Regularization Strategies and Empirical Bayesian Learning for MKL

Ryota Tomioka[1], Taiji Suzuki[1]

[1]Department of Mathematical Informatics, The University of Tokyo

2010-12-11
NIPS2010 Workshop:
New Directions in Multiple Kernel Learning

Our contribution

- Relationships between different regularization strategies
  - Ivanov regularization (kernel weights)
  - Tikhonov regularization (kernel weights)
  - (Generalized) block-norm formulation (no kernel weights)

Are they equivalent? — in which way?

- Empirical Bayesian learning algorithm for MKL
  - Maximizes the marginalized likelihood
  - Can be considered as a non-separable regularization on the kernel weights.

Learning with a fixed kernel combination

Fixed kernel combination $k_{\boldsymbol{d}}(x, x') = \sum_{m=1}^{M} d_m k_m(x, x')$.

$$\underset{\substack{\bar{f}\in\mathcal{H}(\boldsymbol{d}),\\ b\in\mathbb{R}}}{\text{minimize}} \quad \sum_{i=1}^{N} \ell\left(y_i, \bar{f}(x_i) + b\right) + \frac{C}{2}\|\bar{f}\|_{\mathcal{H}(\boldsymbol{d})}^2,$$

($\mathcal{H}(\boldsymbol{d})$ is the RKHS corresponding to the combined kernel $k_{\boldsymbol{d}}$) is equivalent to learning M functions $(f_1, \ldots, f_M)$ as follows:

$$\underset{\substack{f_1\in\mathcal{H}_1,\\ \ldots,f_M\in\mathcal{H}_M,\\ b\in\mathbb{R}}}{\text{minimize}} \quad \sum_{i=1}^{N} \ell\left(y_i, \sum_{m=1}^{M} f_m(x_i) + b\right) + \frac{C}{2}\sum_{m=1}^{M} \frac{\|f_m\|_{\mathcal{H}_m}^2}{d_m} \qquad (1)$$

where $\bar{f}(x) = \sum_{m=1}^{M} f_m(x)$.
See Sec. 6 in Aronszajn (1950), Micchelli & Pontil (2005).

Ivanov regularization

We can *constrain* the size of kernel weights d_m by

$$\underset{\substack{f_1\in\mathcal{H}_1,\dots,f_M\in\mathcal{H}_M,\\ b\in\mathbb{R},\\ d_1\geq 0,\dots,d_M\geq 0}}{\text{minimize}} \sum_{i=1}^{N} \ell\left(y_i, \sum_{m=1}^{M} f_m(x_i) + b\right) + \frac{C}{2}\sum_{m=1}^{M}\frac{\|f_m\|_{\mathcal{H}_m}^2}{d_m}, \quad (2)$$

$$\text{s.t.} \quad \sum_{m=1}^{M} h(d_m) \leq 1 \qquad (h \text{ is convex, increasing}).$$

Equivalent to the more common expression:

$$\underset{\substack{f\in\mathcal{H}(\boldsymbol{d}),\\ b\in\mathbb{R},\\ d_1\geq 0,\dots,d_M\geq 0}}{\text{minimize}} \sum_{i=1}^{N} \ell\left(y_i, f(x_i) + b\right) + \frac{C}{2}\|f\|_{\mathcal{H}(\boldsymbol{d})}^2, \ \text{s.t.} \ \sum_{m=1}^{M} h(d_m) \leq 1.$$

Tikhonov regularization

We can *penalize* the size of kernel weights d_m by

$$\underset{\substack{f_1\in\mathcal{H}_1,\dots,f_M\in\mathcal{H}_M,\\ b\in\mathbb{R},\\ d_1\geq 0,\dots,d_M\geq 0}}{\text{minimize}} \sum_{i=1}^{N} \ell\left(y_i, \textstyle\sum_{m=1}^{M} f_m(x_i) + b\right) + \frac{C}{2}\sum_{m=1}^{M}\left(\frac{\|f_m\|_{\mathcal{H}_m}^2}{d_m} + \mu h(d_m)\right). \tag{3}$$

Note that the above is equivalent to

$$\underset{\substack{f\in\mathcal{H}(\boldsymbol{d}),\\ b\in\mathbb{R},\\ d_1\geq 0,\dots,d_M\geq 0}}{\text{minimize}} \underbrace{\sum_{i=1}^{N} \ell\left(y_i, f(x_i) + b\right)}_{\text{data-fit}} + \underbrace{\frac{C}{2}\|f\|_{\mathcal{H}(\boldsymbol{d})}^2}_{f\text{-prior}} + \underbrace{\frac{C\mu}{2}\sum_{m=1}^{M} h(d_m)}_{d_m\text{-hyper-prior}}.$$

Are these two formulations equivalent?

