Multiple-Kernel Regression with Sparsity Constraints

Shayan Aziznejad and Michael Unser

Biomedical Imaging Group, École Polytechnique Fédérale de Lausanne

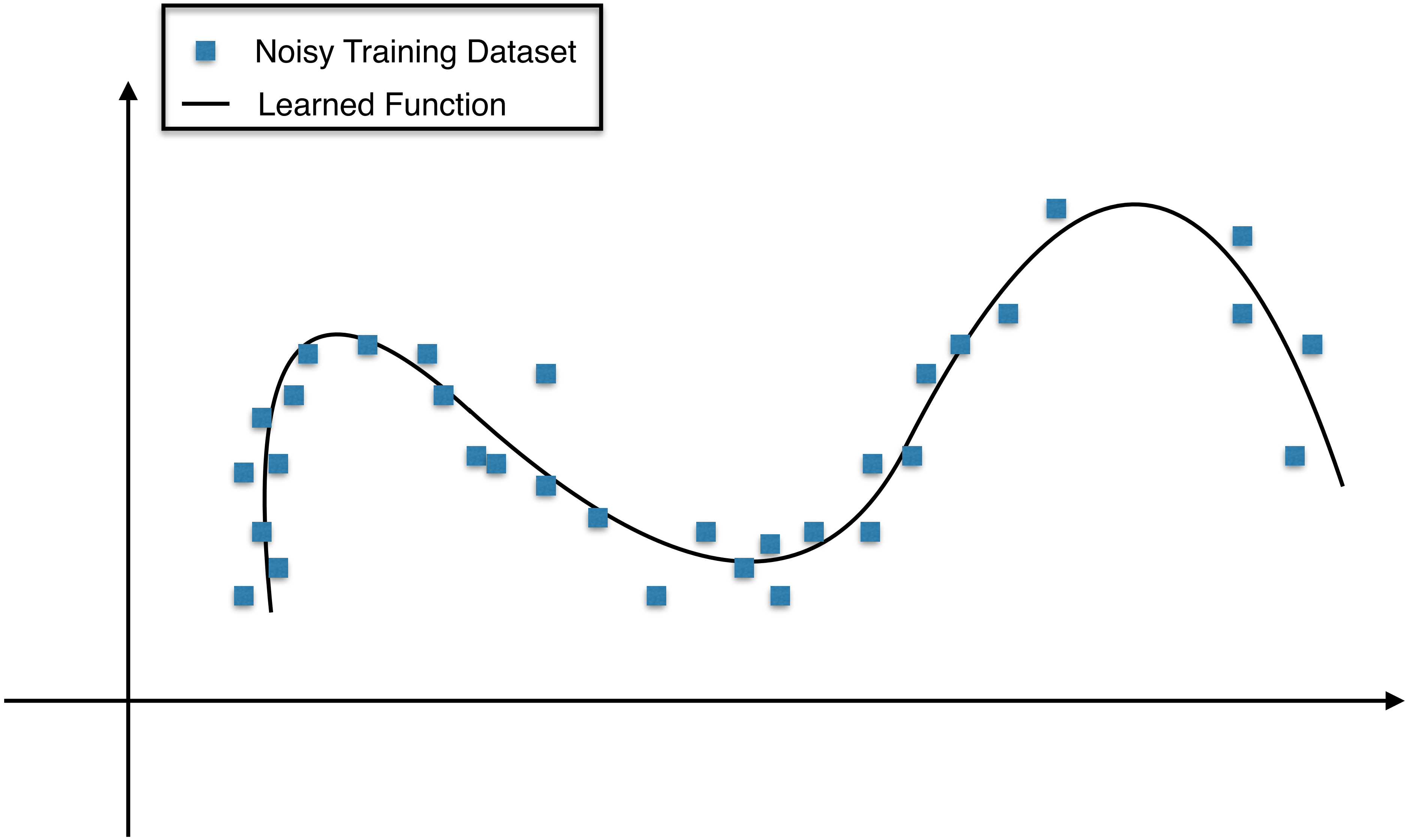
- **Supervised Learning:** Determining an unknown function given its nonuniform samples

- The goal is to recover $f:\mathbb{R}^d \to \mathbb{R}$ based on its noisy samples $y_m \approx f(\boldsymbol{x}_m)$, where $\boldsymbol{x}_m \in \mathbb{R}^d$ and $m = 1, 2, \dots, M$.

