Information Consistency of Nonparametric Gaussian Process Methods

Matthias W. Seeger

Max Planck Institute for Biological Cybernetics
Tübingen, Germany
www.kyb.tuebingen.mpg.de/bs/people/seeger/

with Sham M. Kakade, Dean P. Foster

The Prediction Game

- Conditional sequence prediction under log loss: $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \rightarrow (y_1, \ldots, y_n)$
  - Predict $Q(\cdot|\boldsymbol{y}_{<i}, \boldsymbol{x}_{\leq i})$
  - Incur loss $-\log Q(y_i|\boldsymbol{y}_{<i}, \boldsymbol{x}_{\leq i})$

  Cumulative loss: $\sum_{i=1}^{n} -\log Q(y_i|\boldsymbol{y}_{<i})$
- Fixed $P(y|u)$, function space $\mathcal{F}$. Bayesian strategy:

$$P_{bs}(y_i|\boldsymbol{y}_{<i}) = \mathrm{E}_{P_{bs}(f|\boldsymbol{y}_{<i})}\left[P(y_i|f(\boldsymbol{x}_i))\right],$$

$$\underbrace{dP_{bs}(f|\boldsymbol{y}_{<i})}_{\text{Posterior}} \propto \prod_{j<i} P(y_j|f(\boldsymbol{x}_j)) \underbrace{dP_{bs}(f)}_{\text{Prior}}$$

  Cumulative loss: $L_{bs}(\boldsymbol{y}_{\leq n}) = -\log P_{bs}(\boldsymbol{y}_{\leq n})$
- Expert strategy: For fixed $f \in \mathcal{F}$ predict $P(y_i|f(\boldsymbol{x}_i))$.
  Cumulative loss: $L_f(\boldsymbol{y}_{\leq n}) = -\log P(\boldsymbol{y}_{\leq n}|f)$

The Prediction Game

- Conditional sequence prediction under log loss: $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \to (y_1, \ldots, y_n)$
  - Predict $Q(\cdot|\boldsymbol{y}_{<i}, \boldsymbol{x}_{\leq i})$
  - Incur loss $-\log Q(y_i|\boldsymbol{y}_{<i}, \boldsymbol{x}_{\leq i})$

  Cumulative loss: $\sum_{i=1}^{n} -\log Q(y_i|\boldsymbol{y}_{<i})$
- Fixed $P(y|u)$, function space $\mathcal{F}$. Bayesian strategy:

$$P_{bs}(y_i|\boldsymbol{y}_{<i}) = \mathrm{E}_{P_{bs}(f|\boldsymbol{y}_{<i})}\left[P(y_i|f(\boldsymbol{x}_i))\right],$$

$$\underbrace{dP_{bs}(f|\boldsymbol{y}_{<i})}_{\textit{Posterior}} \propto \prod\nolimits_{j<i} P(y_j|f(\boldsymbol{x}_j)) \underbrace{dP_{bs}(f)}_{\textit{Prior}}$$

  Cumulative loss: $L_{bs}(\boldsymbol{y}_{\leq n}) = -\log P_{bs}(\boldsymbol{y}_{\leq n})$
- Expert strategy: For fixed $f \in \mathcal{F}$ predict $P(y_i|f(\boldsymbol{x}_i))$.
  Cumulative loss: $L_f(\boldsymbol{y}_{\leq n}) = -\log P(\boldsymbol{y}_{\leq n}|f)$

The Prediction Game

- Conditional sequence prediction under log loss: $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \rightarrow (y_1, \ldots, y_n)$
  - Predict $Q(\cdot|\boldsymbol{y}_{<i}, \boldsymbol{x}_{\leq i})$
  - Incur loss $-\log Q(y_i|\boldsymbol{y}_{<i}, \boldsymbol{x}_{\leq i})$

  Cumulative loss: $\sum_{i=1}^{n} -\log Q(y_i|\boldsymbol{y}_{<i})$
- Fixed $P(y|u)$, function space $\mathcal{F}$. Bayesian strategy:

$$P_{bs}(y_i|\boldsymbol{y}_{<i}) = \mathrm{E}_{P_{bs}(f|\boldsymbol{y}_{<i})}\left[P(y_i|f(\boldsymbol{x}_i))\right],$$

$$\underbrace{dP_{bs}(f|\boldsymbol{y}_{<i})}_{\textit{Posterior}} \propto \prod\nolimits_{j<i} P(y_j|f(\boldsymbol{x}_j)) \underbrace{dP_{bs}(f)}_{\textit{Prior}}$$

  Cumulative loss: $L_{bs}(\boldsymbol{y}_{\leq n}) = -\log P_{bs}(\boldsymbol{y}_{\leq n})$
- Expert strategy: For fixed $f \in \mathcal{F}$ predict $P(y_i|f(\boldsymbol{x}_i))$.
  Cumulative loss: $L_f(\boldsymbol{y}_{\leq n}) = -\log P(\boldsymbol{y}_{\leq n}|f)$