Previously thought that...

Yes. But the choice of the pair (C, μ) is complicated.
$\Rightarrow$ In the Tikhonov formulation we have to choose both C and μ!
(Kloft et al., 2010)

We show that...

If you give up the constant 1 in the Ivanov formulation
$\sum_{m=1}^{M} h(d_m) \leq 1$,

- Correspondence via equivalent *block-norm formulations*.
- C and μ can be chosen *independently*.
- The constant 1 has no meaning.

Ivanov $\Rightarrow$ block-norm formulation 1 (known)

Let $h(d_m) = d_m^p$ (ℓ_p-norm MKL); see Kloft et al. (2010).

$$\operatorname*{minimize}_{\substack{f_1\in\mathcal{H}_1,\ldots,f_M\in\mathcal{H}_M,\\ b\in\mathbb{R},\\ d_1\geq 0,\ldots,d_M\geq 0}} \sum_{i=1}^{N} \ell\left(y_i, \textstyle\sum_{m=1}^{M} f_m(x_i) + b\right) + \frac{C}{2}\sum_{m=1}^{M} \frac{\|f_m\|_{\mathcal{H}_m}^2}{d_m},$$

$$\text{s.t.} \quad \sum_{m=1}^{M} d_m^p \leq 1.$$

$$\Downarrow \quad \text{Jensen's inequality}$$

$$\operatorname*{minimize}_{\substack{f_1\in\mathcal{H}_1,\ldots,f_M\in\mathcal{H}_M\\ ,bin\mathbb{R}}} \sum_{i=1}^{N} \ell\left(y_i, \textstyle\sum_{m=1}^{M} f_m(x_i) + b\right) + \frac{C}{2}\left(\textstyle\sum_{m=1}^{M} \|f_m\|_{\mathcal{H}_m}^q\right)^{2/q}.$$

where $q = 2p/(1+p)$. Minimum is attained at $d_m \propto \|f_m\|_{\mathcal{H}_m}^{2/(1+p)}$

Tikhonov ⇒ block-norm formulation 2 (new)

Let $h(d_m) = d_m^p$, $\mu = 1/p$ (ℓ_p-norm MKL)

$$\underset{\substack{f_1 \in \mathcal{H}_1, \ldots, f_M \in \mathcal{H}_M, \\ b \in \mathbb{R}, \\ d_1 \geq 0, \ldots, d_M \geq 0}}{\text{minimize}} \sum_{i=1}^{N} \ell\left(y_i, \textstyle\sum_{m=1}^{M} f_m(x_i) + b\right) + \frac{C}{2} \sum_{m=1}^{M} \left(\frac{\|f_m\|_{\mathcal{H}_m}^2}{d_m} + \frac{d_m^p}{p} \right).$$

⇓ Young's inequality

$$\underset{\substack{f_1 \in \mathcal{H}_1, \ldots, f_M \in \mathcal{H}_M, \\ b \in \mathbb{R}}}{\text{minimize}} \sum_{i=1}^{N} \ell\left(y_i, \textstyle\sum_{m=1}^{M} f_m(x_i) + b\right) + \frac{C}{q} \sum_{m=1}^{M} \|f_m\|_{\mathcal{H}_m}^q.$$

where $q = 2p/(1+p)$. Minimum is attained at $d_m = \|f_m\|_{\mathcal{H}_m}^{2/(1+p)}$.

The two block norm formulations are equivalent

Block norm formulation 1 (from Ivanov):

$$\underset{\substack{f_1\in\mathcal{H}_1,\dots,f_M\in\mathcal{H}_M\\ ,bin\mathbb{R}}}{\text{minimize}} \sum_{i=1}^{N} \ell\left(y_i, \textstyle\sum_{m=1}^{M} f_m(x_i) + b\right) + \frac{\tilde{C}}{2}\left(\textstyle\sum_{m=1}^{M} \|f_m\|_{\mathcal{H}_m}^{q}\right)^{2/q}.$$

Block norm formulation 2 (from Tikhonov):

$$\underset{\substack{f_1\in\mathcal{H}_1,\dots,f_M\in\mathcal{H}_M,\\ b\in\mathbb{R}}}{\text{minimize}} \sum_{i=1}^{N} \ell\left(y_i, \textstyle\sum_{m=1}^{M} f_m(x_i) + b\right) + \frac{C}{q}\sum_{m=1}^{M} \|f_m\|_{\mathcal{H}_m}^{q}.$$

- Just have to map C and $\tilde{C}$.
- The implied kernel weights are normalized/unnormalized.