- General formulation as a minimization problem:

$$\boxed{\min_{f \in \mathcal{F}} \underbrace{\sum_{m=1}^{M} \mathrm{E}(f(\boldsymbol{x}_m), y_m)}_{\text{Data Fidelity}} + \underbrace{\lambda \mathcal{R}(f)}_{\text{Regularization}}}$$

Example



Reproducing Kernel Hilbert Spaces (RKHS)

- $\mathcal{H}(\mathbb{R}^d)$: **Hilbert** space of functions $f : \mathbb{R}^d \to \mathbb{R}$
- $\mathcal{H}(\mathbb{R}^d)$ is an RKHS, if the **sampling functional** is **continuous**.

  Equivalently, $\delta(\cdot - \boldsymbol{x}_0) \in \mathcal{H}'(\mathbb{R}^d)$

- **Unique** reproducing kernel of $\mathcal{H}(\mathbb{R}^d)$: $\mathrm{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that

$$\mathrm{k}(\boldsymbol{x}, \cdot) \in \mathcal{H}(\mathbb{R}^d), \quad \forall f \in \mathcal{H}(\mathbb{R}^d) : f(\boldsymbol{x}) = \langle \mathrm{k}(\boldsymbol{x}, \cdot), f(\cdot) \rangle_{\mathcal{H}}$$

Representer Theorem

- Supervised learning in an **RKHS**:

$$\min_{f \in \mathcal{H}(\mathbb{R}^d)} \sum_{m=1}^{M} \mathrm{E}(f(\boldsymbol{x}_m), y_m) + \lambda \|f\|_{\mathcal{H}}^2$$

- **Representer theorem** [Scholkopf et al. 2001]: If the solution exists, it is in the form of

$$\boxed{f(\boldsymbol{x}) = \sum_{m=1}^{M} a_m \mathrm{k}(\boldsymbol{x}, \boldsymbol{x}_m) \quad a_m \in \mathbb{R}, \boldsymbol{x}_m \in \mathbb{R}^d.}$$

- Linear combination of M kernels on **data points**

Application of Rep. Theorem

- Optimal solution: $f(\boldsymbol{x}) = \sum_{m=1}^{M} a_m \mathrm{k}(\boldsymbol{x}, \boldsymbol{x}_m) \quad a_m \in \mathbb{R}, \boldsymbol{x}_m \in \mathbb{R}^d.$

- Gram matrix: $G \in \mathbb{R}^{M \times M}$ such that $[G]_{m,n} = \mathrm{k}(\boldsymbol{x}_m, \boldsymbol{x}_n) \quad \Longrightarrow \quad \|f\|^2_{\mathcal{H}} = \boldsymbol{a}^T G \boldsymbol{a}$

- **Reduced** minimization problem: $\min_{\boldsymbol{a} \in \mathbb{R}^M} \sum_{m=1}^{M} \mathrm{E}([\mathrm{G}\boldsymbol{a}]_m, y_m) + \lambda \boldsymbol{a}^T \mathrm{G} \boldsymbol{a}$

- **Closed-form** solution if $\mathrm{E}(y, z) = (y - z)^2$ (Tikhonov 1963)

$$\boldsymbol{a} = \left(G^T G + \lambda G\right)^{-1} G^T \boldsymbol{y}$$

SPARSE KERNEL EXPANSION

- **Reducing complexity**: crucial for **large datasets**

- Sparsity-enforcing loss: support-vector machines [Vapnik 1998]

- **Sparsity-promoting** regularizer: **generalized LASSO** [Roth 2004]

$$\min_{\boldsymbol{a}\in\mathbb{R}^M} \sum_{m=1}^{M} \mathrm{E}([\mathrm{G}\boldsymbol{a}]_m, y_m) + \lambda\|\boldsymbol{a}\|_{\ell_1}$$

- Banach-space formulations: gTV regularization

  [Unser et al. 2017, Bach 2017]

