Reproducing Kernels for Pairwise Learning

Xin Guo

The Hong Kong Polytechnic University

International Conference on Computational Harmonic Analysis and Statistical Learning
Hohai University, May 19, 2019

Supported in part by Research Grants Council of Hong Kong

Outline

- Learning and Reproducing Kernels
- Centered Reproducing Kernels
- Reproducing Kernels for Pairwise Learning
- Hypothesis Spaces Generated by Pairwise Kernels

Regression and Classification Learning

- Given a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$, find a function $f_{\mathbf{z}}$ that can predict an output $f_{\mathbf{z}}(x) \in Y \subset \mathbb{R}$ for a new instance x. For example, $Y = \mathbb{R}$ for regression, and $Y = \{\pm 1\}$ for binary classification.
- Empirical risk minimization ($\lambda > 0$)

$$f_{\mathbf{z},\lambda} = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{m} \sum_{i=1}^{m} V(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2 \right\}.$$

- $V(t, y)$: the loss function, usually convex on t. For example,
  - $V(t, y) = (t - y)^2$, used in regularized least squares for regression;
  - $V(t, y) = (1 - \tau)(t - y)\mathbf{1}_{t>y} + \tau(y - t)\mathbf{1}_{y \geq t}$, $0 < \tau < 1$, the τ-pinball loss, for quantile regression;
  - $V(t, y) = \max\{0, 1 - ty\}$, the hinge loss, used in SVM for binary classification.
- $\mathcal{H}$: the hypothesis space, for example a reproducing kernel Hilbert space (RKHS).

Reproducing Kernel Hilbert Space (RKHS)

Let X be a metric space, and $K : X \times X \to \mathbb{R}$. K is called a reproducing kernel on X if it is symmetric ($K(x,u) = K(u,x)$ for any $x, u \in X$), and positive semi-definite (for any $x_1, \ldots, x_m \in X$, the Gram matrix $(K(x_i, x_j))_{1 \le i,j \le m}$ is positive semi-definite). If furthermore, K is continuous, we call it a Mercer kernel.

Examples of Mercer kernels

- linear kernel, $K(x,u) = \langle x, u \rangle$ on $\mathbb{R}^n$;
- Gaussian kernel, $K(x,u) = \exp\left(-\frac{1}{2\sigma^2}\|x-u\|^2\right)$ on $\mathbb{R}^n$, where $\sigma > 0$;
- Polynomial kernel, $K(x,u) = (1 + \langle x, u \rangle)^d$ on $\mathbb{R}^n$, where $d \ge 1$ is an integer;
- Inverse multiquadrics, $K(x,u) = \left(c^2 + \|x-u\|^2\right)^{-\alpha}$ on $\mathbb{R}^n$, for any $c, \alpha > 0$.

- For $x, u \in X$, let $K_x(u) := K(x, u)$. Define inner product $\langle \cdot, \cdot \rangle_K$ on $\mathcal{H}_* := \text{span}\{K_x : x \in X\}$ such that $\langle K_x, K_u \rangle_K = K(x, u)$ for any $x, u \in X$. Let $\|f\|_K^2 := \langle f, f \rangle_K$.
- Let $\mathcal{H}_K$ be the completion of $\mathcal{H}_*$ with respect to $\|\cdot\|_K$. $\mathcal{H}_K$ is referred to as the reproducing kernel Hilbert space (RKHS) defined by K.
- Reproducing property: $\langle f, K_x \rangle_K = f(x)$ for any $f \in \mathcal{H}_K$ and $x \in X$.
- Representer Theorem [Wahba, 1990]

$$f_{\mathbf{z},\lambda} = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{m} \sum_{i=1}^{m} V(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2 \right\}$$
$$\in \text{span}\{K_{x_i} : 1 \le i \le m\}.$$

Centered Reproducing Kernels

Let K be a Mercer kernel defined on a compact metric space X with a Borel probability measure ρ_X.