Generalized block-norm formulation

$$\operatorname*{minimize}_{\substack{f_1 \in \mathcal{H}_1, \\ \ldots, f_M \in \mathcal{H}_M, \\ b \in \mathbb{R}}} \sum_{i=1}^{N} \ell\left(y_i, \textstyle\sum_{m=1}^{M} f_m(x_i) + b\right) + C \sum_{m=1}^{M} g(\|f_m\|_{\mathcal{H}_m}^2), \qquad (4)$$

where g is a concave block-norm-based regularizer.

Example (Elastic-net MKL): $g(x) = (1-\lambda)\sqrt{x} + \frac{\lambda}{2}x$,

$$\operatorname*{minimize}_{\substack{f_1 \in \mathcal{H}_1, \\ \ldots, f_M \in \mathcal{H}_M, \\ b \in \mathbb{R}}} \sum_{i=1}^{N} \ell\left(y_i, \textstyle\sum_{m=1}^{M} f_m(x_i) + b\right)$$

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

Generalized block-norm $\Rightarrow$ Tikhonov regularization

Theorem

Correspondence between the convex *(kernel-weight-based) regularizer* $h(d_m)$ *and the* concave *(block-norm-based) regularizer* $g(x)$ *is given as follows:*

$$\mu h(d_m) = -2g^* \left(\frac{1}{2d_m}\right),$$

where g^* *is the concave conjugate of* g*.*

Proof: Use the concavity of g as

$$\frac{\|f_m\|_{\mathcal{H}_m}^2}{2d_m} \geq g(\|f_m\|_{\mathcal{H}_m}^2) + g^*(1/(2d_m)).$$

See also Palmer et al. (2006).

Examples

Generalized Young's inequality:

$$xy \geq g(x) + g^*(y)$$

where g is concave, and g^* is the concave conjugate of g.

Example 1: let $g(x) = \sqrt{x}$, then $g^*(y) = -1/(4y)$ and

$$\frac{\|f_m\|_{\mathcal{H}_m}^2}{2d_m} + \frac{d_m}{2} \geq \|f_m\|_{\mathcal{H}_m} \qquad \text{(L1-MKL)}.$$

Example 2: let $g(x) = x^{q/2}/q$ $(1 \leq q \leq 2)$, then $g^*(y) = \frac{q-2}{2q}(2y)^{q/(q-2)}$

$$\frac{\|f_m\|_{\mathcal{H}_m}^2}{2d_m} + \frac{d_m^p}{2p} \geq \frac{1}{q}\|f_m\|_{\mathcal{H}_m}^q \qquad (\ell_p\text{-norm MKL}),$$

where $p := q/(2-q)$.

Correspondence

MKL model	block-norm $g(x)$	kern weight $h(d_m)$	reg const μ
block 1-norm MKL	$\sqrt{x}$	d_m	1
ℓ_p-norm MKL	$\frac{1+p}{2p} x^{p/(1+p)}$	d_m^p	$1/p$
Uniform-weight MKL (block 2-norm MKL)	$x/2$	$I_{[0,1]}(d_m)$	$+0$
block q-norm MKL ($q > 2$)	$\frac{1}{q} x^{q/2}$	$d_m^{-q/(q-2)}$	$-(q-2)/q$
Elastic-net MKL	$(1-\lambda)\sqrt{x} + \frac{\lambda}{2} x$	$\frac{(1-\lambda)d_m}{1-\lambda d_m}$	$1-\lambda$

$I_{[0,1]}(x)$ is the indicator function of the closed interval $[0,1]$; i.e., $I_{[0,1]}(x) = 0$ if $x \in [0,1]$, and $+\infty$ otherwise.

