

PAC-Bayesian Bound for Gaussian Process Regression and Multiple Kernel Additive Model

† Taiji Suzuki

†Tokyo Institute of Technology
Department of Mathematical Computing Sciences

15th/March/2014@ISM

Taiji Suzuki: PAC-Bayesian Bound for Gaussian Process Regression and Multiple Kernel Additive Model.

Conference on Learning Theory (COLT2012), *JMLR Workshop and Conference Proceedings 23*, pp. 8.1 – 8.20, 2012.

Taiji Suzuki, and Masashi Sugiyama: Fast learning rate of multiple kernel learning: trade-off between sparsity and smoothness.

The Annals of Statistics, vol. 41, number 3, pp. 1381-1405, 2013.

Taiji Suzuki: Unifying Framework for Fast Learning Rate of Non-Sparse Multiple Kernel Learning.

Advances in Neural Information Processing Systems 24 (NIPS2011). pp.1575–1583.

Outline

- 1 Problem Setting
- 2 Multiple Kernel Learning
- 3 Gaussian Process Regression
- 4 Bayesian MKL
- 5 Convergence Rate of Bayesian MKL
 - PAC-Bayesian Bound
 - Main Result
 - Applications to Some Examples

Outline

- 1 Problem Setting
- 2 Multiple Kernel Learning
- 3 Gaussian Process Regression
- 4 Bayesian MKL
- 5 Convergence Rate of Bayesian MKL
 - PAC-Bayesian Bound
 - Main Result
 - Applications to Some Examples

Sparse estimation [Lasso]

Design matrix $X = (X_{ij}) \in \mathbb{R}^{n \times p}$. p (dimension) $\gg n$ (# of samples).
True coefficients $\beta^* \in \mathbb{R}^p$: only $d (< p)$ elements are non-zeros
(d -sparse).

$$\text{Model : } Y = X\beta^* + \xi.$$

$$\hat{\beta} \leftarrow \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{n} \|X\beta - Y\|_2^2 + \lambda_n \|\beta\|_1.$$

Sparse estimation [Lasso]

Design matrix $X = (X_{ij}) \in \mathbb{R}^{n \times p}$. p (dimension) $\gg n$ (# of samples).
 True coefficients $\beta^* \in \mathbb{R}^p$: only $d (< p)$ elements are non-zeros
 (d -sparse).

$$\text{Model : } Y = X\beta^* + \xi.$$

$$\hat{\beta} \leftarrow \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{n} \|X\beta - Y\|_2^2 + \lambda_n \|\beta\|_1.$$

Theorem (Lasso's convergence rate (Bickel et al., 2009; Zhang, 2009))

If the design matrix satisfies "**Restricted eigenvalue condition**",
 $\max_{i,j} |X_{ij}| \leq 1$, and the noise satisfy $E[e^{\tau \xi_i}] \leq e^{\sigma^2 \tau^2 / 2}$ ($\forall \tau > 0$), then we
 have, with probability $1 - \delta$,

$$\|\hat{\beta} - \beta^*\|_2^2 \leq C \frac{d \log(p/\delta)}{n}.$$

p 's effect is just $\log(p)$, the effective dimension d is dominant .

Restricted eigenvalue condition (Bickel et al., 2009; Zhang, 2009)

$$\phi_b(I) := \sup \left\{ \phi \geq 0 \mid \phi \leq \frac{\beta^\top X^\top X \beta / n}{\sum_{j \in I} \beta_j^2}, \right. \\ \left. \forall \beta \in \mathbb{R}^p \text{ such that } b \sum_{j \in I} |\beta_j| \geq \sum_{j \notin I} |\beta_j| \right\}.$$

Restricted Eigenvalue Condition

There exists a constant $0 < C$ such that

$$0 < C < \min_{I: b_0 \subset I, |I| \leq 2d} \phi_3(I).$$

Motivation: Is there any estimator which does not require this condition to achieve a similar convergence rate?

→ **Bayesian estimator**

Goal of this talk

- Change the risk criterion:

$$\|\hat{\beta} - \beta\|_2^2 \longrightarrow \frac{1}{n} \|\mathbf{X}(\hat{\beta} - \beta^*)\|_2^2$$

- We generalize the model to non-parametric (sparse) additive model:

$$f(x) = \sum_{j=1}^p x_j \beta_j \longrightarrow f(x) = \sum_{m=1}^M f_m(x^{(m)})$$

- Finally, we will derive a risk bound of Bayesian sparse estimator:

$$\|\hat{f} - f^*\|_n^2 \leq ?$$

Goal of this talk

- Change the risk criterion:

$$\|\hat{\beta} - \beta\|_2^2 \longrightarrow \frac{1}{n} \|\mathbf{X}(\hat{\beta} - \beta^*)\|_2^2$$

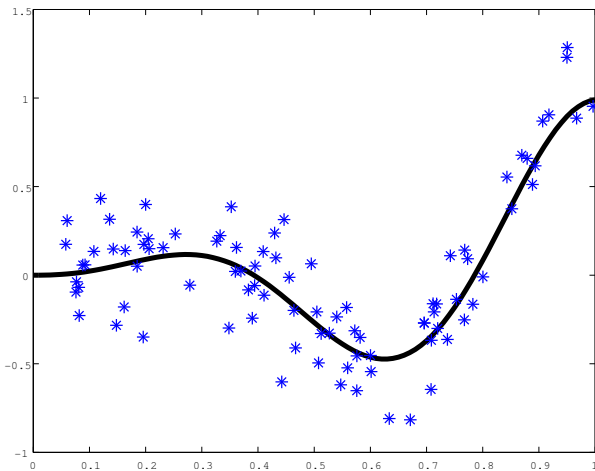
- We generalize the model to non-parametric (sparse) additive model:

$$f(x) = \sum_{j=1}^p x_j \beta_j \longrightarrow f(x) = \sum_{m=1}^M f_m(x^{(m)})$$

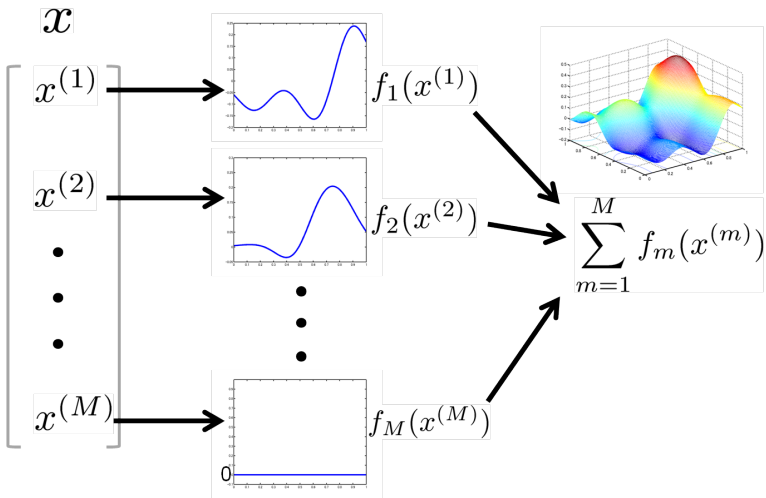
- Finally, we will derive a risk bound of Bayesian sparse estimator:

$$\|\hat{f} - f^*\|_n^2 \leq \sum_{m \in l_0} n^{-\frac{1}{1+s_m}} + \frac{d \log(p/d)}{n}$$