- The centered kernel with respect to ρ_X

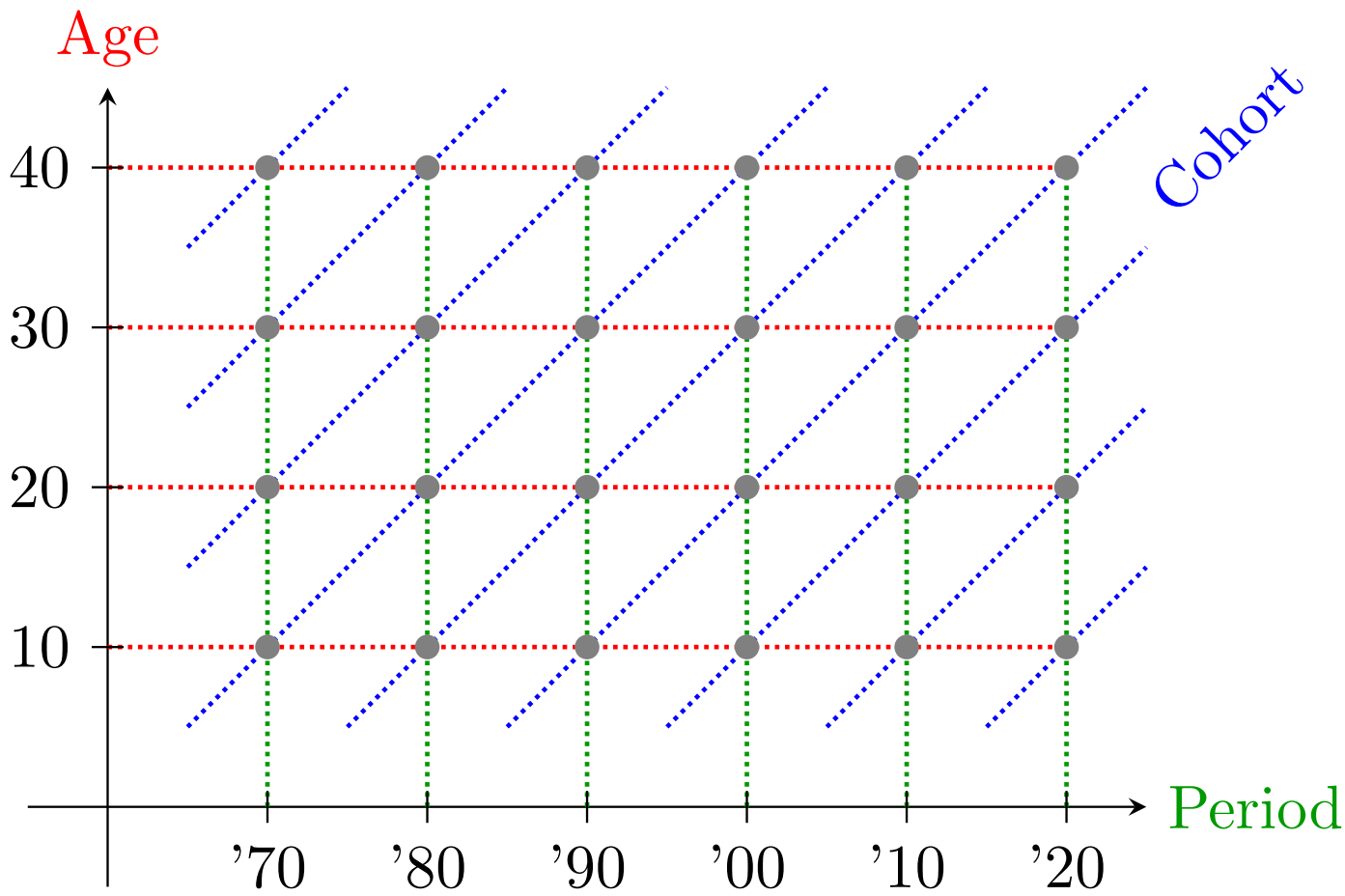
$$\bar{K}(x,u) = K(x,u) - \int_X K(x,\xi)d\rho_X(\xi) - \int_X K(\xi,u)d\rho_X(\xi) + \int_X \int_X K(\xi,\xi')d\rho_X(\xi)d\rho_X(\xi').$$

- The empirical centered kernel with respect to a sample $\{x_i\}_{i=1}^m \subset X$

$$\widehat{K}(x,u) = K(x,u) - \frac{1}{m}\sum_{i=1}^m K(x,x_i) - \frac{1}{m}\sum_{i=1}^m K(x_i,u) + \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m K(x_i,x_j).$$

- $\bar{K}$ and $\widehat{K}$ are both Mercer kernels. $\mathcal{H}_{\bar{K}} \subset \text{span}\{\mathbf{1}\}^\perp$ in $L^2_{\rho_X}$.
- $\widehat{\bar{K}} = \bar{\widehat{K}} = \widehat{K}$, and $\bar{\bar{K}} = \bar{K}$.