GENERALIZED TOTAL VARIATION (gTV)

- **gTV norm** : $\|\mathrm{L}\{f\}\|_{\mathcal{M}}$, where L is an invertible LSI operator
- $\left(\mathcal{M}(\mathbb{R}^d), \|\cdot\|_{\mathcal{M}}\right)$ is a **generalization** of $(L_1(\mathbb{R}^d), \|.\|_{L_1})$:

$$f \in L_1(\mathbb{R}^d) \Rightarrow \|f\|_{\mathcal{M}} = \|f\|_{L_1}$$

$$\delta(\cdot - \boldsymbol{x}_0) \in \mathcal{M}(\mathbb{R}^d), \quad \|\delta(\cdot - \boldsymbol{x}_0)\|_{\mathcal{M}} = 1$$

- The corresponding **native space**:

$$\mathcal{M}_{\mathrm{L}}(\mathbb{R}^d) = \{f : \mathbb{R}^d \to \mathbb{R}; \|\mathrm{L}\{f\}\|_{\mathcal{M}} < +\infty\}$$

- The shift-invariant kernel associated to L :

$$\mathrm{k} = \mathrm{L}^{-1}\{\delta\}$$

Representer Theorem

- **Supervised learning** with **gTV** regularization:

$$\min_{f \in \mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)} \sum_{m=1}^{M} \mathrm{E}(f(\boldsymbol{x}_m), y_m) + \lambda \|\mathrm{L}\{f\}\|_{\mathcal{M}}$$

- **Representer theorem** [Unser et al. 2017]: Under certain assumptions, there exists an optimal solution that admits the kernel expansion

$$f(\boldsymbol{x}) = \sum_{n=1}^{N} a_n \mathrm{k}(\boldsymbol{x} - \boldsymbol{z}_n) \quad a_m \in \mathbb{R}, \boldsymbol{z}_n \in \mathbb{R}^d, N \leq M.$$

- gTV regularization $\Rightarrow$ ℓ_1 penalty on the kernel coefficients
- **Adaptive** positions $\Rightarrow$ **Sparse** representation

Application of Rep. Theorem

- Optimal solution: $f(\cdot) = \sum_{n=1}^{N} a_n \mathrm{k}(\cdot - \boldsymbol{z}_n),$

- Gram matrix: $\mathrm{G}_{\mathrm{Z}} \in \mathbb{R}^{M \times N}$ such that $[\mathrm{G}_{\mathrm{Z}}]_{m,n} = \mathrm{k}(\boldsymbol{x}_m - \boldsymbol{z}_n)$

- **Reduced** minimization problem: $\min_{\boldsymbol{a} \in \mathbb{R}^M, \mathrm{Z} \in \mathbb{R}^{d \times N}} \sum_{m=1}^{M} \mathrm{E}([\mathrm{G}_{\mathrm{Z}} \boldsymbol{a}]_m, y_m) + \lambda \|\boldsymbol{a}\|_{\ell_1}$

- Grid-based algorithms: [Gupta et al. 2017,Debarre et al. 2019]

  Suitable for low dimensional problems

Multiple Kernel Learning

- Motivation: Increasing the model flexibility to have a more accurate regression [Lanckriet et al. 2004][Bach et al. 2004].

- $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_N$ are N RKHS with the kernel functions $\mathrm{k}_1, \mathrm{k}_2, \ldots, \mathrm{k}_N$
- Learn a new positive-definite kernel $\mathrm{k}_{\boldsymbol{\mu}} = \sum_{n=1}^{N} \mu_n \mathrm{k}_n$
- Joint minimization problem:

$$\min_{\boldsymbol{a} \in \mathbb{R}^M, \boldsymbol{\mu} \in \mathrm{R}^N} \sum_{m=1}^{M} \mathrm{E}([\mathrm{G}_{\boldsymbol{\mu}} \boldsymbol{a}]_m, y_m) + \lambda \boldsymbol{a}^T \mathrm{G}_{\boldsymbol{\mu}} \boldsymbol{a}$$