Information Consistency

Barron, Bayes. Stats. 98

- Likelihood $P(y|u)$. Strategy Q. Competitor space $\mathcal{F}_{comp} \subset \mathcal{F}$. Input distribution $d\mu(\boldsymbol{x})$.
- Q information consistent over $\mathcal{F}_{comp}$ w.r.t. μ iff $\forall f \in \mathcal{F}_{comp}$:

$$\frac{\mathrm{E}_{\boldsymbol{x}_{\leq n} \sim \mu^n} \left[\mathrm{D} \left[P(\boldsymbol{y}_{\leq n} | f, \boldsymbol{x}_{\leq n}) \,\|\, Q(\boldsymbol{y}_{\leq n} | \boldsymbol{x}_{\leq n}) \right] \right]}{n} \to 0 \; (n \to \infty)$$

- Link to prediction game:

$$\mathrm{D}[P(\boldsymbol{y}_{\leq n}|f) \,\|\, P_{bs}(\boldsymbol{y}_{\leq n})] = \mathrm{E}[L_{bs}(\boldsymbol{y}_{\leq n}) - L_f(\boldsymbol{y}_{\leq n})]$$

Uniform (over $\boldsymbol{y}_{\leq n}$) regret bound implies information consistency

Information Consistency

Barron, Bayes. Stats. 98

- Likelihood $P(y|u)$. Strategy Q. Competitor space $\mathcal{F}_{comp} \subset \mathcal{F}$. Input distribution $d\mu(\boldsymbol{x})$.
- Q information consistent over $\mathcal{F}_{comp}$ w.r.t. μ iff $\forall f \in \mathcal{F}_{comp}$:

$$\frac{\mathrm{E}_{\boldsymbol{x}_{\leq n} \sim \mu^n} \left[\mathrm{D} \left[P(\boldsymbol{y}_{\leq n} | f, \boldsymbol{x}_{\leq n}) \,\|\, Q(\boldsymbol{y}_{\leq n} | \boldsymbol{x}_{\leq n}) \right] \right]}{n} \to 0 \; (n \to \infty)$$

- Link to prediction game:

$$\mathrm{D}[P(\boldsymbol{y}_{\leq n}|f) \,\|\, P_{bs}(\boldsymbol{y}_{\leq n})] = \mathrm{E}[L_{bs}(\boldsymbol{y}_{\leq n}) - L_f(\boldsymbol{y}_{\leq n})]$$

Uniform (over $\boldsymbol{y}_{\leq n}$) regret bound implies information consistency

Information Consistency

Barron, Bayes. Stats. 98

- Likelihood $P(y|u)$. Strategy Q. Competitor space $\mathcal{F}_{comp} \subset \mathcal{F}$. Input distribution $d\mu(\boldsymbol{x})$.
- Q information consistent over $\mathcal{F}_{comp}$ w.r.t. μ iff $\forall f \in \mathcal{F}_{comp}$:

$$\frac{\mathrm{E}_{\boldsymbol{x}_{\leq n} \sim \mu^n} \left[\mathrm{D} \left[P(\boldsymbol{y}_{\leq n} | f, \boldsymbol{x}_{\leq n}) \parallel Q(\boldsymbol{y}_{\leq n} | \boldsymbol{x}_{\leq n}) \right] \right]}{n} \to 0 \ (n \to \infty)$$

- Link to prediction game:

$$\mathrm{D}[P(\boldsymbol{y}_{\leq n}|f) \parallel P_{bs}(\boldsymbol{y}_{\leq n})] = \mathrm{E}[L_{bs}(\boldsymbol{y}_{\leq n}) - L_f(\boldsymbol{y}_{\leq n})]$$

Uniform (over $\boldsymbol{y}_{\leq n}$) regret bound implies information consistency

Information Consistency (II)

- Our result implies:
  Gaussian process prediction (dP_{bs} GP prior with kernel K)
  is information consistent over reprocuding kernel Hilbert space (RKHS) $\mathcal{H}_K$ w.r.t. any μ, for any stationary K with $K(\boldsymbol{x}, \boldsymbol{x}) < \infty$
- $\mathcal{H}_K$ is (usually) infinite dimensional (nonparametric prediction)
- Our approach allows simple computation of information consistency rate bounds from spectrum of kernel operator
- Main result

$$\underbrace{-\log P_{bs}(\boldsymbol{y}_{\leq n})}_{L_{bs}(\boldsymbol{y}_{\leq n})} \leq \underbrace{-\log P(\boldsymbol{y}_{\leq n}|f)}_{L_f(\boldsymbol{y}_{\leq n})} + \frac{1}{2}\|f\|_K^2 + \frac{1}{2}\underbrace{\log|\boldsymbol{I} + c\boldsymbol{K}|}_{\text{E}[\cdot]:\text{Regret term}}$$

for any $\boldsymbol{y}_{\leq n} \in \mathbb{R}^n$ and any $f \in \mathcal{H}_K$.
$\boldsymbol{K} = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))$ kernel matrix

c bounds $-d^2 \log P(y|u)/du^2$

Information Consistency (II)

