Convergence Rates of Kernel Quadrature Rules

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

NIPS workshop on probabilistic integration - Dec. 2015

Outline

- **Introduction**
  - Quadrature rules
  - Kernel quadrature rules (a.k.a. Bayes-Hermite quadrature)
- **Generic analysis of kernel quadrature rules**
  - Eigenvalues of covariance operator
  - Optimal sampling distribution
- **Extensions**
  - Link with random feature approximations
  - Full function approximation
  - Herding

Quadrature

- Given a square-integrable function $g : \mathcal{X} \to \mathbb{R}$, and a probability measure $d\rho$, approximating

$$\int_{\mathcal{X}} h(x) g(x) d\rho(x) \approx \sum_{i=1}^{n} \alpha_i h(x_i)$$

for all functions $h : \mathcal{X} \to \mathbb{R}$ in a certain function space $\mathcal{F}$

- **Many applications**

- **Main goal**:
  - Choice of support points $x_i \in \mathcal{X}$ and weights $\alpha_i \in \mathbb{R}$
  - Control of error decay as n grows, uniformly over h

Quadrature - Existing rules

- **Generic baseline: Monte-Carlo**
  - x_i sampled from $d\rho(x)$, $\alpha_i = g(x_i)/n$, with error $O(1/\sqrt{n})$

Quadrature - Existing rules

- **Generic baseline: Monte-Carlo**
  - x_i sampled from $d\rho(x)$, $\alpha_i = g(x_i)/n$, with error $O(1/\sqrt{n})$

- **One-dimensional integrals** $\mathcal{X} = [0, 1]$
  - **Trapezoidal or Simpson's rules**: $O(1/n^2)$ for f with uniformly bounded second derivatives (Cruz-Uribe and Neugebauer, 2002)
  - **Gaussian quadrature** (based on orthogonal polynomials): exact on polynomials or degree $2n - 1$ (Hildebrand, 1987)
  - **Quasi-monte carlo**: $O(1/n)$ for functions with bounded variation (Morokoff and Caflisch, 1994)

Quadrature - Existing rules

- **Generic baseline: Monte-Carlo**
  - x_i sampled from $d\rho(x)$, $\alpha_i = g(x_i)/n$, with error $O(1/\sqrt{n})$

- **One-dimensional integrals** $\mathcal{X} = [0, 1]$

- **Multi-dimensional** $\mathcal{X} = [0, 1]^d$
  - All uni-dimensional methods above generalize for small d
  - **Bayes-Hermite quadrature** (O'Hagan, 1991)
  - **Kernel quadrature** (Smola et al., 2007)

Quadrature - Existing rules

- **Generic baseline: Monte-Carlo**
  - x_i sampled from $d\rho(x)$, $\alpha_i = g(x_i)/n$, with error $O(1/\sqrt{n})$
- **One-dimensional integrals** $\mathcal{X} = [0, 1]$
- **Multi-dimensional** $\mathcal{X} = [0, 1]^d$
  - All uni-dimensional methods above generalize for small d
  - **Bayes-Hermite quadrature** (O'Hagan, 1991)
  - **Kernel quadrature** (Smola et al., 2007)
- **Extensions to less-standard sets** $\mathcal{X}$
  - Only require a positive-definite kernel on $\mathcal{X}$

Quadrature - Existing theoretical results

- **Key reference**: Novak (1988)
- **Sobolev space on** $[0,1]$ ($f^{(s)}$ square-integrable)
  - Minimax error decay: $O(n^{-s})$
- **Sobolev space on** $[0,1]^d$
  - All partial derivatives with total order $\leqslant s$ are square-integrable
  - Minimax error decay: $O(n^{-s/d})$
- **Sobolev space on the hypersphere in** $d+1$ **dimensions**
  - Minimax error decay: $O(n^{-2s/d})$

Quadrature - Existing theoretical results

- **Key reference**: Novak (1988)
- **Sobolev space on** $[0,1]$ ($f^{(s)}$ square-integrable)
  - Minimax error decay: $O(n^{-s})$
- **Sobolev space on** $[0,1]^d$
  - All partial derivatives with total order $\leqslant s$ are square-integrable
  - Minimax error decay: $O(n^{-s/d})$
- **Sobolev space on the hypersphere in** $d+1$ **dimensions**
  - Minimax error decay: $O(n^{-2s/d})$
- **A single result for all situations?**