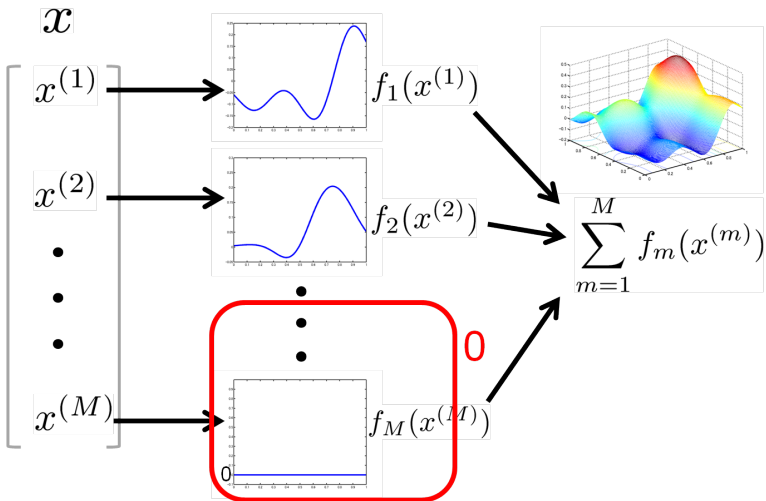
Non-parametric Regression



Sparse Additive Model



Sparse Additive Model



Problem Setting

$$y_i = f^\circ(x_i) + \xi_i, \quad (i = 1, \dots, n),$$

where f° is the true function such that $E[Y|X] = f^\circ(X)$.

f° is well approximated by a function f^* with a sparse representation:

$$f^\circ(x) \simeq f^*(x) = \sum_{m=1}^M f_m^*(x^{(m)}),$$

where only a few components of $\{f_m^*\}_{m=1}^M$ are non-zero.

Outline

- 1 Problem Setting
- 2 Multiple Kernel Learning**
- 3 Gaussian Process Regression
- 4 Bayesian MKL
- 5 Convergence Rate of Bayesian MKL
 - PAC-Bayesian Bound
 - Main Result
 - Applications to Some Examples

Multiple Kernel Learning

$$\min_{f_m \in \mathcal{H}_m} \frac{1}{n} \sum_{i=1}^n \left(y_i - \sum_{m=1}^M f_m(x_i^{(m)}) \right)^2 + C \sum_{m=1}^M \|f_m\|_{\mathcal{H}_m}$$

(\mathcal{H}_m : Reproducing Kernel Hilbert Space (RKHS), explained later)

- Extension of Group Lasso: each group is infinite dimensional
- Sparse solution
- Reduced to finite dimensional optimization problem by the representer theorem (Sonnenburg et al., 2006; Rakotomamonjy et al., 2008; Suzuki & Tomioka, 2009)

Various types of regularization

- L_1 -MKL (Lanckriet et al., 2004; Bach et al., 2004) : **Sparse**

$$\min_{f_m \in \mathcal{H}_m} L \left(\sum_{m=1}^M f_m \right) + C \sum_{m=1}^M \|f_m\|_{\mathcal{H}_m}$$

- L_2 -MKL : **Dense**

$$\min_{f_m \in \mathcal{H}_m} L \left(\sum_{m=1}^M f_m \right) + C \sum_{m=1}^M \|f_m\|_{\mathcal{H}_m}^2$$

Various types of regularization

- L_1 -MKL (Lanckriet et al., 2004; Bach et al., 2004) : **Sparse**

$$\min_{f_m \in \mathcal{H}_m} L \left(\sum_{m=1}^M f_m \right) + C \sum_{m=1}^M \|f_m\|_{\mathcal{H}_m}$$

- L_2 -MKL : **Dense**

$$\min_{f_m \in \mathcal{H}_m} L \left(\sum_{m=1}^M f_m \right) + C \sum_{m=1}^M \|f_m\|_{\mathcal{H}_m}^2$$

- Elasticnet-MKL (Tomioka & Suzuki, 2009)

$$\min_{f_m \in \mathcal{H}_m} L \left(\sum_{m=1}^M f_m \right) + C_1 \sum_{m=1}^M \|f_m\|_{\mathcal{H}_m} + C_2 \sum_{m=1}^M \|f_m\|_{\mathcal{H}_m}^2$$

- ℓ_p -MKL (Kloft et al., 2009)

$$\min_{f_m \in \mathcal{H}_m} L \left(\sum_{m=1}^M f_m \right) + C_1 \left(\sum_{m=1}^M \|f_m\|_{\mathcal{H}_m}^p \right)^{\frac{2}{p}}$$

Sparsity VS accuracy

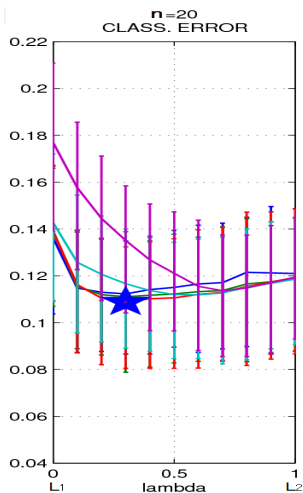


Figure: Relation between accuracy and sparsity of Elasticnet-MKL for caltech

Sparsity VS accuracy

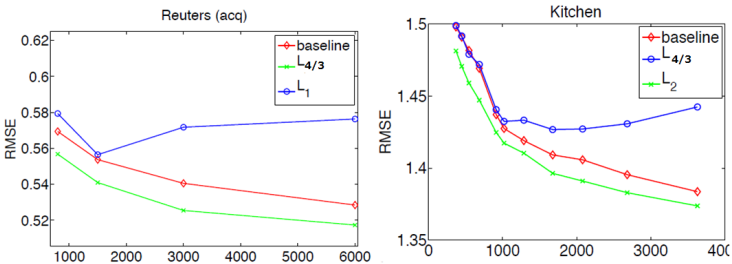


Figure: Relation between accuracy and sparsity of ℓ_p -MKL (Cortes et al., 2009)

Convergence rate of MKL

Suzuki (2011) gave a unifying framework to derive convergence rates of various types of regularizations.