- Our result implies:
  Gaussian process prediction (dP_{bs} GP prior with kernel K)
  is information consistent over reprocuding kernel Hilbert space (RKHS) $\mathcal{H}_K$ w.r.t. any μ, for any stationary K with $K(\boldsymbol{x}, \boldsymbol{x}) < \infty$
- $\mathcal{H}_K$ is (usually) infinite dimensional (nonparametric prediction)
- Our approach allows simple computation of information consistency rate bounds from spectrum of kernel operator
- Main result

$$\underbrace{-\log P_{bs}(\boldsymbol{y}_{\leq n})}_{L_{bs}(\boldsymbol{y}_{\leq n})} \leq \underbrace{-\log P(\boldsymbol{y}_{\leq n}|f)}_{L_f(\boldsymbol{y}_{\leq n})} + \frac{1}{2}\|f\|_K^2 + \frac{1}{2}\underbrace{\log|\boldsymbol{I} + c\boldsymbol{K}|}_{\text{E}[\cdot]:\text{Regret term}}$$

for any $\boldsymbol{y}_{\leq n} \in \mathbb{R}^n$ and any $f \in \mathcal{H}_K$.
$\boldsymbol{K} = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))$ kernel matrix

c bounds $-d^2 \log P(y|u)/du^2$

Information Consistency (II)

- Our result implies:
  Gaussian process prediction (dP_{bs} GP prior with kernel K)
  is information consistent over reprocuding kernel Hilbert space (RKHS) $\mathcal{H}_K$ w.r.t. any μ, for any stationary K with $K(\boldsymbol{x}, \boldsymbol{x}) < \infty$
- $\mathcal{H}_K$ is (usually) infinite dimensional (nonparametric prediction)
- Our approach allows simple computation of information consistency rate bounds from spectrum of kernel operator
- Main result

$$\underbrace{-\log P_{bs}(\boldsymbol{y}_{\leq n})}_{L_{bs}(\boldsymbol{y}_{\leq n})} \leq \underbrace{-\log P(\boldsymbol{y}_{\leq n}|f)}_{L_f(\boldsymbol{y}_{\leq n})} + \frac{1}{2}\|f\|_K^2 + \frac{1}{2}\underbrace{\log|\boldsymbol{I} + c\boldsymbol{K}|}_{\text{E}[\cdot]:\text{Regret term}}$$

for any $\boldsymbol{y}_{\leq n} \in \mathbb{R}^n$ and any $f \in \mathcal{H}_K$.
$\boldsymbol{K} = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))$ kernel matrix

c bounds $-d^2 \log P(y|u)/du^2$

Information Consistency (II)

- Our result implies:
  Gaussian process prediction (dP_{bs} GP prior with kernel K)
  is information consistent over reprocuding kernel Hilbert space (RKHS) $\mathcal{H}_K$ w.r.t. any μ, for any stationary K with $K(\boldsymbol{x}, \boldsymbol{x}) < \infty$
- $\mathcal{H}_K$ is (usually) infinite dimensional (nonparametric prediction)
- Our approach allows simple computation of information consistency rate bounds from spectrum of kernel operator
- Main result

$$\underbrace{-\log P_{bs}(\boldsymbol{y}_{\leq n})}_{L_{bs}(\boldsymbol{y}_{\leq n})} \leq \underbrace{-\log P(\boldsymbol{y}_{\leq n}|f)}_{L_f(\boldsymbol{y}_{\leq n})} + \frac{1}{2}\|f\|_K^2 + \frac{1}{2}\underbrace{\log|\boldsymbol{I} + c\boldsymbol{K}|}_{\text{E}[\cdot]:\text{Regret term}}$$

for any $\boldsymbol{y}_{\leq n} \in \mathbb{R}^n$ and any $f \in \mathcal{H}_K$.
$\boldsymbol{K} = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))$ kernel matrix

c bounds $-d^2 \log P(y|u)/du^2$

Proof Idea

- First $f \in \mathcal{H}_n := \mathrm{span}\{K(\cdot, \boldsymbol{x}_i)\}$. Variational inequality:

$$\underbrace{-\log P_{bs}(\boldsymbol{y}_{\leq n})}_{\text{“Free Energy”}} \leq \underbrace{\mathrm{E}_Q\left[-\log P(\boldsymbol{y}_{\leq n}|u(\cdot))\right]}_{\text{“Energy”}} + \underbrace{\mathrm{D}[Q \,\|\, P_{bs}]}_{\text{“Entropy”}} \quad \forall dQ(f)$$

