Relating Leverage Scores and Density using Regularized Christoffel Functions

Edouard Pauwels[1], Francis Bach[2,4], Jean-Philippe Vert[3,4]   [1]IRIT-AOC [2]INRIA-ENS [3]Google Brain, [4]Mines ParisTech-PSL

Statistical Leverage scores

Randomized linear algebra [3]: Let $A \in \mathbb{R}^{m\times n} = UDV^T$, with $U \in \mathbb{R}^{m\times k}$, $V \in \mathbb{R}^{n\times k}$ and $D \in \mathbb{R}^{k\times k}$ diagonal positive definite, given by the SVD. The statistical leverage score of the i-th row is given by

$$\|U_{i,\cdot}\|^2 = (A(A^TA)^\dagger A^T)_{ii}$$

These can be used for sampling rows of A to approximate $\min_x \|Ax - b\|$, when $m \gg n$.

Kernel ridge regression [1]: Given a kernel matrix $K \in \mathbb{R}^{n\times n}$, observations $y \in \mathbb{R}^n$ and $\lambda > 0$, the kernel ridge regression estimate is

$$\hat{y} = K\left(K + n\lambda I\right)^{-1} y = \hat{H}y.$$

$\hat{H}_{ii}$ is a leverage score which can be used to sub-sample the observations to reduce the training cost with minor degradation of the error.

Main question

- $k\colon \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$: pd kernel, continuous, bounded, integrable.
- p: a bounded integrable density over $\mathbb{R}^d$.
- $\mathcal{H}$ is the RKHS of k (dense in $L^2(p)$), with scalar product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$.

Leverage score [5, 2]: The covariance operator $\Sigma : \mathcal{H} \to \mathcal{H}$ is then defined such that for all $f, g \in \mathcal{H}$, $\langle \Sigma f, g\rangle_{\mathcal{H}} = \int_{\mathbb{R}^d} f(x)g(x)p(x)dx$. The leverage score at $z \in \mathbb{R}^d$ is given by

$$\left\langle k(z,\cdot), (\Sigma + \lambda I)^{-1} k(z,\cdot)\right\rangle_{\mathcal{H}}$$

which is the large sample limit of statistical leverage score for kernel ridge regression. Appears in the analysis of subsampling for learning, random features, quadrature . . .

How do leverage scores relate to p and $\mathcal{H}$?

Main result

First insight: Assume that k is the Laplace kernel: $k\colon (x,y) \mapsto e^{\frac{-\|x-y\|}{l}}$ for $l > 0$. Then there is a constant $q_0(l) > 0$ such that for any $z \in \mathbb{R}^d$, with $p(z) > 0$ and p continuous at z,

$$\left\langle k(z,\cdot), (\Sigma + \lambda I)^{-1} k(z,\cdot)\right\rangle_{\mathcal{H}} \underset{\lambda\to 0}{\sim} q_0(l)\ \lambda^{-\frac{d}{d+1}}\ p(z)^{\frac{-1}{d+1}}$$

Main assumption: k is translation invariant: for any $x, y \in \mathbb{R}^d$, $k(x,y) = q(x-y)$ where $q \in L^1(\mathbb{R}^d)$ is the inverse Fourier transform of $\hat{q} \in L^1(\mathbb{R}^d)$ which is real valued and strictly positive.

Theorem 1 *Assume that for any $\omega \in \mathbb{R}^d$, $\hat{q}(\omega) = \frac{1}{(R(\omega)+Q(\omega))^\gamma}$, where R and Q are multivariate polynomials, $R \geq 1$, Q is $2s$ homogeneous and strictly positive on the unit sphere and $2s\gamma > d$.*
Then for any $z \in \mathbb{R}^d$, with $p(z) > 0$ and p continuous at z,

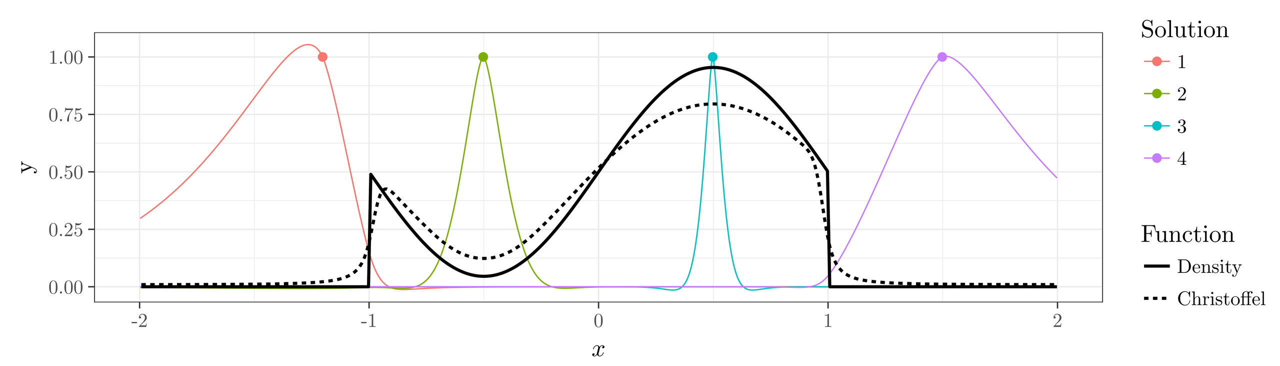
$$\left\langle k(z,\cdot), (\Sigma + \lambda I)^{-1} k(z,\cdot)\right\rangle_{\mathcal{H}} \underset{\lambda\to 0}{\sim} q_0(Q,\gamma)\ \lambda^{\frac{-d}{2s\gamma}}\ p(z)^{\frac{d}{2s\gamma}-1}$$