Examples:

- ℓ_p -MKL: $\|f\|_\psi = \left(\sum_{m=1}^M \|f_m\|_{\mathcal{H}_m}^p\right)^{\frac{1}{p}}$

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left(n^{-\frac{1}{1+s}} M^{1-\frac{2s}{p(1+s)}} R_p^{\frac{2s}{1+s}} + \frac{M \log(M)}{n}\right)$$

- Elasticnet-MKL:

$$\|f\|_\psi = \lambda \sum_{m=1}^M \|f_m\|_{\mathcal{H}_m} + (1 - \lambda) \left(\sum_{m=1}^M \|f_m\|_{\mathcal{H}_m}^2\right)^{\frac{1}{2}}$$

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left(n^{-\frac{1}{1+s}} \frac{M^{1-\frac{s}{1+s}}}{(1-\lambda+\lambda\sqrt{M})^{\frac{2s}{1+s}}} [\lambda \|f^*\|_{\ell_1} + (1-\lambda) \|f^*\|_{\ell_2}]^{\frac{2s}{1+s}} + \frac{M \log(M)}{n}\right)$$

- VSKL: $\|f\|_\psi = \|f\|_{(p,q)} = \left[\sum_{j=1}^{M'} \left(\sum_{k=1}^{M_j} \|f_{j,k}\|_{\mathcal{H}_{j,k}}^p\right)^{\frac{q}{p}}\right]^{\frac{1}{q}}$

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left(\frac{\left(\sum_{j=1}^{M'} M_j\right)^{\frac{1-s}{1+s}}}{n^{\frac{1}{1+s}}} \left\{ \left(\sum_{j=1}^{M'} M_j^{\frac{q^*}{p^*}}\right)^{\frac{1}{q^*}} \left[\sum_{j=1}^{M'} \left(\sum_{k=1}^{M_j} \|f_{j,k}^*\|_{\mathcal{H}_{j,k}}^p\right)^{\frac{q}{p}}\right]^{\frac{1}{q}} \right\}^{\frac{2s}{1+s}} + \frac{M \log(M)}{n}\right)$$

Convergence rate L_1 and elastic-net MKL

$$\min_{f_m \in \mathcal{H}_m} L \left(\sum_{m=1}^M f_m \right) + \sum_{m=1}^M (\lambda_1^{(n)} \|f_m\|_n + \lambda_2^{(n)} \|f_m\|_{\mathcal{H}_m} + \lambda_3^{(n)} \|f_m\|_{\mathcal{H}_m}^2).$$

Theorem (Convergence rate of Mixed-Norm-Elasticnet-MKL (Suzuki & Sugiyama, 2013))

Under the conditions stated above, for sufficiently large n , for appropriately chosen $\lambda_1^{(n)}$, $\lambda_2^{(n)}$, $\lambda_3^{(n)}$, we have

$$(L1) \quad \|\hat{f} - f^*\|_{L_2(\Pi)}^2 \leq C' \left(d^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} R_{1,f^*}^{\frac{2s}{1+s}} + \frac{d \log(M)}{n} \right) \eta(t)^2,$$

$$(Elastic) \quad \|\hat{f} - f^*\|_{L_2(\Pi)}^2 \leq C' \left(d^{\frac{1+q}{1+q+s}} n^{-\frac{1+q}{1+q+s}} R_{2,g^*}^{\frac{2s}{1+q+s}} + \frac{d \log(M)}{n} \right) \eta(t)^2,$$

with probability $1 - e^{-t} - e^{-\zeta_n}$ ($\forall t \geq 1$).

$\eta(t) := \max(\sqrt{t}, t/\sqrt{n})$ and, R_{1,f^*} , R_{2,g^*} are defined as

$$R_{1,f^*} := \sum_{m=1}^M \|f_m^*\|_{\mathcal{H}_m}, \quad R_{2,g^*} := \left(\sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2 \right)^{\frac{1}{2}}$$

(Generalized) Restricted Eigenvalue Condition

To prove a fast convergence rate of MKL, we utilize the following (generalized) *Restricted Eigenvalue Condition* (Bickel et al., 2009; Koltchinskii & Yuan, 2010; Suzuki, 2011; Suzuki & Sugiyama, 2012).

Restricted Eigenvalue Condition

There exists a constant $0 < C$ such that

$$0 < C < \beta_{\sqrt{d}}(I_0).$$

$$\beta_b(I) := \sup \left\{ \beta \geq 0 \mid \beta \leq \frac{\|\sum_{m=1}^M f_m\|_{L_2(\Pi)}^2}{\sum_{m \in I} \|f_m\|_{L_2(\Pi)}^2}, \right. \\ \left. \forall f \text{ such that } b \sum_{m \in I} \|f_m\|_{L_2(\Pi)} \geq \sum_{m \notin I} \|f_m\|_{L_2(\Pi)} \right\}.$$

f_m s are not totally correlated inside I_0 and between I_0 and I_0^c .

We investigate a Bayesian variant of MKL.

We show a fast learning rate of it **without** conditions on the design such as the restricted eigenvalue condition.

Outline

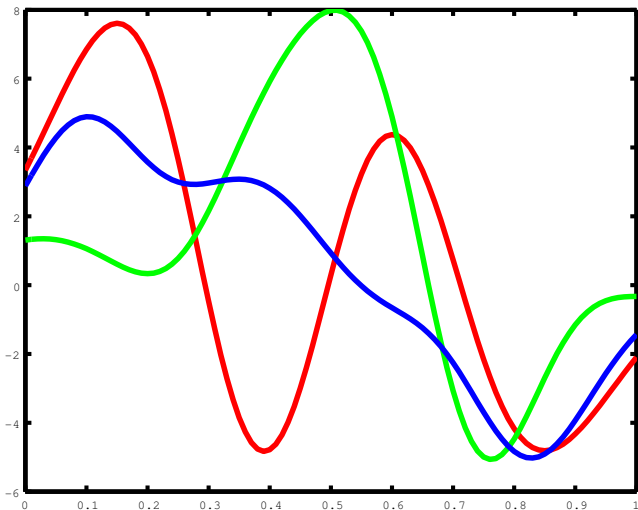
- 1 Problem Setting
- 2 Multiple Kernel Learning
- 3 Gaussian Process Regression**
- 4 Bayesian MKL
- 5 Convergence Rate of Bayesian MKL
 - PAC-Bayesian Bound
 - Main Result
 - Applications to Some Examples

Our proposal = sparse aggregated estimation + Gaussian process

- Aggregated Estimator, Exponential Screening, Model Averaging (Leung & Barron, 2006; Rigollet & Tsybakov, 2011)
- Gaussian Process Regression (Rasmussen & Williams, 2006; van der Vaart & van Zanten, 2008a; van der Vaart & van Zanten, 2008b; van der Vaart & van Zanten, 2011)

Gaussian Process Regression

Gaussian Process Regression



Gaussian Process Regression

Gaussian process prior: a prior on a functions $f = (f(x) : x \in \mathcal{X})$.

$$f \sim \text{GP}$$

means that each finite subset $(f(x_1), f(x_2), \dots, f(x_j))$ ($j = 1, 2, \dots$) obeys a zero-mean multivariate normal distribution.