Bayesian view

Tikhonov regularization as a hierarchical MAP estimation

$$\underset{\substack{f_1\in\mathcal{H}_1,\\ \dots,f_M\in\mathcal{H}_M,\\ d_1\geq 0,\\ \dots,d_M\geq 0}}{\text{minimize}} \underbrace{\sum_{i=1}^{N}\ell\left(y_i,\textstyle\sum_{m=1}^{M} f_m(x_i)\right)}_{\text{likelihood}} + \underbrace{\sum_{m=1}^{M}\frac{\|f_m\|_{\mathcal{H}_m}^2}{2d_m}}_{f_m\text{-prior}} + \underbrace{\mu\sum_{m=1}^{M}h(d_m)}_{d_m\text{-hyper-prior}}.$$

Hyper prior over the kernel weights

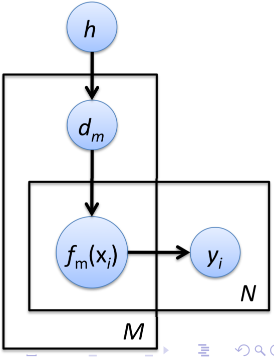
$$d_m \sim \frac{1}{Z_1(\mu)}\exp(-\mu h(d_m)) \qquad (m=1,\dots,M).$$

Gaussian process for the functions

$$f_m \sim GP(f_m;0,d_m k_m) \qquad (m=1,\dots,M).$$

Likelihood

$$y_i \sim \frac{1}{Z_2(x_i)}\exp(-\ell(y_i,\textstyle\sum_{m=1}^{M} f_m(x_i))).$$



Marginalized likelihood

Assume Gaussian likelihood

$$\ell(y,z) = \frac{1}{2\sigma_y^2}(y-z)^2.$$

The marginalized likelihood (omitting hyper-prior for simplicity)

$$-\log p(\boldsymbol{y}|\boldsymbol{d})$$

$$= \underbrace{\frac{1}{2\sigma_y^2}\left\|\boldsymbol{y} - \sum\nolimits_{m=1}^{M} f_m^{\mathrm{MAP}}\right\|^2}_{\text{likelihood}} + \underbrace{\frac{1}{2}\sum_{m=1}^{M}\frac{\|f_m^{\mathrm{MAP}}\|_{\mathcal{H}_m}^2}{d_m}}_{f_m\text{-prior}} + \underbrace{\frac{1}{2}\log\left|\bar{\boldsymbol{K}}(\boldsymbol{d})\right|}_{\text{volume-based regularization}}.$$

- f_m^{MAP}: MAP estimate for a fixed kernel weights d_m $(m = 1,\ldots,M)$.
- $\bar{\boldsymbol{K}}(\boldsymbol{d}) := \sigma_y^2 \boldsymbol{I}_N + \sum_{m=1}^{M} d_m \boldsymbol{K}_m$.

See also Wipf & Nagarajan (2009).

Comparing MAP and empirical Bayes objectives

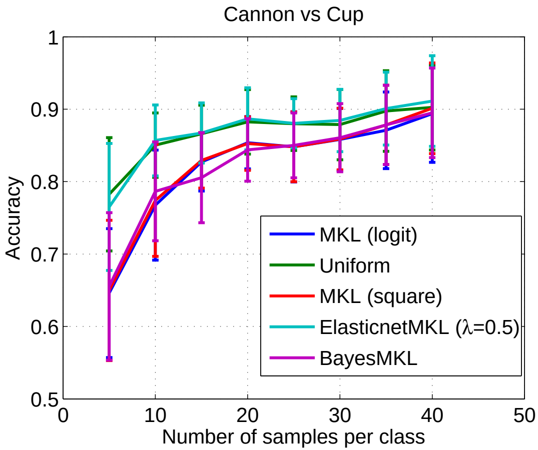
Hyper-prior MAP (MKL):

$$\underbrace{\sum_{i=1}^{N} \ell\left(y_i, \textstyle\sum_{m=1}^{M} f_m(x_i)\right)}_{\text{likelihood}} + \underbrace{\frac{1}{2}\sum_{m=1}^{M} \frac{\|f_m\|_{\mathcal{H}_m}^2}{d_m}}_{f_m\text{-prior}} + \underbrace{\mu \sum_{m=1}^{M} h(d_m)}_{\substack{d_m\text{-hyper-prior} \\ \text{(separable)}}}.$$

Empirical Bayes:

$$\underbrace{\frac{1}{2\sigma_y^2}\left\|\boldsymbol{y} - \textstyle\sum_{m=1}^{M} f_m^{\mathrm{MAP}}\right\|^2}_{\text{likelihood}} + \underbrace{\frac{1}{2}\sum_{m=1}^{M} \frac{\|f_m^{\mathrm{MAP}}\|_{\mathcal{H}_m}^2}{d_m}}_{f_m\text{-prior}} + \underbrace{\frac{1}{2}\log\left|\bar{\boldsymbol{K}}(\boldsymbol{d})\right|}_{\substack{\text{volume-based} \\ \text{regularization} \\ \text{(non-separable)}}}.$$