Kernels and reproducing kernel Hilbert spaces

- **Input space** $\mathcal{X}$
- **Positive definite kernel** $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
- **Reproducing kernel Hilbert space (RKHS)** $\mathcal{F}$
  - Space of functions $f : \mathcal{X} \to \mathbb{R}$ "spanned" by $\Phi(x) = k(\cdot, x)$, $x \in \mathcal{X}$
  - Reproducing properties

$$\langle f, k(\cdot, x) \rangle_{\mathcal{F}} = f(x) \quad \text{and} \quad \langle k(\cdot, y), k(\cdot, x) \rangle_{\mathcal{F}} = k(x, y)$$

- **Example**: Sobolev spaces, e.g., $k(x, y) = \exp(-|x - y|)$

Quadrature in RKHS

- **Goal**: given a distribution $d\rho$ on $\mathcal{X}$ and $g \in L_2(d\rho)$, estimation of

$$\int_{\mathcal{X}} h(x)g(x)d\rho(x) \text{ by } \sum_{j=1}^{n} \alpha_j h(x_j) \text{ for any } h \in \mathcal{F}$$

Quadrature in RKHS

- **Goal**: given a distribution $d\rho$ on $\mathcal{X}$ and $g \in L_2(d\rho)$, estimation of

$$\int_{\mathcal{X}} h(x) g(x) d\rho(x) \text{ by } \sum_{j=1}^{n} \alpha_j h(x_j) \text{ for any } h \in \mathcal{F}$$

$$\int_{\mathcal{X}} \langle k(\cdot, x), h \rangle_{\mathcal{F}} g(x) d\rho(x) \text{ by } \sum_{j=1}^{n} \alpha_j \langle k(\cdot, x_j), h \rangle_{\mathcal{F}} \text{ for any } h \in \mathcal{F}$$

- Error $= \left\langle h, \int_{\mathcal{X}} k(x, \cdot) g(x) d\rho(x) - \sum_{j=1}^{n} \alpha_j k(x_j, \cdot) \right\rangle$

$$\leqslant \|h\|_{\mathcal{F}} \left\| \int_{\mathcal{X}} k(x, \cdot) g(x) d\rho(x) - \sum_{j=1}^{n} \alpha_j k(x_j, \cdot) \right\|_{\mathcal{F}}$$

Quadrature in RKHS

- **Goal**: given a distribution $d\rho$ on $\mathcal{X}$ and $g \in L_2(d\rho)$, estimation of

$$\int_{\mathcal{X}} h(x)g(x)d\rho(x) \text{ by } \sum_{j=1}^{n} \alpha_j h(x_j) \text{ for any } h \in \mathcal{F}$$

- **Worst-case bound**: $\displaystyle \sup_{\|h\|_{\mathcal{F}} \leqslant 1} \bigg| \int_{\mathcal{X}} h(x)g(x)d\rho(x) - \sum_{j=1}^{n} \alpha_j h(x_j) \bigg|$ is equal to (Smola et al., 2007)

$$\bigg\| \int_{\mathcal{X}} k(x,\cdot)g(x)d\rho(x) - \sum_{j=1}^{n} \alpha_j k(x_j,\cdot) \bigg\|_{\mathcal{F}}$$

Quadrature in RKHS

- **Goal**: find $x_1, \dots, x_n$ such that
$$\left\| \int_{\mathcal{X}} k(x,\cdot) g(x) d\rho(x) - \sum_{j=1}^n \alpha_j k(x_j,\cdot) \right\|_{\mathcal{F}}$$
is as small as possible

- **Computation of weights** $\alpha \in \mathbb{R}^n$ given $x_i \in \mathcal{X}$, $i = 1, \dots, n$
  – Need precise evaluations of $\mu(y) = \int_{\mathcal{X}} k(x,y) g(x) d\rho(x)$ (Smola et al., 2007; Oates and Girolami, 2015)
  – Minimize $-2 \sum_{i=1}^n \mu(x_i)\alpha_i + \sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j)$