We assume that $\sup_x |f(x)| < \infty$ and $f : \Omega \rightarrow \ell_\infty(\mathcal{X})$ is tight and Borel measurable.

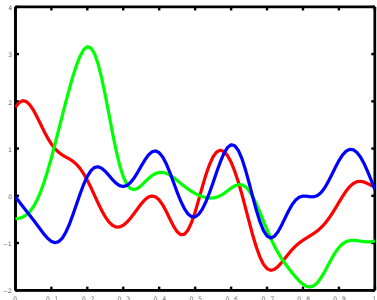
Kernel function:

$$k(x, x') := \mathbb{E}_{f \sim \text{GP}}[f(x)f(x')].$$

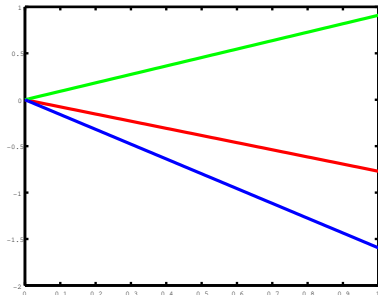
Examples:

- Linear kernel: $k(x, x') = x^\top x'$.
- Gaussian kernel: $k(x, x') = \exp(-\|x - x'\|^2/2\sigma^2)$.
- polynomial kernel: $k(x, x') = (1 + x^\top x')^d$.

Gaussian Process Prior



(a) Gaussian kernel



(b) Linear kernel

Estimation

Suppose $\{y_i\}_{i=1}^n$ are generated from the following model:

$$y_i = f^\circ(x_i) + \xi_i,$$

where ξ_i is i.i.d. Gaussian noise $\mathcal{N}(0, \sigma^2)$.

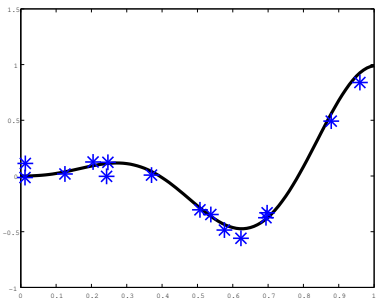
Posterior distribution: let $\mathbf{f} = (f(x_1), \dots, f(x_n))$, then

$$\begin{aligned} p(\mathbf{f}|D_n) &= \frac{1}{C} \exp\left(-n \frac{\|\mathbf{f} - Y\|_n^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2} \mathbf{f}^\top K^{-1} \mathbf{f}\right) \\ &= \frac{1}{C} \exp\left(-\frac{1}{2} \left\| \mathbf{f} - (K + \sigma^2 I_n)^{-1} K Y \right\|_{(K^{-1} + I_n/\sigma^2)}^2\right), \end{aligned}$$

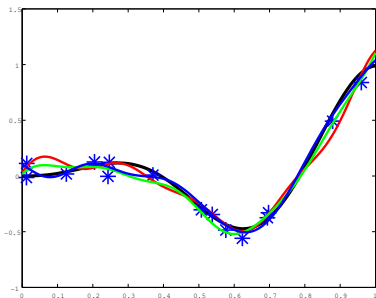
where $K \in \mathbb{R}^{n \times n}$ is the Gram matrix ($K_{i,j} = k(x_i, x_j)$).

- posterior mean: $\hat{\mathbf{f}} = (K + \sigma^2 I_n)^{-1} K Y$.
- posterior covariance: $K - K(K + \sigma^2 I_n)^{-1} K$.

Gaussian Process Posterior

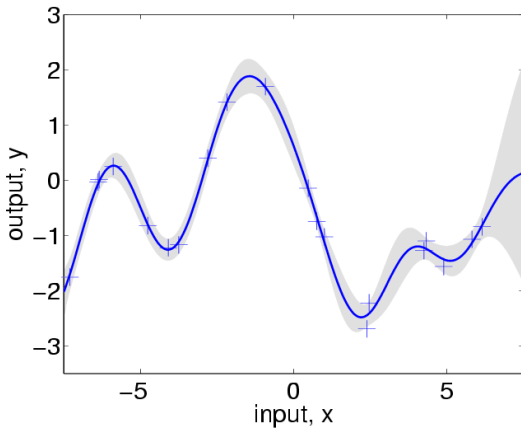


(c) Training Data



(d) Posterior Sample

Gaussian Process Posterior 2



Our interest

How fast does the posterior concentrate around the true?

Reproducing Kernel Hilbert Space (RKHS)

The kernel function defines Reproducing Kernel Hilbert Space (RKHS) \mathcal{H} as a completion of the linear space spanned by all functions

$$x \mapsto \sum_{i=1}^I \alpha_i k(x_i, x), \quad (I = 1, 2, \dots)$$

relative to the RKHS norm $\|\cdot\|_{\mathcal{H}}$ induced by the inner product

$$\left\langle \sum_{i=1}^I \alpha_i k(x_i, \cdot), \sum_{j=1}^J \alpha'_j k(x'_j, \cdot) \right\rangle_{\mathcal{H}} = \sum_{i=1}^I \sum_{j=1}^J \alpha_i \alpha'_j k(x_i, x'_j).$$

Reproducibility: for $f \in \mathcal{H}$, the function value at x is recovered as

$$f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}.$$

Example: Matérn prior

Matérn prior: for a *smoothness* parameter $\alpha > 0$, we define

$$k(x, x') = \int_{\mathbb{R}^d} e^{i\lambda^\top (x-x')} \psi(\lambda) d\lambda,$$

where $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$ is the spectral density given by

$$\psi(\lambda) = \frac{1}{(1 + \|\lambda\|^2)^{\alpha+d/2}}.$$

- The support is included in a Hölder space $C^{\alpha'} [0, 1]^d$ for any $\alpha' < \alpha$.
- The RKHS \mathcal{H} is included in a Sobolev space $W^{\alpha+d/2} [0, 1]^d$ with the regularity $\alpha + d/2$.

For infinite dimensional RKHS \mathcal{H} , the support of the prior is typically much larger than \mathcal{H} .

Convergence rate of posterior: Matérn prior

Let \hat{f} be the posterior mean.

Theorem (van der Vaart and van Zanten (2011))

Let $f^ \in C^\beta[0, 1]^d \cap W^\beta[0, 1]^d$ for $\beta > 0$, then for Matérn prior with parameter α , we have*

$$E[\|\hat{f} - f^*\|_n^2] \leq O\left(n^{-\frac{\alpha \wedge \beta}{\alpha + d/2}}\right).$$

- The optimal rate is $O\left(n^{-\frac{\beta}{\beta + d/2}}\right)$.
- The optimal rate is achieved only when $\alpha = \beta$.
- The rate $n^{-\frac{\alpha \wedge \beta}{\alpha + d/2}}$ is tight.
 → If $f^* \in \mathcal{H}$ ($\beta = \alpha + d/2$), then GP does not achieve the optimal rate.