where $q_0(Q,\gamma) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \frac{1}{1+Q(\omega)^\gamma} d\omega.$

Divergence rate matches known estimates for degrees of freedom [5, 2]. The leverage score tends to take high values in low density regions. Laplace kernel: $R = 1$, $Q(\cdot) \sim \|\cdot\|^2$ and $\gamma = (d+1)/2$.

Other examples: Matèrn kernels, Sobolev spaces of functions with squared integrable partial derivatives up to order $s > d/2$, various norms.

Proof using regularized Christoffel function



Definition 1 *The* regularized Christoffel function, *is given for any $\lambda > 0$, $z \in \mathbb{R}^d$ by*

$$C_\lambda(z) = \inf_{f\in\mathcal{H}} \int_{\mathbb{R}^d} f(x)^2 p(x)dx + \lambda\|f\|^2_{\mathcal{H}} \quad \text{subject to} \quad f(z) = 1\,. \qquad (1)$$

Lemma 1 $C_\lambda(z) = \left\langle k(z,\cdot), (\Sigma + \lambda I)^{-1} k(z,\cdot)\right\rangle^{-1}_{\mathcal{H}}$, *for any $z \in \mathbb{R}^d$. Furthermore, replacing integration by finite sample average in (1) leads to kernel ridge regression statistical leverage score.*

Formulation (1) and connection with p are related to orthogonal polynomials [6, 4].

A simpler problem:

$$D(\lambda) := \min_{f\in\mathcal{H}} \int_{\mathbb{R}^d} f(x)^2 dx + \lambda\|f\|^2_{\mathcal{H}} \quad \text{subject to } f(\mathbf{0}) = 1. \qquad (2)$$

Lemma 2 *For any $\lambda > 0$, $D(\lambda) = \frac{(2\pi)^d}{\int_{\mathbb{R}^d}\frac{\hat{q}(\omega)}{\lambda+\hat{q}(\omega)}d\omega}$, and this value is attained by the function*

$$f_\lambda\colon x \mapsto D(\lambda)\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\frac{\hat{q}(\omega)e^{i\omega^\top x}}{\hat{q}(\omega)+\lambda}d\omega.$$

Theorem 2 *Suppose that there exists $\varepsilon\colon \mathbb{R}_+ \to \mathbb{R}_+$ such that, as $\lambda \to 0$, $\varepsilon(\lambda) \to 0$, and*

$$\int_{\|x\|\geq\varepsilon(\lambda)} f^2_\lambda(x)dx = o(\lambda D(\lambda)). \qquad (3)$$

Then, for any $z \in \mathbb{R}^d$ such that $p(z) > 0$ and p is continuous at z, we have