Quadrature in RKHS

- **Goal**: find $x_1, \dots, x_n$ such that
$$\left\| \int_{\mathcal{X}} k(x,\cdot) g(x) d\rho(x) - \sum_{j=1}^n \alpha_j k(x_j,\cdot) \right\|_{\mathcal{F}}$$
is as small as possible

- **Computation of weights** $\alpha \in \mathbb{R}^n$ given $x_i \in \mathcal{X}$, $i = 1, \dots, n$
  - Need precise evaluations of $\mu(y) = \int_{\mathcal{X}} k(x,y) g(x) d\rho(x)$ (Smola et al., 2007; Oates and Girolami, 2015)
  - Minimize $-2 \sum_{i=1}^n \mu(x_i)\alpha_i + \sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j)$

- **Choice of support points** x_i
  - Optimization (Chen et al., 2010; Bach et al., 2012)
  - Sampling

Outline

- **Introduction**
  - Quadrature rules
  - Kernel quadrature rules (a.k.a. Bayes-Hermite quadrature)

- **Generic analysis of kernel quadrature rules**
  - Eigenvalues of covariance operator
  - Optimal sampling distribution

- **Extensions**
  - Link with random feature approximations
  - Full function approximation
  - Herding

Generic analysis of kernel quadrature rules

- **Support points** x_i sampled i.i.d. from a density q w.r.t. $d\rho$

- **Importance weighted** quadrature and error bound

$$\Big\| \sum_{i=1}^{n} \frac{\beta_i}{q(x_i)^{1/2}} k(\cdot, x_i) - \int_{\mathcal{X}} k(\cdot, x) g(x) d\rho(x) \Big\|_{\mathcal{F}}^2$$

- **Robustness to noise**: $\|\beta\|_2^2$ small

Generic analysis of kernel quadrature rules

- **Support points** x_i sampled i.i.d. from a density q w.r.t. $d\rho$

- **Importance weighted** quadrature and error bound

$$\left\| \sum_{i=1}^{n} \frac{\beta_i}{q(x_i)^{1/2}} k(\cdot, x_i) - \int_{\mathcal{X}} k(\cdot, x) g(x) d\rho(x) \right\|_{\mathcal{F}}^2$$

- **Robustness to noise**: $\|\beta\|_2^2$ small

- **Approximation of a function** $\mu = \int_{\mathcal{X}} k(\cdot, x) g(x) d\rho(x)$ **by random elements from an RKHS**
  - Classical tool: eigenvalues of covariance operator
  - Mercer decomposition (Mercer, 1909)

$$k(x, y) = \sum_{m \geqslant 1} \mu_m e_m(x) e_m(y)$$

Upper-bound

- **Assumptions**
  - $x_1, \ldots, x_n \in \mathcal{X}$ sampled i.i.d. with density $q(x) = \sum_{m \geqslant 1} \dfrac{\mu_m}{\mu_m + \lambda} e_m(x)^2$
  - Degrees of freedom $d(\lambda) = \sum_{m \geqslant 1} \frac{\mu_m}{\mu_m + \lambda}$

Upper-bound

- **Assumptions**
  - $x_1, \dots, x_n \in \mathcal{X}$ sampled i.i.d. with density $q(x) = \sum_{m \geqslant 1} \frac{\mu_m}{\mu_m + \lambda} e_m(x)^2$
  - Degrees of freedom $d(\lambda) = \sum_{m \geqslant 1} \frac{\mu_m}{\mu_m + \lambda}$

- **Bound**: for any $\delta > 0$, if $n \geqslant 4 + 6 d(\lambda) \log \frac{4 d(\lambda)}{\delta}$, with probability greater than $1 - \delta$, we have

$$\sup_{\|g\|_{L_2(d\rho)} \leqslant 1} \inf_{\|\beta\|_2^2 \leqslant \frac{4}{n}} \left\| \sum_{i=1}^n \frac{\beta_i}{q(x_i)^{1/2}} k(\cdot, x_i) - \int_{\mathcal{X}} k(\cdot, x) g(x) d\rho(x) \right\|_{\mathcal{F}}^2 \leqslant 4\lambda$$