→ **Scale mixture** is useful (van der Vaart & van Zanten, 2009).

Summary of existing results

- GP is optimal only in quite restrictive situations ($\alpha = \beta$).
 - In particular, if $f^\circ \in \mathcal{H}$, the optimal rate can not be achieved.
- The analysis was given only for restricted classes such as Sobolev and Hölder classes.

Outline

- 1 Problem Setting
- 2 Multiple Kernel Learning
- 3 Gaussian Process Regression
- 4 Bayesian MKL**
- 5 Convergence Rate of Bayesian MKL
 - PAC-Bayesian Bound
 - Main Result
 - Applications to Some Examples

Bayesian-MKL = sparse aggregated estimation + Gaussian process

- **Condition on design:** Does not require any special conditions such as restricted eigenvalue condition.
- **Optimality:** Adaptively achieves the optimal rate for a wide class of true functions. In particular, even if $f^o \in \mathcal{H}$, it achieves the optimal rate.
- **Generality:** The analysis is given for a general class of spaces utilizing the notion of *interpolation spaces* and *the metric entropy*.

Bayesian MKL

We estimate f^o in a Bayesian manner. Let $f = (f_1, \dots, f_M)$.

Prior of Bayesian MKL:

$$\Pi(df) = \sum_{J \subseteq \{1, \dots, M\}} \pi_J \cdot \prod_{m \in J} \int_{\lambda_m \in \mathbb{R}_+} \text{GP}_m(df_m | \lambda_m) \mathcal{G}(d\lambda_m) \cdot \prod_{m \notin J} \delta_0(df_m)$$

Bayesian MKL

We estimate f° in a Bayesian manner. Let $f = (f_1, \dots, f_M)$.

Prior of Bayesian MKL:

$$\Pi(df) = \sum_{J \subseteq \{1, \dots, M\}} \pi_J \cdot \prod_{m \in J} \int_{\lambda_m \in \mathbb{R}_+} \text{GP}_m(df_m | \lambda_m) \mathcal{G}(d\lambda_m) \cdot \prod_{m \notin J} \delta_0(df_m)$$

- $\text{GP}_m(\cdot | \lambda_m)$ with a scale parameter λ_m is a *scaled Gaussian process* corresponding to the kernel function \tilde{k}_{m, λ_m} where

$$\tilde{k}_{m, \lambda_m} = \frac{k_m}{\lambda_m},$$

for some fixed kernel function k_m .

Bayesian MKL

We estimate f° in a Bayesian manner. Let $f = (f_1, \dots, f_M)$.

Prior of Bayesian MKL:

$$\Pi(df) = \sum_{J \subseteq \{1, \dots, M\}} \pi_J \cdot \prod_{m \in J} \int_{\lambda_m \in \mathbb{R}_+} \text{GP}_m(df_m | \lambda_m) \mathcal{G}(d\lambda_m) \cdot \prod_{m \notin J} \delta_0(df_m)$$

- $\text{GP}_m(\cdot | \lambda_m)$ with a scale parameter λ_m is a *scaled Gaussian process* corresponding to the kernel function \tilde{k}_{m, λ_m} where

$$\tilde{k}_{m, \lambda_m} = \frac{k_m}{\lambda_m},$$

for some fixed kernel function k_m .

- $\mathcal{G}(\lambda_m) = \exp(-\lambda_m)$ (Gamma distribution: conjugate prior)
→ **scale mixture**.

Bayesian MKL

We estimate f^o in a Bayesian manner. Let $f = (f_1, \dots, f_M)$.

Prior of Bayesian MKL:

$$\Pi(df) = \sum_{J \subseteq \{1, \dots, M\}} \pi_J \cdot \prod_{m \in J} \int_{\lambda_m \in \mathbb{R}_+} \text{GP}_m(df_m | \lambda_m) \mathcal{G}(d\lambda_m) \cdot \prod_{m \notin J} \delta_0(df_m)$$

- $\text{GP}_m(\cdot | \lambda_m)$ with a scale parameter λ_m is a *scaled Gaussian process* corresponding to the kernel function \tilde{k}_{m, λ_m} where

$$\tilde{k}_{m, \lambda_m} = \frac{k_m}{\lambda_m},$$

for some fixed kernel function k_m .

- $\mathcal{G}(\lambda_m) = \exp(-\lambda_m)$ (Gamma distribution: conjugate prior)
→ **scale mixture**.
- Set all components f_m for $m \notin J$ as 0.

Bayesian MKL

We estimate f° in a Bayesian manner. Let $f = (f_1, \dots, f_M)$.

Prior of Bayesian MKL:

$$\Pi(df) = \sum_{J \subseteq \{1, \dots, M\}} \pi_J \cdot \prod_{m \in J} \int_{\lambda_m \in \mathbb{R}_+} \text{GP}_m(df_m | \lambda_m) \mathcal{G}(d\lambda_m) \cdot \prod_{m \notin J} \delta_0(df_m)$$

- $\text{GP}_m(\cdot | \lambda_m)$ with a scale parameter λ_m is a *scaled Gaussian process* corresponding to the kernel function \tilde{k}_{m, λ_m} where

$$\tilde{k}_{m, \lambda_m} = \frac{k_m}{\lambda_m},$$

for some fixed kernel function k_m .

- $\mathcal{G}(\lambda_m) = \exp(-\lambda_m)$ (Gamma distribution: conjugate prior)
→ **scale mixture**.
- Set all components f_m for $m \notin J$ as 0.
- Put a prior π_J on each sub-model J .

Bayesian MKL

We estimate f^o in a Bayesian manner. Let $f = (f_1, \dots, f_M)$.

Prior of Bayesian MKL:

$$\Pi(df) = \sum_{J \subseteq \{1, \dots, M\}} \pi_J \cdot \prod_{m \in J} \int_{\lambda_m \in \mathbb{R}_+} \text{GP}_m(df_m | \lambda_m) \mathcal{G}(d\lambda_m) \cdot \prod_{m \notin J} \delta_0(df_m)$$

- π_J is given as

$$\pi_J = \frac{\zeta^{|J|}}{\sum_{j=1}^M \zeta^j} \binom{M}{|J|}^{-1},$$

with some $\zeta \in (0, 1)$.

The estimator

The posterior: For some constant $\beta > 0$, the posterior probability measure is given as

$$\Pi(df|D_n) := \frac{\exp\left(-\frac{\sum_{i=1}^n (y_i - \sum_{m=1}^M f_m(x_i))^2}{\beta}\right)}{\int \exp\left(-\frac{\sum_{i=1}^n (y_i - \sum_{m=1}^M \tilde{f}_m(x_i))^2}{\beta}\right) \Pi(d\tilde{f})} \Pi(df),$$

for $f = (f_1, \dots, f_M)$.

