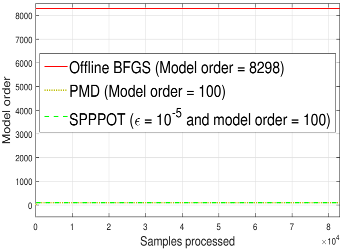
$\Rightarrow$ Computational time and complexity

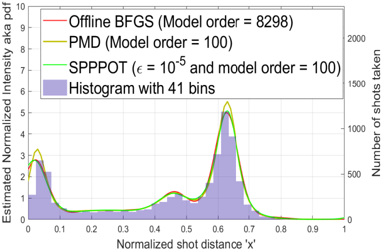
[13] S. Flaxman, Y.W. Teh, D. Sejdinovic et al., "Poisson intensity estimation with reproducing kernels," Electronic Journal of Statistics, vol. 11, no. 2, pp. 5081–5104, 2017.

[14] 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.

Simulation results







Conclusion

SPPPOT beats the state of the art

$\Rightarrow$ Offline BFGS has high computational time/complexity

$\Rightarrow$ PMD employs fixed grid points, cannot extrapolate

$\Rightarrow$ SPPPOT has comparable complexity as PMD

$\rightarrow$ superior performance

$\Rightarrow$ SPPPOT $\Rightarrow$ guarantees w/ compressed dictionary

$\Rightarrow$ PMD does not characterize error of fixed subspace approx.

$\rightarrow$ additional experiments, Quasi-Newton variant in the journal

References

⇒ A. Chakraborty, K. Rajawat, and A. Koppel, "Projected Pseudo-Mirror Descent in Reproducing Kernel Hilbert Space," in, Asilomar Conference on Signals, Systems and Computers. IEEE, 2021.

→ A. Chakraborty, K. Rajawat, and A. Koppel, "Sparse representations of positive functions via projected pseudo-mirror descent," arXiv preprint arXiv:2011.07142, 2020.