Caltech 101 dataset (classification)

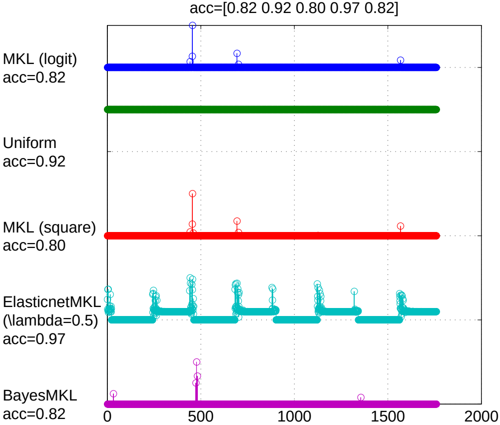


- Regularization constant C chosen by 2×4-fold cross validation on the training-set.

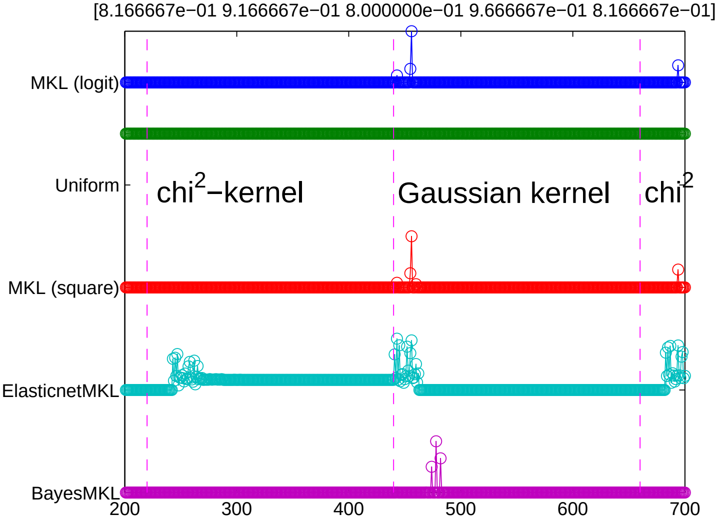
Caltech 101 dataset: kernel weights

1,760 kernel functions.

- 4 SIFT features (hsvsift, sift, sift4px, sift8px)
- 22 spacial decompositions (including spatial pyramid kernel)
- 2 kernel functions (Gaussian and χ^2)
- 10 kernel parameters



Caltech 101 dataset: kernel weights (detail)



Summary

- Two regularized kernel weight learning formulations
  - Ivanov regularization.
  - Tikhonov regularization.

  are equivalent. No additional tuning parameter!
- Both formulations reduce to block-norm formulations via Jensen's inequality / (generalized) Young's inequality.
- Probabilistic view of MKL: hierarchical Gaussian process model.
- Elastic-net MKL performs similarly to uniform weight MKL, but shows grouping of mutually depended kernels.
- Empirical-Bayes MKL and L1-MKL seem to make the solution overly sparse, but often they choose slightly different set of kernels.
- Code for Elastic-net-MKL available from

  `http://www.simplex.t.u-tokyo.ac.jp/~s-taiji/software/SpicyMKL`

Acknowledgements

We would like to thank Hisashi Kashima and Shinichi Nakajima for helpful discussions. This work was supported in part by MEXT KAKENHI 22700138, 22700289, and NTT Communication Science Laboratories.

A brief proof

- Minimize the Lagrangian:

$$\min_{\substack{f_1 \in \mathcal{H}_1, \\ \dots, f_M \in \mathcal{H}_M}} \frac{1}{2} \sum_{m=1}^{M} \frac{\|f_m\|_{\mathcal{H}_m}^2}{d_m} + \left\langle g, \underbrace{\bar{f} - \sum\nolimits_{m=1}^{M} f_m}_{\text{equality const.}} \right\rangle_{\mathcal{H}(\boldsymbol{d})},$$

where $g \in \mathcal{H}(\boldsymbol{d})$ is a Lagrangian multiplier.