Multiple Kernel Regression with gTV

- **Multi-component** model for the target function
- **Sparsity**-promoting regularization

$$\min_{f_n \in \mathcal{M}_{\mathrm{L}_n}(\mathbb{R}^d)} \sum_{m=1}^{M} \mathrm{E}(f(\boldsymbol{x}_m), y_m) + \lambda \sum_{n=1}^{N} \|\mathrm{L}_n\{f_n\}\|_{\mathcal{M}} \quad s.t. \quad f = \sum_{n=1}^{N} f_n$$

- **Representer theorem**: Under certain assumptions, there exists an optimal solution that admits the kernel expansion

$$f(\boldsymbol{x}) = \sum_{n=1}^{N} \sum_{j=1}^{M_n} a_{n,j} \mathrm{k}_n(\boldsymbol{x} - \boldsymbol{z}_{n,j}), \quad a_{n,j} \in \mathbb{R}, \boldsymbol{z}_{n,j} \in \mathbb{R}^d, \quad \sum_{n=1}^{N} M_n \leq M$$

- The number of **active kernels** is upper bounded by M

Independent of the number of components!!

ADMISSIBLE KERNELS

- Proposition: Any function $\mathrm{k} : \mathbb{R}^d \to \mathbb{R}$ with the following properties is an **admissible** kernel with respect to the gTV regularization:

    1. Non vanishing, smooth and slowly growing frequency response $\widehat{\mathrm{k}}(\boldsymbol{\omega})$

    2. Heavy-tailed frequency response; $\widehat{\mathrm{k}}(\boldsymbol{\omega}) \geq C(\|\boldsymbol{\omega}\| + 1)^{-\alpha}$

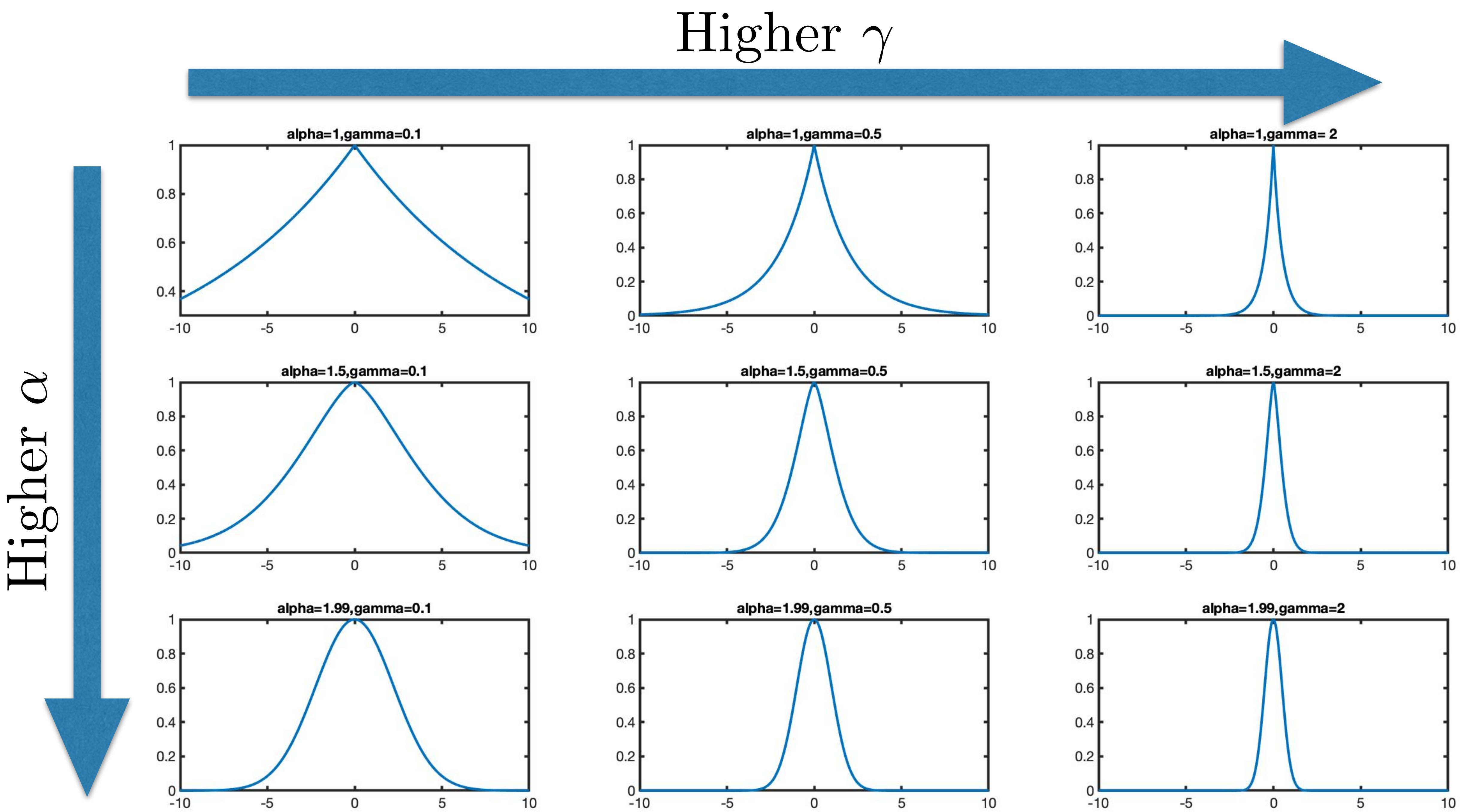
- Example: Characteristic function of heavy-tailed distributions.