The Age-Period-Cohort (APC) Model



$$\text{Cohort} = \text{Period} - \text{Age}, \qquad R_{a,p} = \mu + f_a + f_p + f_c + \varepsilon_{a,p}$$

[Fu 2016; Fu, Land, Yang, 2011; Yang, Fu, Land, 2004]

- Define $L_K : L^2_\mu \to L^2_\mu$

$$(L_K f)(x) = \int_X f(u)K(u,x)d\mu(x).$$

  Define $L_{\bar{K}}$ in the same way with $\bar{K}$.
- Write $\{(\lambda_i, \phi_i)\}_i$ as the pair of eigenvalues and eigenfunctions with $\|\phi_i\|_{\mathcal{H}_K} = 1$. Write $\{(\bar{\lambda}_i, \bar{\phi}_i)\}_i$ as the eigen system for $L_{\bar{K}}$ with $\|\bar{\phi}_i\|_{\mathcal{H}_K} = 1$.

Theorem (G, Hu, Zhou, preprint)

$$\lambda_1 \geq \bar{\lambda}_1 \geq \lambda_2 \geq \bar{\lambda}_2 \geq \cdots \geq \lambda_n \geq \bar{\lambda}_n \geq \cdots .$$

- Effective dimension [Zhang, 2002] of L_K: Let $\lambda > 0$.

$$\mathcal{N}_K(\lambda) := \mathsf{Trace}(L_K(L_K + \lambda I)^{-1})$$

$\mathcal{N}_{\bar{K}}$ is similarly defined for $L_{\bar{K}}$.

Corollary (G, Hu, Zhou, preprint)

For any $\lambda > 0$,

$$\mathcal{N}_K(\lambda) \geq \mathcal{N}_{\bar{K}}(\lambda) \geq \mathcal{N}_K(\lambda) - \frac{\lambda_1}{\lambda_1 + \lambda}.$$

Consider a sample for regression $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$, with $x_i \in X$ and $y_i \in \mathbb{R}$. Define the following variations of regularized least squares

$$(f_{\mathbf{z},\lambda}, b_{\mathbf{z},\lambda}) = \arg \min_{f \in \mathcal{H}_K, b \in \mathbb{R}} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) + b - y_i)^2 + \lambda \|f\|_K^2 \right\}$$

$$f^*_{\mathbf{z},\lambda} = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \lambda \|f\|_K^2 \right\}$$

$$f^\dagger_{\mathbf{z},\lambda} = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i + \bar{y})^2 + \lambda \|f\|_K^2 \right\}, \ \bar{y} := \frac{1}{m} \sum_{i=1}^m y_i$$

$$(\hat{f}_{\mathbf{z},\lambda}, \hat{b}_{\mathbf{z},\lambda}) = \arg \min_{f \in \mathcal{H}_{\hat{K}}, b \in \mathbb{R}} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) + b - y_i)^2 + \lambda \|f\|_{\hat{K}}^2 \right\}$$

- $\hat{b}_{\mathbf{z},\lambda} = \bar{y}$.
- For any $x \in X$,

$$\hat{f}_{\mathbf{z},\lambda}(x) + \hat{b}_{\mathbf{z},\lambda} = f_{\mathbf{z},\lambda}(x) + b_{\mathbf{z},\lambda}$$

Theorem (Wang, Wu, G, preprint)

Assume $|y| \le M$ *almost surely,* $f_\rho = c + L_{\bar{K}}^r h$ *holds for some* $h \in L^2_{\rho_X}$, $c \in \mathbb{R}$, *and* $0 < r \le 1$. *Assume* $\mathcal{N}_K(\lambda) = O(\lambda^{-s})$ *for some* $0 < s \le 1$.