- Fréchet derivative

$$\left\langle h_m, \frac{f_m}{d_m} - \langle g, k_m \rangle_{\mathcal{H}(\boldsymbol{d})} \right\rangle_{\mathcal{H}_m} = 0 \;\Rightarrow\; f_m(x) = \langle g, d_m k_m(\cdot, x) \rangle_{\mathcal{H}(\boldsymbol{d})}.$$

- Maximize the dual

$$\max_{g \in \mathcal{H}(\boldsymbol{d})} -\frac{1}{2}\|g\|_{\mathcal{H}(\boldsymbol{d})}^2 + \langle g, \bar{f} \rangle_{\mathcal{H}(\boldsymbol{d})} = \frac{1}{2}\|\bar{f}\|_{\mathcal{H}(\boldsymbol{d})}^2$$

References

- Aronszajn. Theory of Reproducing Kernels. TAMS, 1950.
- Lanckriet et al. Learning the Kernel Matrix with Semidefinite Programming. JMLR, 2004.
- Bach et al. Multiple kernel learning, conic duality, and the SMO algorithm. ICML 2004.
- Micchelli & Pontil. Learning the kernel function via regularization. JMLR, 2005.
- Cortes. Can learning kernels help performance? ICML, 2009.
- Cortes et al. Generalization Bounds for Learning Kernels. ICML, 2010.
- Kloft et al. Efficient and accurate lp-norm multiple kernel learning. NIPS 22, 2010.
- Tomioka & Suzuki. Sparsity-accuracy trade-off in MKL. arxiv, 2010.
- Varma & Babu. More Generality in Efficient Multiple Kernel Learning. ICML, 2009.
- Gehler & Nowozin. Let the kernel figure it out; principled learning of pre-processing for kernel classifiers. CVPR, 2009.
- Tipping. Sparse bayesian learning and the relevance vector machine. JMLR, 2001.
- Palmer et al. Variational EM Algorithms for Non-Gaussian Latent Variable Models. NIPS, 2006.
- Wipf & Nagarajan. A new view of automatic relevance determination. NIPS, 2008.

Method A: upper-bounding the log det term

- Use the upper bound

$$\log \left|\bar{\boldsymbol{K}}(\boldsymbol{d})\right| \leq \sum_{m=1}^{M} z_m d_m - \psi^*(\boldsymbol{z})$$

- Eliminate the kernels weights by explicit minimization (AGM ineq.)

Update f_m as

$$(\boldsymbol{f}_m)_{m=1}^M \leftarrow \operatorname*{argmin}_{(\boldsymbol{f}_m)_{m=1}^M} \left(\frac{1}{2\sigma_y^2} \left\| \boldsymbol{y} - \sum_{m=1}^{M} \boldsymbol{f}_m \right\|^2 + \sum_{m=1}^{M} \sqrt{z_m} \|\boldsymbol{f}_m\|_{\boldsymbol{K}_m}$$

Update z_m as (tighten the upper bound)

$$z_m \leftarrow \mathrm{Tr}\left((\sigma_y^2 \boldsymbol{I}_N + \sum_{m=1}^{M} d_m \boldsymbol{K}_m)^{-1} \boldsymbol{K}_m \right),$$

where $d_m = \|f_m\|_{\mathcal{H}_m} / \sqrt{z_m}$.

- Each update step is a *reweighted L1-MKL problem*.
- Each update step minimizes an upper bound of the

Method B: MacKay update

- Use the fixed point condition for the update of the weights:

$$-\frac{\|\boldsymbol{f}_m^{\mathrm{FKL}}\|_{\boldsymbol{K}_m}^2}{d_m^2} + \mathrm{Tr}\left((\sigma^2 \boldsymbol{I}_N + \sum_{m=1}^{M} d_m \boldsymbol{K}_m)^{-1} \boldsymbol{K}_m\right) = 0.$$

Update f_m as

$$(\boldsymbol{f}_m)_{m=1}^M \leftarrow \underset{(f_m)_{m=1}^M}{\mathrm{argmin}} \left(\frac{1}{2\sigma_y^2} \left\| \boldsymbol{y} - \sum_{m=1}^{M} \boldsymbol{f}_m \right\|^2 + \frac{1}{2} \sum_{m=1}^{M} \frac{\|\boldsymbol{f}_m\|_{\boldsymbol{K}_m}^2}{d_m} \right)$$

Update the kernel weights d_m as

$$d_m \leftarrow \frac{\|\boldsymbol{f}_m\|_{\boldsymbol{K}_m}^2}{\mathrm{Tr}\left((\sigma^2 \boldsymbol{I}_N + \sum_{m=1}^{M} d_m \boldsymbol{K}_m)^{-1} d_m \boldsymbol{K}_m\right)}.$$

- Each update step is a *fixed kernel weight leraning problem* (easy).
- Convergence empirically OK (e.g., RVM).