Upper-bound

- **Assumptions**
  - $x_1, \dots, x_n \in \mathcal{X}$ sampled i.i.d. with density $q(x) = \sum_{m \geqslant 1} \frac{\mu_m}{\mu_m + \lambda} e_m(x)^2$
  - Degrees of freedom $d(\lambda) = \sum_{m \geqslant 1} \frac{\mu_m}{\mu_m + \lambda}$

- **Bound**: for any $\delta > 0$, if $n \geqslant 4 + 6 d(\lambda) \log \frac{4 d(\lambda)}{\delta}$, with probability greater than $1 - \delta$, we have

$$\sup_{\|g\|_{L_2(d\rho)} \leqslant 1} \inf_{\|\beta\|_2^2 \leqslant \frac{4}{n}} \left\| \sum_{i=1}^n \frac{\beta_i}{q(x_i)^{1/2}} k(\cdot, x_i) - \int_{\mathcal{X}} k(\cdot, x) g(x) d\rho(x) \right\|_{\mathcal{F}}^2 \leqslant 4\lambda$$

- **Proof technique** (Bach, 2013; El Alaoui and Mahoney, 2014)
  - Explicit β obtained by regularizing by $\lambda \|\beta\|_2^2$
  - Concentration inequalities in Hilbert space

Upper-bound

- **Assumptions**
  - $x_1, \dots, x_n \in \mathcal{X}$ sampled i.i.d. with density $q(x) = \displaystyle\sum_{m \geqslant 1} \frac{\mu_m}{\mu_m + \lambda} e_m(x)^2$
  - Degrees of freedom $d(\lambda) = \sum_{m \geqslant 1} \frac{\mu_m}{\mu_m + \lambda}$

- **Bound**: for any $\delta > 0$, if $n \geqslant 5d(\lambda) \log \dfrac{16 d(\lambda)}{\delta}$, with probability greater than $1 - \delta$, we have

$$\sup_{\|g\|_{L_2(d\rho)} \leqslant 1} \quad \inf_{\|\beta\|_2^2 \leqslant \frac{4}{n}} \left\| \sum_{i=1}^n \frac{\beta_i}{q(x_i)^{1/2}} k(\cdot, x_i) - \int_{\mathcal{X}} k(\cdot, x) g(x) d\rho(x) \right\|_{\mathcal{F}}^2 \leqslant 4\lambda$$

- **Matching lower bound** (any family of x_i)

- **Key features**: (a) degrees of freedom $d(\lambda)$ and (b) distribution q

Degrees of freedom

- Degrees of freedom
$$d(\lambda) = \sum_{m \geqslant 1} \frac{\mu_m}{\mu_m + \lambda}$$

- Traditional quantity for analysis of kernel methods (Hastie and Tibshirani, 1990)

- If eigenvalues decay as $\mu_m \approx m^{-\alpha}$, $\alpha > 1$, then
$$d(\lambda) \approx \#\{m, \mu_m \geqslant \lambda\} \approx \lambda^{-1/\alpha}$$
  - Sobolev spaces in dimension d and order s: $\alpha = 2s/d$

- **Take-home** : need $\#\{m, \mu_m \geqslant \lambda\}$ features for squared error λ

Degrees of freedom

- Degrees of freedom
$$d(\lambda) = \sum_{m \geqslant 1} \frac{\mu_m}{\mu_m + \lambda}$$

- Traditional quantity for analysis of kernel methods (Hastie and Tibshirani, 1990)

- If eigenvalues decay as $\mu_m \approx m^{-\alpha}$, $\alpha > 1$, then
$$d(\lambda) \approx \#\{m, \mu_m \geqslant \lambda\} \approx \lambda^{-1/\alpha}$$
  - Sobolev spaces in dimension d and order s: $\alpha = 2s/d$

- **Take-home** : need $\#\{m, \mu_m \geqslant \lambda\}$ features for squared error λ

- **Take-home** : After n sampled points, squared error μ_n

Optimized sampling distribution

- **Density** $q(x) = \sum_{m \geqslant 1} \frac{\mu_m}{\mu_m + \lambda} e_m(x)^2$

- **Relationship to leverage scores** (Mahoney, 2011)
  - Hard to compute in generic situations
  - Possible approximations (Drineas et al., 2012)