  Fenchel Duality. Similar to PAC-Bayesian technique

- Use specific GP Q with $\mathrm{E}_Q[u(\cdot)] = f(\cdot)$:

$$\inf_{f \in \mathcal{H}_n} -\log P(\boldsymbol{y}_{\leq n}|f) + \frac{1}{2}\|f\|_K^2 \underbrace{\geq -\log P_{bs}(\boldsymbol{y}_{\leq n}) - \frac{1}{2}\log|\boldsymbol{I} + c\boldsymbol{K}|}_{\text{Equality for } P(y|u)=N(y|u,c^{-1})}$$

- $-\log P(\boldsymbol{y}_{\leq n}|f)$ depends on $f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)$ only
  $\Rightarrow \inf_{\mathcal{H}_n}[\cdot] = \inf_{\mathcal{H}}[\cdot]$ by representer theorem

Proof Idea

- First $f \in \mathcal{H}_n := \mathrm{span}\{K(\cdot, \boldsymbol{x}_i)\}$. Variational inequality:

$$\underbrace{-\log P_{bs}(\boldsymbol{y}_{\leq n})}_{\text{“Free Energy”}} \leq \underbrace{\mathrm{E}_Q\left[-\log P(\boldsymbol{y}_{\leq n}|u(\cdot))\right]}_{\text{“Energy”}} + \underbrace{\mathrm{D}[Q \,\|\, P_{bs}]}_{\text{“Entropy”}} \quad \forall dQ(f)$$

  Fenchel Duality. Similar to PAC-Bayesian technique

- Use specific GP Q with $\mathrm{E}_Q[u(\cdot)] = f(\cdot)$:

$$\inf_{f \in \mathcal{H}_n} -\log P(\boldsymbol{y}_{\leq n}|f) + \frac{1}{2}\|f\|_K^2 \underbrace{\geq -\log P_{bs}(\boldsymbol{y}_{\leq n}) - \frac{1}{2}\log|\boldsymbol{I} + c\boldsymbol{K}|}_{\text{Equality for } P(y|u)=N(y|u,c^{-1})}$$

- $-\log P(\boldsymbol{y}_{\leq n}|f)$ depends on $f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)$ only
  $\Rightarrow \inf_{\mathcal{H}_n}[\cdot] = \inf_{\mathcal{H}}[\cdot]$ by representer theorem

Proof Idea

- First $f \in \mathcal{H}_n := \mathrm{span}\{K(\cdot, \boldsymbol{x}_i)\}$. Variational inequality:

$$\underbrace{-\log P_{bs}(\boldsymbol{y}_{\leq n})}_{\text{“Free Energy”}} \leq \underbrace{\mathrm{E}_Q\left[-\log P(\boldsymbol{y}_{\leq n}|u(\cdot))\right]}_{\text{“Energy”}} + \underbrace{\mathrm{D}[Q \,\|\, P_{bs}]}_{\text{“Entropy”}} \quad \forall dQ(f)$$

  Fenchel Duality. Similar to PAC-Bayesian technique

- Use specific GP Q with $\mathrm{E}_Q[u(\cdot)] = f(\cdot)$:

$$\inf_{f \in \mathcal{H}_n} -\log P(\boldsymbol{y}_{\leq n}|f) + \frac{1}{2}\|f\|_K^2 \underbrace{\geq -\log P_{bs}(\boldsymbol{y}_{\leq n}) - \frac{1}{2}\log|\boldsymbol{I} + c\boldsymbol{K}|}_{\text{Equality for } P(y|u)=N(y|u,c^{-1})}$$

- $-\log P(\boldsymbol{y}_{\leq n}|f)$ depends on $f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)$ only
  $\Rightarrow \inf_{\mathcal{H}_n}[\cdot] = \inf_{\mathcal{H}}[\cdot]$ by representer theorem

The Regret Term

- $R = \log |\boldsymbol{I} + c\boldsymbol{K}|$. Information consistency if $\mathrm{E}[R]/n \to 0$

$$R = \log \left| \boldsymbol{I} + cn\boldsymbol{U}\hat{\boldsymbol{\Lambda}}\boldsymbol{U}^T \right| = \sum_{i=1}^{n} \log(1 + cn\hat{\lambda}_i)$$

$\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_n$. $\hat{\lambda}_j = 0$ for $j > n$

- Stationary kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = K(\boldsymbol{x}_i - \boldsymbol{x}_j)$, probability $d\mu$: Kernel operator on $\mathrm{L}_2(d\mu)$, discrete spectrum $\lambda_1 \geq \lambda_2 \geq \ldots$ Summable: $\sum_{i \geq 1} \lambda_i = \int K(\boldsymbol{x}, \boldsymbol{x})\, d\mu(\boldsymbol{x}) = K(\boldsymbol{0}) < \infty$
- Split sum into two parts:

$$\mathrm{E}[R] = S_1 + S_2, \quad S_1 = \mathrm{E}\left[\sum_{i=1}^{s_0} \log(1 + cn\hat{\lambda}_i)\right] = O(s_0 \log n),$$

$$S_2 = \mathrm{E}\left[\sum_{i > s_0} \log(1 + cn\hat{\lambda}_i)\right] \underbrace{\leq cn\mathrm{E}\left[\sum_{i > s_0} \hat{\lambda}_i\right]}_{\log(1+x) \leq x}$$