- Example: Bessel potentials [Aronszajn1961]

$$G_s(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{F}^{-1}\left\{\frac{1}{(1 + \|\boldsymbol{\omega}\|_2^2)^{\frac{s}{2}}}\right\}(\boldsymbol{x} - \boldsymbol{y})$$

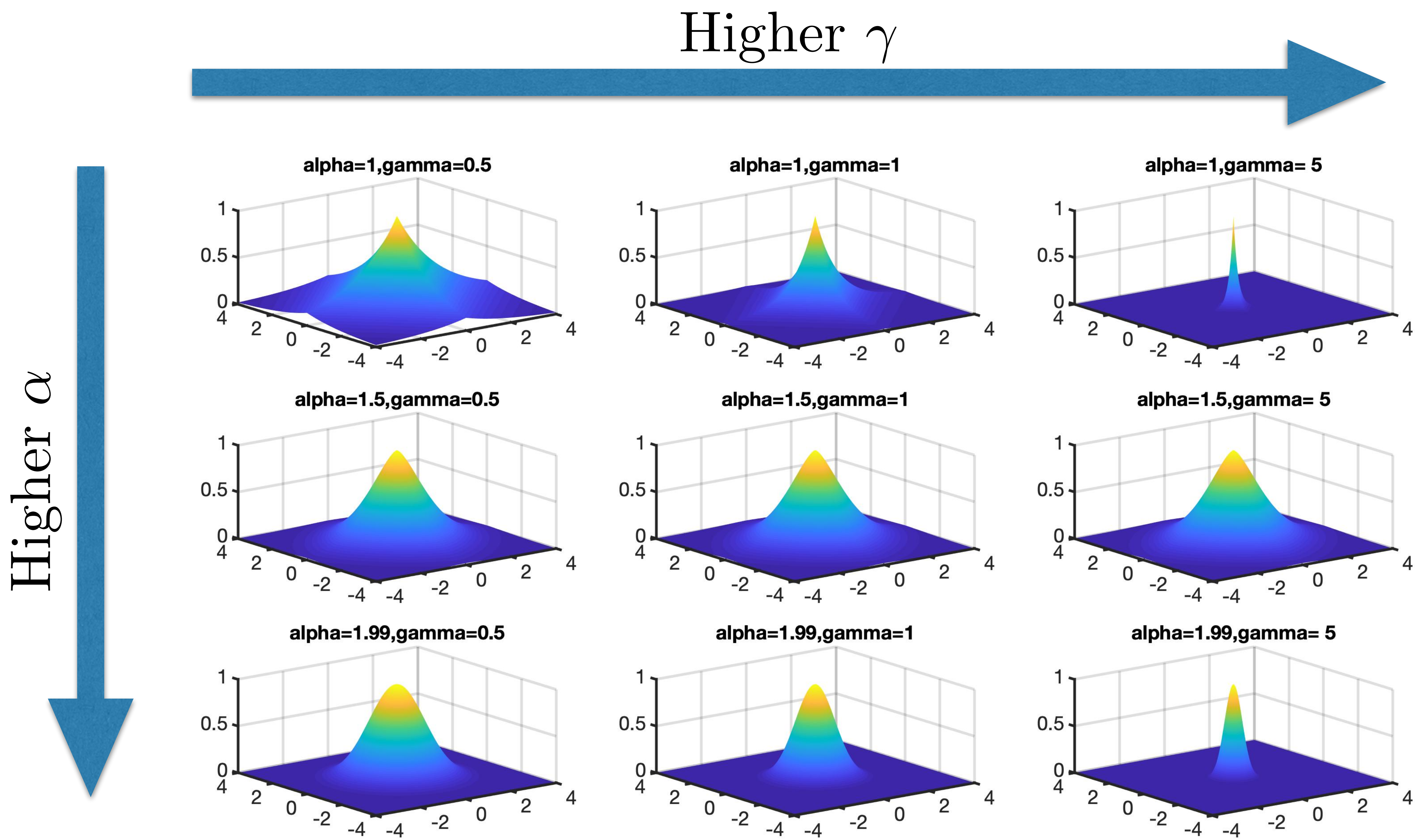
EXAMPLES: SYMMETRIC-ALPHA-STABLES

$$\mathrm{k}_{\alpha,\gamma}(\boldsymbol{x},\boldsymbol{y}) = \exp(-\gamma\|\boldsymbol{x}-\boldsymbol{y}\|_{\alpha}^{\alpha}), \quad \alpha \in (0,2), \gamma \in \mathbb{R}^{+}$$



2D Kernels

$$\mathrm{k}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{d} k(x_i - y_i).$$



Conclusion

- We considered a Banach space framework for supervised learning.
- Our framework suggests an adaptive kernel expansion for the learned function.
- We proposed a multi-component model for the target function with gTV regularization.
- Our representer theorem shows that the number of active kernels in the multi-component model is upper-bounded by the number of data points.
- We identified some classes of kernel functions that are admissible to our theory.

Challenge: Lack of algorithms for high dimensional data

References

- Schölkopf, Bernhard, Ralf Herbrich, and Alex J. Smola. "A generalized representer theorem." *International conference on computational learning theory*. Springer, Berlin, Heidelberg, 2001.
- Tikhonov, Andrei Nikolaevich. "On the solution of ill-posed problems and the method of regularization." *Doklady Akademii Nauk*. Vol. 151. No. 3. Russian Academy of Sciences, 1963.