Optimized sampling distribution

- **Density** $q(x) = \sum_{m \geqslant 1} \frac{\mu_m}{\mu_m + \lambda} e_m(x)^2$
- **Relationship to leverage scores** (Mahoney, 2011)
  - Hard to compute in generic situations
  - Possible approximations (Drineas et al., 2012)
- **Sobolev spaces on** $[0, 1]^d$ **or hypersphere**
  - Equal to uniform distribution
  - Matches known minimax rates

Outline

- **Introduction**
  - Quadrature rules
  - Kernel quadrature rules (a.k.a. Bayes-Hermite quadrature)
- **Generic analysis of kernel quadrature rules**
  - Eigenvalues of covariance operator
  - Optimal sampling distribution
- **Extensions**
  - Link with random feature approximations
  - Full function approximation
  - Herding

Link with random feature expansions

- Some kernels are naturally expressed as **expectations**

$$k(x,y) = \mathbb{E}_v\big[\varphi(v,x)\varphi(v,y)\big] \approx \frac{1}{n}\sum_{i=1}^n \varphi(v_i,x)\varphi(v_i,y)$$

  - Neural networks with infinitely random hidden units (Neal, 1995)
  - Fourier features (Rahimi and Recht, 2007)
  - **Main question**: minimal number n of features for a given approximation quality

Link with random feature expansions

- Some kernels are naturally expressed as **expectations**

$$k(x,y) = \mathbb{E}_v\big[\varphi(v,x)\varphi(v,y)\big] \approx \frac{1}{n}\sum_{i=1}^{n}\varphi(v_i,x)\varphi(v_i,y)$$

  - Neural networks with infinitely random hidden units (Neal, 1995)
  - Fourier features (Rahimi and Recht, 2007)
  - **Main question**: minimal number n of features for a given approximation quality

- **Kernel quadrature is a subcase of random feature expansions**
  - Mercer decomposition: $k(x,y) = \sum_{m\geqslant 1}\mu_m e_m(x)e_m(y)$
  - $\varphi(v,x) = \sum_{m\geqslant 1}\mu_m^{1/2}e_m(x)e_m(v)$, with $v \in \mathcal{X}$ sampled from $d\rho$

$$\mathbb{E}_v\big[\varphi(v,x)\varphi(v,y)\big] = \sum_{m,m'\geqslant 1}(\mu_m\mu_{m'})^{1/2}e_m(x)e_{m'}(y)\big(\mathbb{E}e_m(v)e_{m'}(v)\big)$$

Full function approximation

- **Fact**: given support points x_i, quadrature rule for $\int_{\mathcal{X}} h(x)g(x)d\rho(x)$ has weights which are linear in g, that is $\alpha_i = \langle a_i, g \rangle_{L_2(d\rho)}$

$$\int_{\mathcal{X}} h(x)g(x)d\rho(x) - \sum_{i=1}^{n} \alpha_i h(x_i) = \Big\langle g, h - \sum_{i=1}^{n} h(x_i)a_i \Big\rangle_{L_2(d\rho)}$$

Full function approximation

- **Fact**: given support points x_i, quadrature rule for $\int_{\mathcal{X}} h(x)g(x)d\rho(x)$ has weights which are linear in g, that is $\alpha_i = \langle a_i, g\rangle_{L_2(d\rho)}$

$$\int_{\mathcal{X}} h(x)g(x)d\rho(x) - \sum_{i=1}^{n} \alpha_i h(x_i) = \Big\langle g, h - \sum_{i=1}^{n} h(x_i) a_i \Big\rangle_{L_2(d\rho)}$$

- **Uniform bound for** $\|g\|_{L_2(d\rho)} \leqslant 1 \Rightarrow$ **approximation of** h **in** $L_2(d\rho)$
  - Recover result from Novak (1988)
  - Approximation in RKHS norm not possible
  - Approximation in L_∞-norm incur loss of performance of $\sqrt{n}$

Full function approximation

- **Fact**: given support points x_i, quadrature rule for $\int_{\mathcal{X}} h(x)g(x)d\rho(x)$ has weights which are linear in g, that is $\alpha_i = \langle a_i, g \rangle_{L_2(d\rho)}$