The estimator: The Bayesian estimator \hat{f} (Bayesian-MKL estimator) is given as the expectation of the posterior:

$$\hat{f} = \int \sum_{m=1}^M f_m \Pi(df|y_1, \dots, y_n).$$

Point

- Model averaging
- Scale mixture of Gaussian process prior

Outline

- 1 Problem Setting
- 2 Multiple Kernel Learning
- 3 Gaussian Process Regression
- 4 Bayesian MKL
- 5 Convergence Rate of Bayesian MKL**
 - PAC-Bayesian Bound
 - Main Result
 - Applications to Some Examples

Mean Squared Error

We want to bound the mean squared error:

$$\|f^o - \hat{f}\|_n^2 := \frac{1}{n} \sum_{i=1}^n (f^o(x_i) - \hat{f}(x_i))^2,$$

where \hat{f} is the Bayesian estimator. We utilize a *PAC-Bayesian bound*.

PAC-Bayesian Bound

Under some conditions for the noise (explained in the next slide), we have the following theorem.

Theorem (Dalalyan and Tsybakov (2008))

For all probability measure ρ , we have

$$\mathbb{E}_{\mathbf{Y}_{1:n}|\mathbf{X}_{1:n}} \left[\|\hat{f} - f^\circ\|_n^2 \right] \leq \int \|f - f^\circ\|_n^2 d\rho(f) + \frac{\beta \mathcal{K}(\rho, \Pi)}{n},$$

where $\mathcal{K}(\rho, \Pi)$ is the KL-divergence between ρ and Π :

$$\mathcal{K}(\rho, \Pi) := \int \log \left(\frac{d\rho}{d\Pi} \right) d\rho.$$

Noise Condition

Let

$$m_\xi(z) := -\mathbb{E}[\xi_1 \mathbf{1}\{\xi_1 \leq z\}],$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. Then we impose the following assumption on m_ξ .

Assumption

$\mathbb{E}[\xi_1^2] < \infty$ and the measure $m_\xi(z)dz$ is absolutely continuous with respect to the density function $p_\xi(z)$ with a bounded Radon-Nikodym derivative, i.e., there exists a bounded function $g_\xi : \mathbb{R} \rightarrow \mathbb{R}_+$ such that

$$\int_a^b m_\xi(z)dz = \int_a^b g_\xi(z)p_\xi(z)dz, \quad \forall a, b \in \mathbb{R}.$$

- The Gaussian noise $\mathcal{N}(0, \sigma^2)$ satisfies the assumption with $g_\xi(z) = \sigma^2$,
- The uniform distribution on $[-a, a]$ satisfies the assumption with $g_\xi(z) = \max(a^2 - z^2, 0)/2$.

Concentration Function

We define the **concentration function** as

$$\phi_{f_m^*}^{(m)}(\epsilon, \lambda_m) := \underbrace{\inf_{h \in \mathcal{H}_m: \|h - f_m^*\|_n \leq \epsilon} \|h\|_{\mathcal{H}_m, \lambda_m}^2}_{\text{bias}} \underbrace{- \log \text{GP}_m(\{f : \|f\|_n \leq \epsilon\} | \lambda_m)}_{\text{variance}}.$$

It is known that $\phi_{f_m^*}^{(m)}(\epsilon, \lambda_m) \sim -\log \text{GP}_m(\{f : \|f_m^* - f\|_n \leq \epsilon\} | \lambda_m)$ (van der Vaart & van Zanten, 2011; van der Vaart & van Zanten, 2008b).

General Result

Let $l_0 := \{m \mid f_m^* \neq 0\}$, $\check{l}_0 := \{m \in l_0 \mid f_m^* \notin \mathcal{H}_m\}$, and $\kappa := \zeta(1 - \zeta)$.

Theorem (Convergence rate of Bayesian-MKL)

The convergence rate of Bayesian-MKL is bounded as

$$\begin{aligned} \mathbb{E}_{Y_{1:n}|X_{1:n}} \left[\|\hat{f} - f^o\|_n^2 \right] &\leq 2\|f^o - f^*\|_n^2 \\ &+ C_1 \inf_{\epsilon_m, \lambda_m > 0} \left\{ \sum_{m \in l_0} \left(\epsilon_m^2 + \frac{1}{n} \phi_{f_m^*}^{(m)}(\epsilon_m, \lambda_m) + \frac{\lambda_m}{n} - \frac{\log(\lambda_m)}{n} \right) \right. \\ &\quad \left. + \sum_{\substack{m, m' \in \check{l}_0: \\ m \neq m'}} \epsilon_m \epsilon_{m'} + \frac{|l_0|}{n} \log \left(\frac{Me}{\kappa |l_0|} \right) \right\}. \end{aligned}$$

Interpretation of the theorem

Let $\hat{\epsilon}_m^2 = \inf_{\epsilon_m, \lambda_m > 0} \left(\epsilon_m^2 + \frac{1}{n} \phi_{f_m^*}^{(m)}(\epsilon_m, \lambda_m) + \frac{\lambda_m}{n} - \frac{\log(\lambda_m)}{n} \right)$ and suppose $f^\circ = f^*$. Typically $\hat{\epsilon}_m^2$ achieves the optimal learning rate for the *single kernel learning*.

- (Correctly specified) If $f_m^* \in \mathcal{H}_m$ for all m , then we have

$$\mathbb{E}_{Y_{1:n}|X_{1:n}} \left[\|\hat{f} - f^\circ\|_n^2 \right] = O \left[\sum_{m \in I_0} \hat{\epsilon}_m^2 + \frac{|I_0|}{n} \log \left(\frac{Me}{\kappa |I_0|} \right) \right].$$

Optimal learning rate for MKL.

Note that we imposed no condition on the design such as restricted eigenvalue condition.