The Regret Term

- $R = \log|\boldsymbol{I} + c\boldsymbol{K}|$. Information consistency if $\mathrm{E}[R]/n \to 0$

$$R = \log\left|\boldsymbol{I} + cn\boldsymbol{U}\hat{\boldsymbol{\Lambda}}\boldsymbol{U}^T\right| = \sum_{i=1}^{n} \log(1 + cn\hat{\lambda}_i)$$

  $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_n$. $\hat{\lambda}_j = 0$ for $j > n$
- Stationary kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = K(\boldsymbol{x}_i - \boldsymbol{x}_j)$, probability $d\mu$:
  Kernel operator on $\mathrm{L}_2(d\mu)$, discrete spectrum $\lambda_1 \geq \lambda_2 \geq \ldots$
  Summable: $\sum_{i\geq 1} \lambda_i = \int K(\boldsymbol{x}, \boldsymbol{x})\, d\mu(\boldsymbol{x}) = K(\boldsymbol{0}) < \infty$
- Split sum into two parts:

$$\mathrm{E}[R] = S_1 + S_2, \quad S_1 = \mathrm{E}\left[\sum_{i=1}^{s_0} \log(1 + cn\hat{\lambda}_i)\right] = O(s_0 \log n),$$

$$S_2 = \mathrm{E}\left[\sum_{i>s_0} \log(1 + cn\hat{\lambda}_i)\right] \underbrace{\leq cn\mathrm{E}\left[\sum_{i>s_0} \hat{\lambda}_i\right]}_{\log(1+x)\leq x}$$

The Regret Term

- $R = \log|\boldsymbol{I} + c\boldsymbol{K}|$. Information consistency if $\mathrm{E}[R]/n \to 0$

$$R = \log\left|\boldsymbol{I} + cn\boldsymbol{U}\hat{\boldsymbol{\Lambda}}\boldsymbol{U}^T\right| = \sum_{i=1}^{n} \log(1 + cn\hat{\lambda}_i)$$

  $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_n$. $\hat{\lambda}_j = 0$ for $j > n$
- Stationary kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = K(\boldsymbol{x}_i - \boldsymbol{x}_j)$, probability $d\mu$:
  Kernel operator on $\mathrm{L}_2(d\mu)$, discrete spectrum $\lambda_1 \geq \lambda_2 \geq \ldots$
  Summable: $\sum_{i\geq 1} \lambda_i = \int K(\boldsymbol{x}, \boldsymbol{x})\, d\mu(\boldsymbol{x}) = K(\boldsymbol{0}) < \infty$
- Split sum into two parts:

$$\mathrm{E}[R] = S_1 + S_2, \quad S_1 = \mathrm{E}\left[\sum_{i=1}^{s_0} \log(1 + cn\hat{\lambda}_i)\right] = O(s_0 \log n),$$

$$S_2 = \mathrm{E}\left[\sum_{i>s_0} \log(1 + cn\hat{\lambda}_i)\right] \underbrace{\leq cn\mathrm{E}\left[\sum_{i>s_0} \hat{\lambda}_i\right]}_{\log(1+x)\leq x}$$

The Regret Term (II)

- Empirical eigenvalues are overestimates — Shawe-Taylor *et.al.*, IEEE IT 05

$$\sum_{i\geq 1} \hat{\lambda}_i = n^{-1} \operatorname{tr} \boldsymbol{K} = K(\mathbf{0}) = \sum_{i\geq 1} \lambda_i,$$
$$\sum_{i=1}^{s_0} \hat{\lambda}_i \geq \sum_{i=1}^{s_0} \lambda_i$$

- For any s_0:

$$S_2 \leq cn \sum_{i>s_0} \hat{\lambda}_i \stackrel{!}{\leq} cn \sum_{i>s_0} \lambda_i$$

- All in all: For some $C \approx 1$:

$$\mathrm{E}[R]/n = (S_1 + S_2)/n \leq Cs_0(\log n)/n + c \sum_{i>s_0} \lambda_i$$

Choose $s_0(n) \to \infty$, $s_0(n)(\log n)/n \to 0$.
$\Rightarrow$ Information consistency

Thanks to Ingo Steinwart

The Regret Term (II)

- Empirical eigenvalues are overestimates — Shawe-Taylor *et.al.*, IEEE IT 05

$$\sum_{i\geq 1} \hat{\lambda}_i = n^{-1} \operatorname{tr} \boldsymbol{K} = K(\mathbf{0}) = \sum_{i\geq 1} \lambda_i,$$
$$\sum_{i=1}^{s_0} \hat{\lambda}_i \geq \sum_{i=1}^{s_0} \lambda_i$$