- *If* $\frac{1}{2} \le r \le 1$, *then with the choice* $\lambda = m^{-\frac{1}{2r+s}}$, *we have*

$$\mathbb{E}\left[\|\hat{f}_{D,\lambda} + \hat{b}_{D,\lambda} - f_\rho\|_\rho\right] = O(m^{-\frac{r}{2r+s}});$$

- *If* $r < \frac{1}{2}$ *and* $2r + s \ge 1$, *then with the choice* $\lambda = m^{-\frac{1}{2r+s}}$, *we have*

$$\mathbb{E}\left[\|\hat{f}_{D,\lambda} + \hat{b}_{D,\lambda} - f_\rho\|_\rho\right] = O(m^{-\frac{r}{2r+s}});$$

- *If* $r < \frac{1}{2}$ *and* $2r + s < 1$, *then with the choice* $\lambda = m^{-\frac{1}{1+s}}$, *we have*

$$\mathbb{E}\left[\|\hat{f}_{D,\lambda} + \hat{b}_{D,\lambda} - f_\rho\|_\rho\right] = O(m^{-\frac{r}{1+s}}).$$

Theorem (G, Hu, Zhou, preprint)

- *For any* $f \in L^2(X,\mu)$ *and* $\frac{1}{2} \le r \le 1$ *(or* $0 < r < \infty$ *if the constant function* $\mathbf{1}$ *is an eigenfunction of* L_K*), there is a real constant* c *and a function* $g \in L^2(X,\mu)$*, with* $\|g\|_\mu \le \|f\|_\mu$*, such that*

$$L_K^r g = L_{\bar{K}}^r f + c.$$

- *For any* $g \in L^2(X,\mu)$ *and* $0 < r \le \frac{1}{2}$ *(or* $0 < r < \infty$ *if the constant function* $\mathbf{1}$ *is an eigenfunction of* L_K*), there is a real constant* c *and a function* $f \in L^2(X,\mu)$*, with* $\|f\|_\mu \le \|g\|_\mu$*, such that*

$$L_{\bar{K}}^r f = L_K^r g + c.$$

Pairwise Learning

- Data: $\{(x_i, x_i', a_i)\}_{i=1}^m$, where $x_1, \ldots, x_m, x_1', \ldots, x_m'$ are sampled from some domain X, and $a_1, \ldots, a_m \in \mathbb{R}$ are the labels. Target: to find some function F to predict the label a for a new pair (x, x') coming in future.
- In particular, data may take the form $\{(x_i, y_i)\}_{i=1}^m$, where for each pair (x_i, x_j), the label a_{ij} is given by $a_{ij} = y_i - y_j$.
- Example: Ranking [Cao et al., 2006; Rudin, 2006; Cossock, Zhang, 2006; Clemencon et al., 2008; Freund et al., 2003; Agarwal, Niyogi, 2009; Rejchel, 2012; Jiang, Lim, Yao, Ye, 2011; ...], in particular, bipartite ranking [Freund et al., 2003; Agarwal et al., 2005; ...]. Data: input $\{x_i\}_{i=1}^m \subset X$ and labels $\{a_{i,j}\}_{i,j=1}^m \subset \mathbb{R}$. Sometimes only a part of the labels $\{a_{i,j}\}_{i,j=1}^m$ are available.
  For any given pair (x_i, x_j), $a_{i,j}$ is random, and is more likely to be positive if x_i is "better" than x_j.

- Other problems of pairwise learning: AUC maximization [Zhao et al., 2011; Cortes, Mohri, 2004; Ying et al., 2016; Rakotomamonjy, 2004], metric and similarity learning [Bellet, Habrard, 2014; Cao et al., 2016; Chechik et al., 2010; Weinberger, Saul, 2009; Ying, Li, 2012], and a minimum error entropy principle [Hu et al., 2015].

Loss Functions and Hypothesis Spaces for Pairwise Learning

- Loss function $l(F, x, x', a) = \varphi(aF(x, x'))$
- Examples
  - misranking loss $\varphi(t) = \chi_{\{t<0\}}$
  - least squares loss $\varphi(t) = (1-t)^2$
  - logistic loss $\varphi(t) = \log(1+e^{-t})$
  - hinge loss $\varphi(t) = \max\{0, 1-t\}$
- Risk functional

$$R(F) = \int l(F, x, x', a) d\rho(x, x', a).$$

- Hypothesis space defined by Mercer kernels
  Let $\widetilde{K} : X^2 \times X^2 \to \mathbb{R}$ be a Mercer kernel, and $\mathcal{H}_{\widetilde{K}}$ be the reproducing kernel Hilbert space defined by $\widetilde{K}$

Empirical Risk Minimization (ERM)