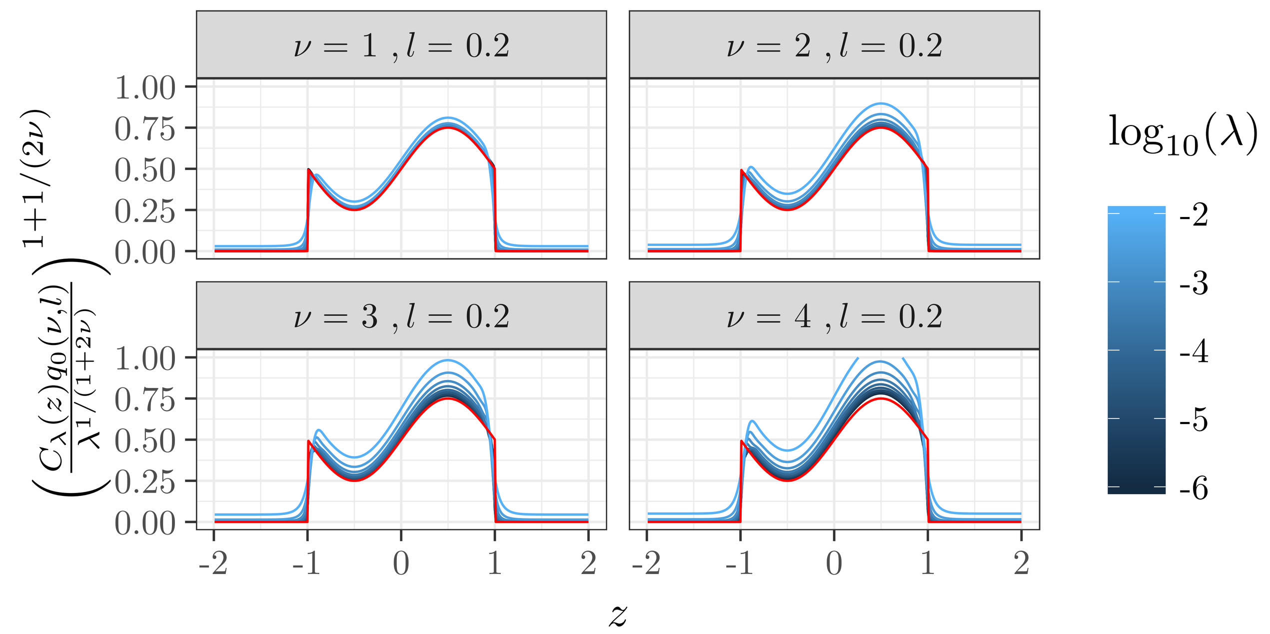
$$C_\lambda(z) \underset{\lambda\to 0,\,\lambda>0}{\sim} p(z)D\left(\frac{\lambda}{p(z)}\right).$$

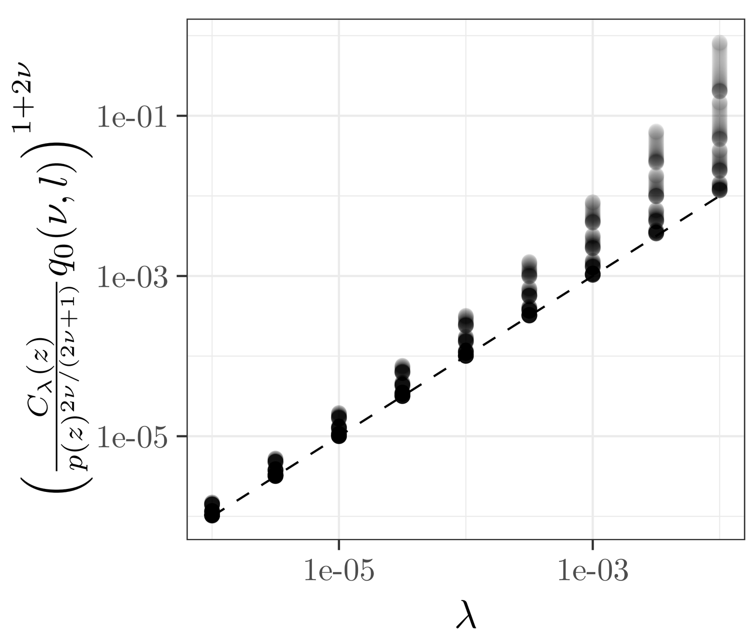
Proof sketch: Test f_λ in (1), assumption (3) leads to $C_\lambda(z) \leq p(z)D\left(\frac{\lambda}{p(z)}\right) + o\left(D\left(\frac{\lambda}{p(z)}\right)\right)$. Restricting (1) to a ball of radius $\varepsilon(\lambda)$ assumption (3) leads to $C_\lambda(z) \geq p(z)D\left(\frac{\lambda}{p(z)}\right) + o\left(D\left(\frac{\lambda}{p(z)}\right)\right)$.

Proof of Theorem 1: Check (3) and compute $D(\lambda)$ for the special choice of $\hat{q}$.

Numerical simulations

Univariate density with Matèrn kernel: bandwidth l and regularity parameter ν which corresponds to $s = 1$ and $\gamma = \nu + \frac{d}{2}$. **Left:** Comparison with the density. **Right:** Validation of the convergence rate.





Future Research

Finite sample plugin estimates and tuning of λ. Estimation of leverage scores, support, density. Broader classes of RKHS. Beyond $\mathbb{R}^d$.

Support acknowledgement

ERC-COG SEQUOIA 724063 *European Research Council*. CIMI Labex.

References

[1] Alaoui, Mahoney, NIPS 2015.
[2] Bach, JMLR 2017.
[3] Mahoney, Foundations and Trends in ML 2011.
[4] Máté, Nevai, Totik, Annals of Maths, 1991.
[5] Rudi, Camoriano, Rosasco, NIPS 2015.
[6] Szegö, AMS, 1974.