- Vlamimir Vapnik. "Statistical learning theory Wiley." *New York* (1998): 156-160.
- Roth, Volker. "The generalized LASSO." *IEEE transactions on neural networks* 15.1 (2004): 16-28.
- Unser, Michael, Julien Fageot, and John Paul Ward. "Splines are universal solutions of linear inverse problems with generalized TV regularization." *SIAM Review* 59.4 (2017): 769-793.
- Bach, Francis. "Breaking the curse of dimensionality with convex neural networks." *The Journal of Machine Learning Research* 18.1 (2017): 629-681.
- Gupta, Harshit, Julien Fageot, and Michael Unser. "Continuous-domain solutions of linear inverse problems with Tikhonov versus generalized TV regularization." *IEEE Transactions on Signal Processing* 66.17 (2018): 4670-4684.
- Debarre, Thomas, et al. "B-Spline-Based Exact Discretization of Continuous-Domain Inverse Problems with Generalized TV Regularization." *IEEE Transactions on Information Theory* (2019).
- Lanckriet, Gert RG, et al. "Learning the kernel matrix with semidefinite programming." *Journal of Machine learning research* 5.Jan (2004): 27-72.
- Bach, Francis R., Gert RG Lanckriet, and Michael I. Jordan. "Multiple kernel learning, conic duality, and the SMO algorithm." *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004.
- Aronszajn, Nachman, and Kennan T. Smith. "Theory of Bessel potentials. I." *Annales de l'institut Fourier*. Vol. 11. 1961.

THANKS FOR YOUR ATTENTION!