- Consider the pairwise learning problem with data $\{(x_i, y_i)\}_{i=1}^m \overset{\text{iid}}{\sim} (X \times \mathbb{R}, \rho)$. Define the risk

$$R(F) = \iint \varphi((y - y')F(x, x'))d\rho(x, y)d\rho(x', y').$$

- The following ERM scheme has widely been studied [Agarwal; Niyogi, 2009; Rejchel, 2012; Bellet, Habrard, 2014; Cao et al., 2016; Christmann, Zhou, 2016]

$$F_{\mathbf{z},\lambda} = \arg\min_{F \in \mathcal{H}_{\widetilde{K}}} \left\{ \frac{1}{\binom{m}{2}} \sum_{1 \le i < j \le m} \varphi((y_i - y_j)F(x_i, x_j)) + \lambda \|F\|_{\widetilde{K}}^2 \right\}$$
$$\in \text{span}\{K((x_i, x_j), \cdot) : 1 \le i, j \le m\}$$

The hypothesis spaces

- For any $F \in \mathcal{H}_{\widetilde{K}}$, we hope that F is skew-symmetric: $F(x, x') = -F(x', x)$ (as demanded by, e.g., ranking and AUC maximization, etc.). So we consider the following kernel [Vert, Qiu, Nobel, 2007; Ying, Zhou, 2016]

$$\widetilde{K}((x, x'), (u, u')) = K(x, u) + K(x', u') - K(x, u') - K(x', u),$$

  where K itself is a Mercer kernel on X.
- The skew-symmetry of $F \in \mathcal{H}_{\widetilde{K}}$ is guaranteed by the obvious fact

$$\widetilde{K}((x, x'), (u, u')) = -\widetilde{K}((x, x'), (u', u)).$$

The coefficient complexity is reduced. For any

$$\sum_{i,j} c_{i,j}\widetilde{K}((x_i,x_j),(u,u')) \in \text{span}\{K((x_i,x_j),\cdot) : 1 \le i,j \le m\},$$

Since $\widetilde{K}((x_i,x_j),(u,u')) = -\widetilde{K}((x_j,x_i),(u,u'))$, we restrict $c_{i,j} = -c_{j,i}$ to give

$$\sum_{i,j} c_{i,j}\widetilde{K}((x_i,x_j),(u,u')) = \sum_{i} \theta_i \left[K(x_i,u) - K(x_i,u')\right],$$

where

$$\theta_i = 2\sum_{j=1}^{m} c_{i,j}, \qquad 1 \le i \le m.$$

In fact, for any $F \in \mathcal{H}_{\widetilde{K}}$ there exists some $f \in \mathcal{H}_K$ so that $F(x,u) = f(x) - f(u)$ and $\|F\|_{\widetilde{K}} = \|f\|_K$. The opposite direction is not always true.

Furthermore, we have

Proposition (G, Hu, Zhou, preprint)

Let $G : X^2 \times X^2 \to \mathbb{R}$ be a reproducing kernel and $\mathcal{H}_G$ the associated RKHS, then the following two items are equivalent.

1. *For any $f \in \mathcal{H}_G$ and any $x, u, v, w \in X$, $f(x,u) = -f(u,x)$ and $f(x,u) + f(v,w) = f(x,w) + f(v,u)$.*
2. *There exists a reproducing kernel $W : X \times X \to \mathbb{R}$ such that for any $x, u, v, w \in X$,*
   $G((x,u),(v,w)) = W(x,v) + W(u,w) - W(x,w) - W(u,v)$.

The pairwise kernel is linked to the centered kernel through the following equation.

$$\begin{aligned}\widetilde{K}((x,x'),(u,u')) =& K(x,u) + K(x',u') - K(x,u') - K(x',u)\\ =& \widehat{K}(x,u) + \widehat{K}(x',u') - \widehat{K}(x,u') - \widehat{K}(x',u)\\ =& \bar{K}(x,u) + \bar{K}(x',u') - \bar{K}(x,u') - \bar{K}(x',u)\end{aligned}$$

OPERA (Ying, Zhou, 2015, 2016)

$$F_{t+1} = F_t - \frac{\gamma_t}{t-1}\sum_{j=1}^{t-1}(F_t(x_t, x_j) - y_t + y_j)K((x_t, x_j), (\cdot, \cdot)).$$

Theorem (Ying, Zhou, 2016)