- (Misspecified) If $f_m^* \notin \mathcal{H}_m$ for all $m \in I_0$, then we have

$$\mathbb{E}_{Y_{1:n}|X_{1:n}} \left[\|\hat{f} - f^\circ\|_n^2 \right] = O \left[\left(\sum_{m \in I_0} \hat{\epsilon}_m \right)^2 + \frac{|I_0|}{n} \log \left(\frac{Me}{\kappa |I_0|} \right) \right].$$

Outline of the Proof

Fix $\epsilon_m, \lambda_m > 0$ arbitrary. Next we define a “representer” element $\tilde{h}_m \in \mathcal{H}_m$ that is close to f_m^* . If $f_m^* \in \mathcal{H}_m$, then set $\tilde{h}_m = f_m^*$. Otherwise, we take $\tilde{h}_m \in \mathcal{H}_{m, \lambda_m}$ such that

$\|\tilde{h}_m\|_{\mathcal{H}_{m, \lambda_m}}^2 \leq 2 \inf_{h \in \mathcal{H}_m: \|h - f_m^*\|_n \leq \epsilon_m} \|h\|_{\mathcal{H}_{m, \lambda_m}}^2$. We substitute the following “dummy” posterior into ρ :

$$\rho(df) = \prod_{m \in I_0} \frac{\int_{\frac{\lambda_m}{2} \leq \tilde{\lambda}_m \leq \lambda_m} \frac{\text{GP}_m(df_m - \tilde{h}_m | \tilde{\lambda}_m) \mathbf{1}\{\|f_m - \tilde{h}_m\|_n \leq \epsilon_m\}}{\text{GP}_m(\{\Delta f_m: \|\Delta f_m\|_n \leq \epsilon_m\} | \tilde{\lambda}_m)} \mathcal{G}(d\tilde{\lambda}_m)}{\mathcal{G}(\{\tilde{\lambda}_m: \frac{\lambda_m}{2} \leq \tilde{\lambda}_m \leq \lambda_m\})} \cdot \prod_{m \notin I_0} \delta_0(df_m).$$

One can show that the KL-divergence between ρ and the prior Π is bounded as

$$\frac{1}{n} \mathcal{K}(\rho, \Pi) \leq C'_1 \sum_{m \in I_0} \left(\frac{1}{n} \phi_{f_m^*}^{(m)}(\epsilon_m, \lambda_m) + \frac{1}{n} \lambda_m - \frac{1}{n} \log(\lambda_m) \right) + \frac{|I_0|}{n} \log \left(\frac{Me}{|I_0| \kappa} \right),$$

where C'_1 is a universal constant. A key to prove this is an infinite dimensional extension of the Brascamp and Lieb inequality (Brascamp & Lieb, 1976; Hargé, 2004). Since $\{\epsilon_m, \lambda_m\}_{m=1}^M$ are arbitrary, this gives the assertion.

Example 1: Metric Entropy Characterization (Correctly specified)

Define the ϵ -covering number $N(\mathcal{B}_{\mathcal{H}_m}, \epsilon, \|\cdot\|_n)$ as the number of $\|\cdot\|_n$ -norm balls covering the unit ball $\mathcal{B}_{\mathcal{H}_m}$ in \mathcal{H}_m .

The metric entropy is its logarithm:

$$\log N(\mathcal{B}_{\mathcal{H}_m}, \epsilon, \|\cdot\|_n).$$

Example 1: Metric Entropy Characterization

We assume that there exists a real value $0 < s_m < 1$ such that

$$\log N(\mathcal{B}_{\mathcal{H}_m}, \epsilon, \|\cdot\|_n) = O(\epsilon^{-2s_m}),$$

where $\mathcal{B}_{\mathcal{H}_m}$ is the unit ball of the RKHS \mathcal{H}_m .

Theorem (Correctly specified)

If $f_m^* \in \mathcal{H}_m$ for all $m \in I_0$, then we have

$$\mathbb{E}_{Y_{1:n}|X_{1:n}} \left[\|\hat{f} - f^o\|_n^2 \right] \leq C \left\{ \sum_{m \in I_0} n^{-\frac{1}{1+s_m}} + \frac{|I_0|}{n} \log \left(\frac{Me}{|I_0|\kappa} \right) \right\} + 2\|f^o - f^*\|_n^2$$

It is known that, if there is no scale mixture prior, the optimal rate $n^{-\frac{1}{1+s_m}}$ can not be achieved (Castillo, 2008).

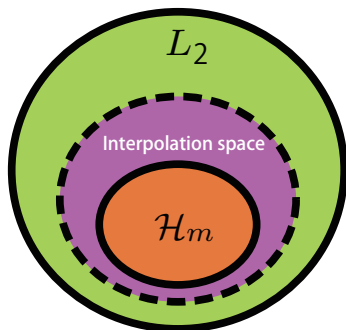
Example 1: Metric Entropy Characterization (Misspecified)

Let $[L_2(P_n), \mathcal{H}_m]_{\theta, \infty}$ be the *interpolation space* equipped with the norm,

$$\|f\|_{\theta, \infty}^{(m)} := \sup_{t>0} t^{-\theta} \inf_{g_m \in \mathcal{H}_m} \{ \|f - g_m\|_n + t \|g_m\|_{\mathcal{H}_m} \}.$$

One has

$$\mathcal{H}_m \hookrightarrow [L_2(P_n), \mathcal{H}_m]_{\theta, \infty} \hookrightarrow L_2(P_n).$$



Example 1: Metric Entropy Characterization (Misspecified)

Let $[L_2(P_n), \mathcal{H}_m]_{\theta, \infty}$ be the *interpolation space* equipped with the norm,

$$\|f\|_{\theta, \infty}^{(m)} := \sup_{t>0} t^{-\theta} \inf_{g_m \in \mathcal{H}_m} \{ \|f - g_m\|_n + t \|g_m\|_{\mathcal{H}_m} \}.$$

One has

$$\mathcal{H}_m \hookrightarrow [L_2(P_n), \mathcal{H}_m]_{\theta, \infty} \hookrightarrow L_2(P_n).$$

Theorem (Misspecified)

If $f_m^* \in [L_2(P_n), \mathcal{H}_m]_{\theta, \infty}$ with $0 < \theta \leq 1$, then we have

$$\mathbb{E}_{Y_{1:n}|X_{1:n}} \left[\|\hat{f} - f^\circ\|_n^2 \right] \leq C \left\{ \left(\sum_{m \in I_0} n^{-\frac{1}{2(1+s_m/\theta)}} \right)^2 + \frac{|I_0|}{n} \log \left(\frac{Me}{|I_0|\kappa} \right) \right\} + 2\|f^\circ - f^*\|_n^2$$

Thanks to the scale mixture prior, the estimator adaptively achieves the optimal rate for $\theta \in (s_m, 1]$.

Example 2: Matérn prior

Suppose that $\mathcal{X}_m = [0, 1]^{d_m}$. The Matérn priors on \mathcal{X}_m correspond to the kernel function defined as

$$k_m(z, z') = \int_{\mathbb{R}^{d_m}} e^{is^\top(z-z')} \psi_m(s) ds,$$

where $\psi_m(s)$ is the spectral density given by $\psi_m(s) = (1 + \|s\|^2)^{-(\alpha_m + d_m/2)}$, for a smoothness parameter $\alpha_m > 0$.