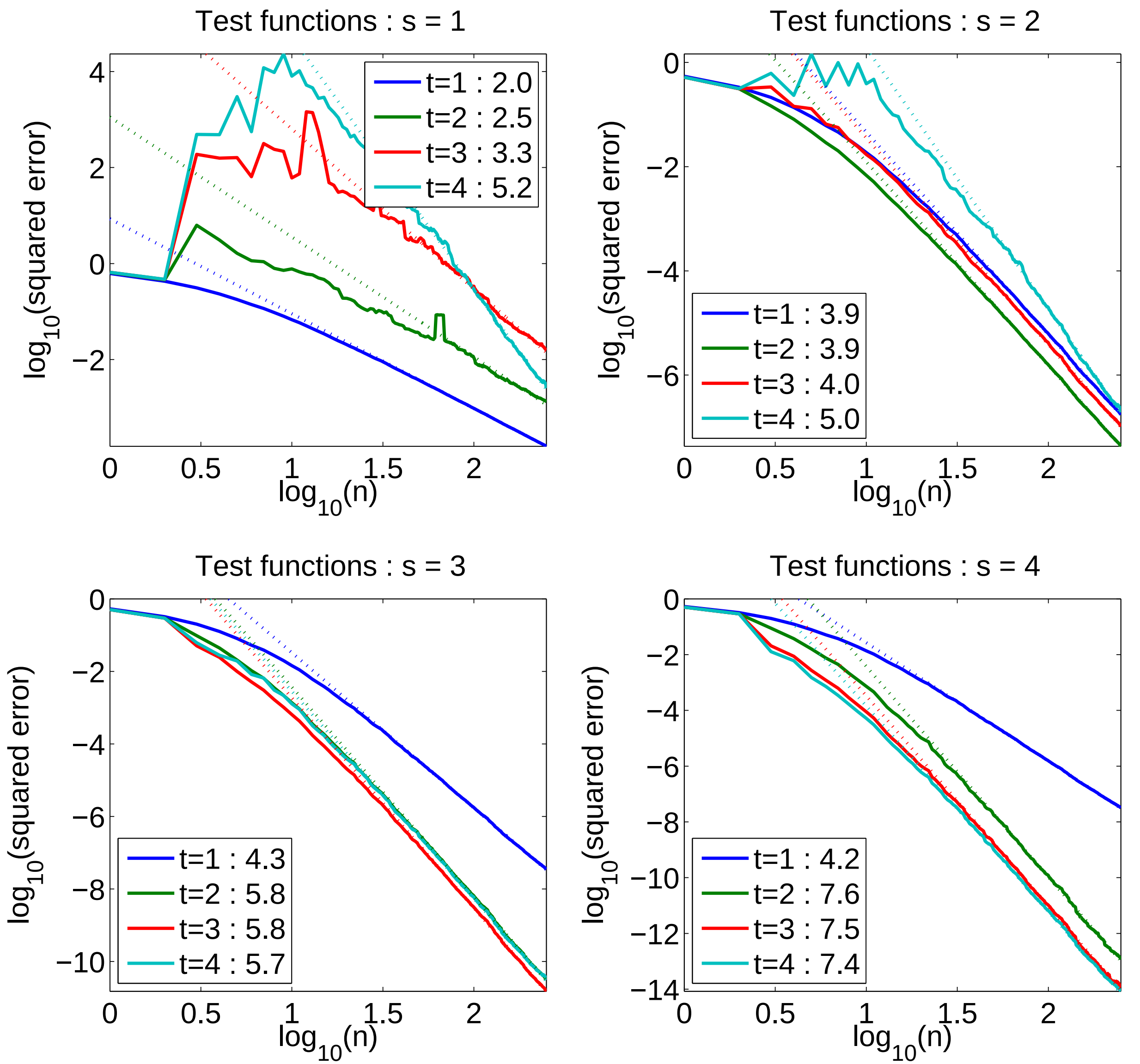
$$\int_{\mathcal{X}} h(x)g(x)d\rho(x) - \sum_{i=1}^{n} \alpha_i h(x_i) = \Big\langle g, h - \sum_{i=1}^{n} h(x_i)a_i \Big\rangle_{L_2(d\rho)}$$