Assume $\widetilde{F}_\rho \in L^r_{\widetilde{K}}(L^2_{\rho_X^2})$ *with some* $r > 0$ *(here* $\widetilde{F}_\rho(x, x') := f_\rho(x) - f_\rho(x')$*). Then with probability no less than* $1 - \delta$ *(and some properly selected step sizes* γ_t*'s), one has*

$$\|F_{T+1} - \widetilde{F}_\rho\|_\rho = O(T^{-\min(\frac{r}{2r+2}, \frac{1}{6})} \log T \log \frac{8T}{\delta}).$$

In literature, the relations between $L^r_{\widetilde{K}}(L^2_{\rho_X^2})$ and $L^r_K(L^2_{\rho_X})$ is not clear. Centered reproducing kernels can be used to characterize the relations.

Theorem (G, Hu, Zhou, preprint)

Let $L_{\bar{K}}$ be the integral operator defined with $\bar{K}$ and $\{(\bar{\lambda}_i, \bar{\phi}_i)\}_i^\infty$ be the eigenpairs of $L_{\bar{K}}$ with $\bar{\lambda}_1 \geq \bar{\lambda}_2 \geq \ldots \geq 0$ and $\|\bar{\phi}_i\|_{\bar{K}} = 1$. Let $L_{\widetilde{K}}$ be the integral operator defined with $\widetilde{K}$ and $\{(\widetilde{\lambda}_i, \widetilde{\phi}_i)\}_i^\infty$ be the eigenpairs of $L_{\widetilde{K}}$ with $\widetilde{\lambda}_1 \geq \widetilde{\lambda}_2 \geq \ldots \geq 0$ and $\|\widetilde{\phi}_i\|_{\widetilde{K}} = 1$. Then for all $i \geq 1$,

$$\widetilde{\lambda}_i = 2\bar{\lambda}_i,$$

and one can use

$$\widetilde{\phi}_i(x, u) = \bar{\phi}_i(x) - \bar{\phi}_i(u).$$

Corollary (G, Hu, Zhou, preprint)

The eigenvalues of L_K and $\frac{1}{2}L_{\widetilde{K}}$ are interlaced:

$$\lambda_1 \geq \frac{1}{2}\widetilde{\lambda}_1 \geq \lambda_2 \geq \frac{1}{2}\widetilde{\lambda}_2 \geq \cdots \geq \lambda_n \geq \frac{1}{2}\widetilde{\lambda}_n \geq \lambda_{n+1} \geq \cdots .$$

Therefore

$$\mathcal{N}_K(\lambda) - \frac{\lambda_1}{\lambda_1 + \lambda} \leq \mathcal{N}_{\widetilde{K}}(2\lambda) = N_{\bar{K}}(\lambda) \leq \mathcal{N}_K(\lambda).$$

Corollary (G, Hu, Zhou, preprint)

- *For any* $1/2 \le r \le 1$ *(*$r > 0$ *if* 1 *is eigenfunction of* L_K*), and any* $F \in L^2_{\rho_X^2}$*, there exists some* f *in* $L^2_{\rho_X}$ *such that*

$$(L^r_{\widetilde{K}} F)(x,u) = (L^r_K f)(x) - (L^r_K f)(u), \quad \text{and } \|f\|_\rho \le \|F\|_{\rho^2}.$$

- *For any* $0 < r \le 1/2$ *(*$r > 0$ *if* 1 *is eigenfunction of* L_K*), and any* $f \in L^2_{\rho_X}$*, there exists some* $F \in L^2_{\rho_X^2}$ *such that*

$$(L^r_{\widetilde{K}} F)(x,u) = (L^r_K f)(x) - (L^r_K f)(u), \quad \text{and } \|F\|_{\rho^2} \le \|f\|_\rho.$$

Corollary (G, Hu, Zhou, preprint)

Assume $\widetilde{F}_\rho(x,x') := f_\rho(x) - f_\rho(x')$ *with* $f_\rho \in L_K^r(L^2_{\rho_X})$ *for some* $0 < r \leq \frac{1}{2}$. *Let* $\{F_T\}$ *be the output of OPERA. Then with probability no less than* $1-\delta$ *(and some properly selected step sizes* γ_t*'s), one has*

$$\|F_{T+1} - \widetilde{F}_\rho\|_\rho = O(T^{-\min(\frac{r}{2r+2}, \frac{1}{6})} \log T \log \frac{8T}{\delta}).$$

Thank you!