- For any s_0:

$$S_2 \leq cn \sum_{i>s_0} \hat{\lambda}_i \stackrel{!}{\leq} cn \sum_{i>s_0} \lambda_i$$

- All in all: For some $C \approx 1$:

$$\mathrm{E}[R]/n = (S_1 + S_2)/n \leq Cs_0(\log n)/n + c \sum_{i>s_0} \lambda_i$$

Choose $s_0(n) \to \infty$, $s_0(n)(\log n)/n \to 0$.
$\Rightarrow$ Information consistency

Thanks to Ingo Steinwart

The Regret Term (II)

- Empirical eigenvalues are overestimates — Shawe-Taylor *et.al.*, IEEE IT 05

$$\sum_{i\geq 1} \hat{\lambda}_i = n^{-1} \operatorname{tr} \boldsymbol{K} = K(\mathbf{0}) = \sum_{i\geq 1} \lambda_i,$$
$$\sum_{i=1}^{s_0} \hat{\lambda}_i \geq \sum_{i=1}^{s_0} \lambda_i$$

- For any s_0:

$$S_2 \leq cn \sum_{i>s_0} \hat{\lambda}_i \stackrel{!}{\leq} cn \sum_{i>s_0} \lambda_i$$

- All in all: For some $C \approx 1$:

$$\mathrm{E}[R]/n = (S_1 + S_2)/n \leq Cs_0(\log n)/n + c \sum_{i>s_0} \lambda_i$$

Choose $s_0(n) \to \infty$, $s_0(n)(\log n)/n \to 0$.
$\Rightarrow$ Information consistency

Thanks to Ingo Steinwart

Information Consistency Rates

- Is R concentrated ($\mathrm{Var}[R] \leq \mathrm{E}[R]$)?
  - $\mathrm{E}[R]$ can be small. Standard global approach (McDiarmid) not strong enough
  - Local results on $\sum_{i>s_0} \hat{\lambda}_i$ could be useful

  Zwald *et.al.*, Mach. Learn. 07
- For fixed stationary K, μ:
  Bounds on $\sum_{i>s_0} \lambda_i \Rightarrow$ Information consistency rate bounds
- Gaussian $K(\boldsymbol{r}) = \exp(-b\|\boldsymbol{r}\|^2)$, Gaussian $\mu(\boldsymbol{x}) = N(\boldsymbol{x}|\boldsymbol{0}, (4a)^{-1}\boldsymbol{I})$, $\boldsymbol{r}, \boldsymbol{x} \in \mathbb{R}^d$: Spectrum known explicitly

  Zhu *et.al.*, NATO ASI 98

$$\mathrm{E}[R] = O\left((\log n)^{d+1}\right)$$

Very small (infinite dimensional $\mathcal{H}_K$):
Smoothness through kernel for nonparametric statistic $\leftrightarrow$ dimensionality for parametric statistics.
$\Rightarrow$ Gaussian kernel extremely smoothing

Information Consistency Rates

- Is R concentrated ($\mathrm{Var}[R] \leq \mathrm{E}[R]$)?
  - $\mathrm{E}[R]$ can be small. Standard global approach (McDiarmid) not strong enough
  - Local results on $\sum_{i>s_0} \hat{\lambda}_i$ could be useful — Zwald *et.al.*, Mach. Learn. 07
- For fixed stationary K, μ:
  Bounds on $\sum_{i>s_0} \lambda_i \Rightarrow$ Information consistency rate bounds
- Gaussian $K(\boldsymbol{r}) = \exp(-b\|\boldsymbol{r}\|^2)$, Gaussian $\mu(\boldsymbol{x}) = N(\boldsymbol{x}|\boldsymbol{0}, (4a)^{-1}\boldsymbol{I})$, $\boldsymbol{r}, \boldsymbol{x} \in \mathbb{R}^d$: Spectrum known explicitly — Zhu *et.al.*, NATO ASI 98

$$\mathrm{E}[R] = O\left((\log n)^{d+1}\right)$$

Very small (infinite dimensional $\mathcal{H}_K$):
Smoothness through kernel for nonparametric statistic $\leftrightarrow$ dimensionality for parametric statistics.
$\Rightarrow$ Gaussian kernel extremely smoothing

Information Consistency Rates

- Is R concentrated ($\mathrm{Var}[R] \leq \mathrm{E}[R]$)?
  - $\mathrm{E}[R]$ can be small. Standard global approach (McDiarmid) not strong enough
  - Local results on $\sum_{i>s_0} \hat{\lambda}_i$ could be useful