Theorem (Matérn prior, correctly specified)

If $f_m^* \in \mathcal{H}_m$, then we have that

$$\mathbb{E}_{Y_{1:n}|X_{1:n}} \left[\|\hat{f} - f^o\|_n^2 \right] \leq C \left\{ \sum_{m \in I_0} n^{-\frac{1}{1 + \frac{d_m}{2\alpha_m + d_m}}} + \frac{|I_0|}{n} \log \left(\frac{Me}{|I_0|\kappa} \right) \right\} + 2\|f^o - f^*\|_n^2$$

Example 2: Matérn prior

Theorem (Matérn prior, Misspecified)

If $f_m^* \in C^{\beta_m}[0, 1]^{d_m} \cap W^{\beta_m}[0, 1]^{d_m}$ and $\beta_m < \alpha_m + d_m/2$, then we have that

$$\mathbb{E}_{Y_{1:n}|X_{1:n}} \left[\|\hat{f} - f^o\|_n^2 \right] \leq C \left\{ \left(\sum_{m \in I_0} n^{-\frac{\beta_m}{2\beta_m + d_m}} \right)^2 + \frac{|I_0|}{n} \log \left(\frac{Me}{|I_0|\kappa} \right) \right\} + 2\|f^o - f^*\|_n^2$$

Although $f_m^* \notin \mathcal{H}_m$, the convergence rate achieves the optimal rate adaptively.

Example 3: Group Lasso

\mathcal{X}_m is a finite dimensional Euclidean space: $\mathcal{X}_m = \mathbb{R}^{d_m}$. The kernel function corresponding to the Gaussian process prior is $k_m(x, x') = x^\top x'$:

$$f_m(x) = \mu^\top x, \quad \mu \sim \mathcal{N}(0, I_{d_m}).$$

Theorem (Group Lasso)

If $f_m^* = \mu_m^\top x$, then we have that

$$\mathbb{E}_{Y_{1:n}|X_{1:n}} \left[\|\hat{f} - f^o\|_n^2 \right] = C \left\{ \frac{\sum_{m \in I_0} d_m \log(n)}{n} + \frac{|I_0|}{n} \log \left(\frac{Me}{|I_0| \kappa} \right) \right\} + 2\|f^o - f^*\|_n^2$$

This is rate optimal up to $\log(n)$ -order.

Gaussian correlation conjecture

We use an infinite dimensional version of the following inequality (Brascamp-Lieb inequality (Brascamp & Lieb, 1976; Hargé, 2004)):

$$\mathbb{E}[\langle X, \phi \rangle^2 | X \in A] \leq \mathbb{E}[\langle X, \phi \rangle^2],$$

where $X \sim \mathcal{N}(0, \Sigma)$ and A is a symmetric convex set centered on the origin.

Gaussian correlation conjecture:

$$\mu(A \cap B) \geq \mu(A)\mu(B),$$

where μ is any centered Gaussian measure and A and B are any two symmetric convex sets.

Brascamp-Lieb inequality can be seen as an application of a particular case of the Gaussian correlation conjecture. See the survey by Li and Shao (2001) for more details.

Conclusion

- We developed a PAC-Bayesian bound for Gaussian process model and generalized it to sparse additive model.
- The optimal rate is achieved *without* any conditions on the design.
- We have observed that Gaussian processes with scale mixture adaptively achieve the minimax optimal rate on both correctly-specified and misspecified situations.

- Bach, F., Lanckriet, G., & Jordan, M. (2004). Multiple kernel learning, conic duality, and the SMO algorithm. *the 21st International Conference on Machine Learning* (pp. 41–48).
- Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37, 1705–1732.
- Brascamp, H. J., & Lieb, E. H. (1976). On extensions of the brunn-minkowski and prékopa-leindler theorem, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22, 366–389.
- Castillo, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electronic Journal of Statistics*, 2, 1281–1299.
- Cortes, C., Mohri, M., & Rostamizadeh, A. (2009). L_2 regularization for learning kernels. *the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*. Montréal, Canada.
- Dalalyan, A., & Tsybakov, A. B. (2008). Aggregation by exponential weighting sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72, 39–61.
- Hargé, G. (2004). A convex/log-concave correlation inequality for gaussian measure and an application to abstract wiener spaces. *Probability Theory and Related Fields*, 130, 415–440.

- Kloft, M., Brefeld, U., Sonnenburg, S., Laskov, P., Müller, K.-R., & Zien, A. (2009). Efficient and accurate ℓ_p -norm multiple kernel learning. *Advances in Neural Information Processing Systems 22* (pp. 997–1005). Cambridge, MA: MIT Press.
- Koltchinskii, V., & Yuan, M. (2010). Sparsity in multiple kernel learning. *The Annals of Statistics*, 38, 3660–3695.
- Lanckriet, G., Cristianini, N., Ghahoui, L. E., Bartlett, P., & Jordan, M. (2004). Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research*, 5, 27–72.
- Leung, G., & Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52, 3396–3410.
- Li, W. V., & Shao, Q.-M. (2001). Gaussian processes: inequalities, small ball probabilities and applications. *Stochastic Processes: Theory and Methods*, 19, 533–597.
- Rakotomamonjy, A., Bach, F., Canu, S., & Y., G. (2008). SimpleMKL. *Journal of Machine Learning Research*, 9, 2491–2521.
- Rasmussen, C. E., & Williams, C. (2006). *Gaussian processes for machine learning*. MIT Press.

- Rigollet, P., & Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39, 731–771.
- Sonnenburg, S., Rätsch, G., Schäfer, C., & Schölkopf, B. (2006). Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7, 1531–1565.
- Suzuki, T. (2011). Unifying framework for fast learning rate of non-sparse multiple kernel learning. *Advances in Neural Information Processing Systems 24* (pp. 1575–1583). NIPS2011.
- Suzuki, T., & Sugiyama, M. (2012). Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoothness. *JMLR Workshop and Conference Proceedings 22* (pp. 1152–1183). Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS2012).
- Suzuki, T., & Sugiyama, M. (2013). Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoothness. *The Annals of Statistics*, 41, 1381–1405.
- Suzuki, T., & Tomioka, R. (2009). SpicyMKL. arXiv:0909.5026.
- Tomioka, R., & Suzuki, T. (2009). Sparsity-accuracy trade-off in MKL. *NIPS 2009 Workshop:: Understanding Multiple Kernel Learning Methods*. Whistler. arXiv:1001.2615.

- van der Vaart, A. W., & van Zanten, J. H. (2008a). Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36, 1435–1463.
- van der Vaart, A. W., & van Zanten, J. H. (2008b). Reproducing kernel Hilbert spaces of Gaussian priors. *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, 3, 200–222. IMS Collections.
- van der Vaart, A. W., & van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *The Annals of Statistics*, 37, 2655–2675.
- van der Vaart, A. W., & van Zanten, J. H. (2011). Information rates of nonparametric gaussian process methods. *Journal of Machine Learning Research*, 12, 2095–2119.
- Zhang, T. (2009). Some sharp performance bounds for least squares regression with l_1 regularization. *The Annals of Statistics*, 37, 2109–2144.