- **Uniform bound for $\|g\|_{L_2(d\rho)} \leqslant 1 \Rightarrow$ approximation of h in $L_2(d\rho)$**
  - Recover result from Novak (1988)
  - Approximation in RKHS norm not possible
  - Approximation in L_∞-norm incur loss of performance of $\sqrt{n}$

- **Adaptivity to smoother functions**
  - If h is (a bit) smoother, then the rate is still optimal

Sobolev spaces on $[0, 1]$

- Quadrature rule obtained from order t, applied to order s



Herding

- **Choosing points through optimization** (Chen et al., 2010)

- **Interpretation as Frank-Wolfe optimization** (Bach, Lacoste-Julien, and Obozinski, 2012)
  - Convex weights α
  - Extra-projection step (Briol et al., 2015)

Herding

- **Choosing points through optimization** (Chen et al., 2010)

- **Interpretation as Frank-Wolfe optimization** (Bach, Lacoste-Julien, and Obozinski, 2012)
  - Convex weights α
  - Extra-projection step (Briol et al., 2015)

- **Open problem**: No "true" convergence rate (Bach et al., 2012)
  - **In finite dimension**: exponential convergence rate depending on the existence of a certain constant $c > 0$
  - **In infinite dimension**: the constant c is provably equal to zero (exponential would contradict lower bounds)

Conclusion

- **Sharp analysis of kernel quadrature rules**
  - Spectrum of the covariance operator (μ_m)
  - n points sampled i.i.d. from a well chosen distribution
  - Error of $\sqrt{\mu_n}$
  - Applies to all $\mathcal{X}$ with a positive definite kernel

Conclusion

- **Sharp analysis of kernel quadrature rules**
  - Spectrum of the covariance operator (μ_m)
  - n points sampled i.i.d. from a well chosen distribution
  - Error of $\sqrt{\mu_n}$
  - Applies to all $\mathcal{X}$ with a positive definite kernel

- **Extensions**
  - Computationally efficient ways to sample from optimized distribution (Drineas et al., 2012)
  - Anytime sampling
  - From quadrature to maximization (Novak, 1988)
  - Quasi-random sampling (Yang et al., 2014; Oates and Girolami, 2015)

References

F. Bach. Sharp analysis of low-rank kernel matrix approximations. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2013.

Francis Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms. *arXiv preprint arXiv:1203.4523*, 2012.

François-Xavier Briol, Chris J Oates, Mark Girolami, and Michael A Osborne. Frank-wolfe bayesian quadrature: Probabilistic integration with theoretical guarantees. In *Adv. NIPS*, 2015.

Y. Chen, M. Welling, and A. Smola. Super-samples from kernel herding. In *Proc. UAI*, 2010.

D. Cruz-Uribe and C. J. Neugebauer. Sharp error bounds for the trapezoidal rule and Simpson's rule. *Journal of Inequalities in Pure and Applied Mathematics*, 3(4), 2002.

P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13:3475–3506, 2012.

A. El Alaoui and M. W. Mahoney. Fast randomized kernel methods with statistical guarantees. Technical Report 1411.0306, arXiv, 2014.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1990.

F. B. Hildebrand. *Introduction to Numerical Analysis*. Courier Dover Publications, 1987.

M. W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

J. Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London, Series A.*, 209:415–446, 1909.

W. J. Morokoff and R. E. Caflisch. Quasi-random sequences and their discrepancies. *SIAM Journal on Scientific Computing*, 15(6):1251–1279, 1994.

R. M. Neal. *Bayesian Learning for Neural Networks*. PhD thesis, University of Toronto, 1995.

E. Novak. *Deterministic and Stochastic Error Bounds in Numerical Analysis*. Springer-Verlag, 1988.

C. J. Oates and M. Girolami. Variance reduction for quasi-Monte-Carlo. Technical Report 1501.03379, arXiv, 2015.

A. O'Hagan. Bayes-Hermite quadrature. *Journal of statistical planning and inference*, 29(3):245–260, 1991.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20:1177–1184, 2007.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

J. Yang, V. Sindhwani, H. Avron, and M. Mahoney. Quasi-Monte Carlo feature maps for shift-invariant kernels. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.