  Zwald *et.al.*, Mach. Learn. 07
- For fixed stationary K, μ:
  Bounds on $\sum_{i>s_0} \lambda_i \Rightarrow$ Information consistency rate bounds
- Gaussian $K(\boldsymbol{r}) = \exp(-b\|\boldsymbol{r}\|^2)$, Gaussian $\mu(\boldsymbol{x}) = N(\boldsymbol{x}|\boldsymbol{0}, (4a)^{-1}\boldsymbol{I})$, $\boldsymbol{r}, \boldsymbol{x} \in \mathbb{R}^d$: Spectrum known explicitly

  Zhu *et.al.*, NATO ASI 98

$$\mathrm{E}[R] = O\left((\log n)^{d+1}\right)$$

Very small (infinite dimensional $\mathcal{H}_K$):
Smoothness through kernel for nonparametric statistic $\leftrightarrow$ dimensionality for parametric statistics.
$\Rightarrow$ Gaussian kernel extremely smoothing

Information Consistency Rates

- Is R concentrated ($\mathrm{Var}[R] \leq \mathrm{E}[R]$)?
  - $\mathrm{E}[R]$ can be small. Standard global approach (McDiarmid) not strong enough
  - Local results on $\sum_{i>s_0} \hat{\lambda}_i$ could be useful

  Zwald *et.al.*, Mach. Learn. 07
- For fixed stationary K, μ:
  Bounds on $\sum_{i>s_0} \lambda_i \Rightarrow$ Information consistency rate bounds
- Gaussian $K(\boldsymbol{r}) = \exp(-b\|\boldsymbol{r}\|^2)$, Gaussian $\mu(\boldsymbol{x}) = N(\boldsymbol{x}|\boldsymbol{0}, (4a)^{-1}\boldsymbol{I})$, $\boldsymbol{r}, \boldsymbol{x} \in \mathbb{R}^d$: Spectrum known explicitly

  Zhu *et.al.*, NATO ASI 98

$$\mathrm{E}[R] = O\left((\log n)^{d+1}\right)$$

Very small (infinite dimensional $\mathcal{H}_K$):
Smoothness through kernel for nonparametric statistic $\leftrightarrow$ dimensionality for parametric statistics.
$\Rightarrow$ Gaussian kernel extremely smoothing

Information Consistency Rates (II)

- Widom (Trans. AMS 63): Asymptotic expressions for λ_i if
  - Spectral density (Fourier transform) of K decays polynomially
  - $d\mu$ has density with bounded support
- Matérn class: Kernels with Student's t spectral densities: Ornstein-Uhlenbeck ($\nu = 1/2$) $\leftrightarrow$ Gaussian ($\nu \to \infty$)
- K Matérn (d.o.f. $\nu > 0$), $\mu(\boldsymbol{x})$ bounded support, $\boldsymbol{x} \in \mathbb{R}^d$:

$$\mathrm{E}[R] = O\left(n^{d/(2\nu+d)}(\log n)^{2\nu/(2\nu+d)}\right)$$

  Drawback: Constant as large as $|\mathrm{supp}\mu|^{2\nu+d}$
- K Matérn (d.o.f. ν), $\tilde{\mu}$ Student's t (d.o.f. $\nu_2 > \nu$), $\mu(\boldsymbol{x}) \propto \tilde{\mu}(\boldsymbol{x})\mathrm{I}_{\{\|\boldsymbol{x}\| \leq T\}}$:

$$\mathrm{E}[R] = O\left(n^{d/(2\nu+d)}(\log n)^{2\nu/(2\nu+d)}\right)$$

  Constant independent of T.
  But speed of convergence may depend on T

Information Consistency Rates (II)

- Widom (Trans. AMS 63): Asymptotic expressions for λ_i if
  - Spectral density (Fourier transform) of K decays polynomially
  - $d\mu$ has density with bounded support
- Matérn class: Kernels with Student's t spectral densities: Ornstein-Uhlenbeck ($\nu = 1/2$) $\leftrightarrow$ Gaussian ($\nu \to \infty$)
- K Matérn (d.o.f. $\nu > 0$), $\mu(\boldsymbol{x})$ bounded support, $\boldsymbol{x} \in \mathbb{R}^d$:

$$\mathrm{E}[R] = O\left(n^{d/(2\nu+d)}(\log n)^{2\nu/(2\nu+d)}\right)$$

  Drawback: Constant as large as $|\mathrm{supp}\mu|^{2\nu+d}$
- K Matérn (d.o.f. ν), $\tilde{\mu}$ Student's t (d.o.f. $\nu_2 > \nu$), $\mu(\boldsymbol{x}) \propto \tilde{\mu}(\boldsymbol{x})\mathrm{I}_{\{\|\boldsymbol{x}\| \leq T\}}$:

$$\mathrm{E}[R] = O\left(n^{d/(2\nu+d)}(\log n)^{2\nu/(2\nu+d)}\right)$$

  Constant independent of T.
  But speed of convergence may depend on T

Information Consistency Rates (II)

- Widom (Trans. AMS 63): Asymptotic expressions for λ_i if
  - Spectral density (Fourier transform) of K decays polynomially
  - $d\mu$ has density with bounded support
- Matérn class: Kernels with Student's t spectral densities: Ornstein-Uhlenbeck ($\nu = 1/2$) $\leftrightarrow$ Gaussian ($\nu \to \infty$)
- K Matérn (d.o.f. $\nu > 0$), $\mu(\boldsymbol{x})$ bounded support, $\boldsymbol{x} \in \mathbb{R}^d$:

$$\mathrm{E}[R] = O\left(n^{d/(2\nu+d)}(\log n)^{2\nu/(2\nu+d)}\right)$$

  Drawback: Constant as large as $|\mathrm{supp}\mu|^{2\nu+d}$
- K Matérn (d.o.f. ν), $\tilde{\mu}$ Student's t (d.o.f. $\nu_2 > \nu$), $\mu(\boldsymbol{x}) \propto \tilde{\mu}(\boldsymbol{x})\mathrm{I}_{\{\|\boldsymbol{x}\| \le T\}}$:

$$\mathrm{E}[R] = O\left(n^{d/(2\nu+d)}(\log n)^{2\nu/(2\nu+d)}\right)$$

  Constant independent of T.
  But speed of convergence may depend on T

Information Consistency Rates (II)

- Widom (Trans. AMS 63): Asymptotic expressions for λ_i if
  - Spectral density (Fourier transform) of K decays polynomially
  - $d\mu$ has density with bounded support
- Matérn class: Kernels with Student's t spectral densities: Ornstein-Uhlenbeck ($\nu = 1/2$) $\leftrightarrow$ Gaussian ($\nu \to \infty$)
- K Matérn (d.o.f. $\nu > 0$), $\mu(\boldsymbol{x})$ bounded support, $\boldsymbol{x} \in \mathbb{R}^d$:

$$\mathrm{E}[R] = O\left(n^{d/(2\nu+d)}(\log n)^{2\nu/(2\nu+d)}\right)$$

  Drawback: Constant as large as $|\mathrm{supp}\mu|^{2\nu+d}$
- K Matérn (d.o.f. ν), $\tilde{\mu}$ Student's t (d.o.f. $\nu_2 > \nu$), $\mu(\boldsymbol{x}) \propto \tilde{\mu}(\boldsymbol{x})\mathrm{I}_{\{\|\boldsymbol{x}\| \leq T\}}$:

$$\mathrm{E}[R] = O\left(n^{d/(2\nu+d)}(\log n)^{2\nu/(2\nu+d)}\right)$$

  Constant independent of T.
  But speed of convergence may depend on T

Conclusions

- Nonparametric information consistency of GP prediction for wide range of stationary kernels
- Information consistency rates reduced to kernel operator spectrum (entropy and covering numbers can be used; I. Steinwart)
- Nonparametrics: Smoothness more important than dimensionality
- Nonparametric MDL beyond Shtarkov Kakade *et.al.*, NIPS 06; Grünwald 07, Sect. 13.5
- Open problems
  - Concentration of regret term
  - Regret term for unbounded support $d\mu$

Conclusions

- Nonparametric information consistency of GP prediction for wide range of stationary kernels
- Information consistency rates reduced to kernel operator spectrum (entropy and covering numbers can be used; I. Steinwart)
- Nonparametrics: Smoothness more important than dimensionality
- Nonparametric MDL beyond Shtarkov Kakade *et.al.*, NIPS 06; Grünwald 07, Sect. 13.5
- Open problems
  - Concentration of regret term
  - Regret term for unbounded support $d\mu$

Conclusions

- Nonparametric information consistency of GP prediction for wide range of stationary kernels
- Information consistency rates reduced to kernel operator spectrum (entropy and covering numbers can be used; I. Steinwart)
- Nonparametrics: Smoothness more important than dimensionality
- Nonparametric MDL beyond Shtarkov Kakade *et.al.*, NIPS 06; Grünwald 07, Sect. 13.5
- Open problems
  - Concentration of regret term
  - Regret term for unbounded support $d\mu$

Conclusions

- Nonparametric information consistency of GP prediction for wide range of stationary kernels
- Information consistency rates reduced to kernel operator spectrum (entropy and covering numbers can be used; I. Steinwart)
- Nonparametrics: Smoothness more important than dimensionality
- Nonparametric MDL beyond Shtarkov Kakade *et.al.*, NIPS 06; Grünwald 07, Sect. 13.5
- Open problems
  - Concentration of regret term
  - Regret term for unbounded support $d\mu$

Conclusions

- Nonparametric information consistency of GP prediction for wide range of stationary kernels
- Information consistency rates reduced to kernel operator spectrum (entropy and covering numbers can be used; I. Steinwart)
- Nonparametrics: Smoothness more important than dimensionality
- Nonparametric MDL beyond Shtarkov Kakade *et.al.*, NIPS 06; Grünwald 07, Sect. 13.5
- Open problems
  - Concentration of regret term
  - Regret term for unbounded support $d\mu$