SpicyMKL
Efficient multiple kernel learning method using dual augmented Lagrangian

Taiji Suzuki
Ryota Tomioka

The University of Tokyo
Graduate School of Information Science and Technology
Department of Mathematical Informatics

23rd Oct., 2009@ISM

Multiple Kernel Learning

Fast Algorithm

SpicyMKL

Kernel Learning

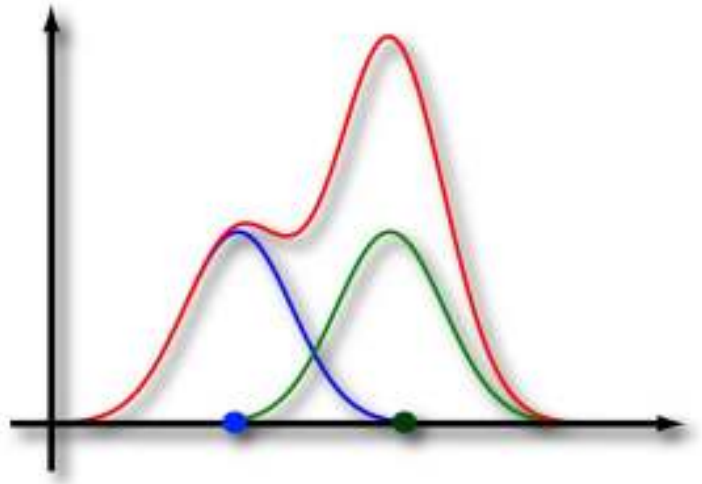
$$x \xrightarrow{f(x)} y$$

Regression, Classification

$$f(x) = \sum_{i=1}^{N} \alpha_i k(x_i, x)$$

Kernel Function

$$k(x, x')$$



- Gaussian, polynomial, chi-square, ….
  - Parameters: Gaussian width, polynomial degree

Thousands of kernels

- Features
  - Computer Vision: color, gradient, sift (sift, hsvsift, huesift, scaling of sift), Geometric Blur, image regions, . . .

Multiple Kernel Learning

(Lanckriet et al. 2004)

Select important kernels and combine them

Multiple Kernel Learning (MKL)

Single Kernel

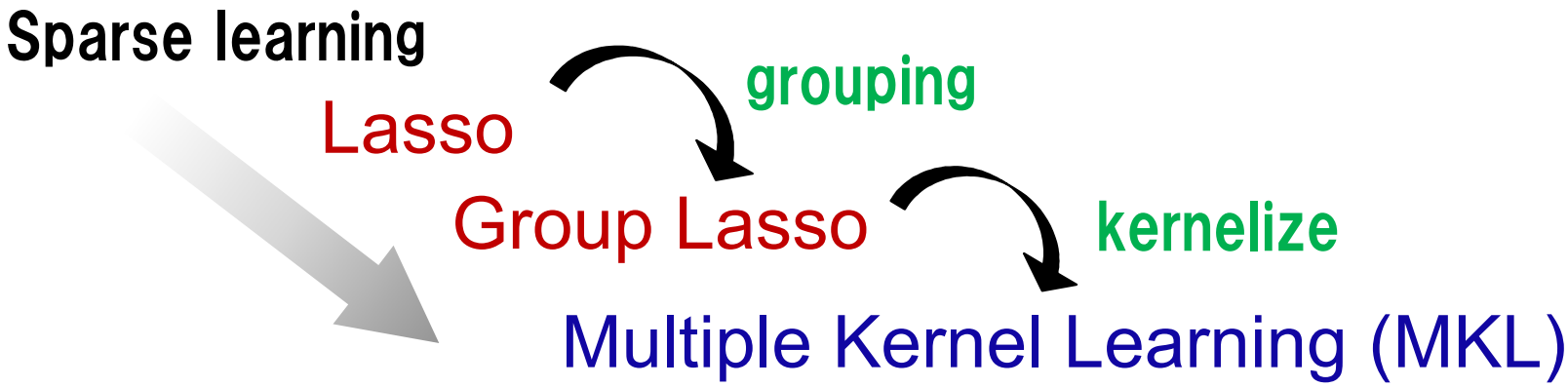
$$f(x) = \sum_{i=1}^{N} \alpha_i k(x_i, x)$$

Multiple Kernel

$$f(x) = \sum_{i=1}^{N} \alpha_i \left(\sum_{m=1}^{M} d_m k_m(x_i, x) \right)$$

d_m:convex combination,

sparse (many 0 components)



$$\underset{f_m \in \mathcal{H}_m}{\text{minimize}} \quad \sum_{i=1}^{N} \ell\left(y_i, \sum_{m=1}^{M} f_m(x_i)\right) + C \sum_{m=1}^{M} \|f_m\|_{\mathcal{H}_m}$$

L_1-regularization
→**sparse**

$\{(x_i, y_i)\}_{i=1}^{N}$: N samples

$\{k_m(x, x')\}_{m=1}^{M}$: M kernels

$\mathcal{H}_m$: RKHS associated with k_m

$\ell(y, f)$: Convex loss (hinge, square, logistic)

$$\underset{f_m \in \mathcal{H}_m}{\text{minimize}} \quad \sum_{i=1}^{N} \ell\left(y_i, \sum_{m=1}^{M} f_m(x_i)\right) + C \sum_{m=1}^{M} \|f_m\|_{\mathcal{H}_m}$$

Representer theorem

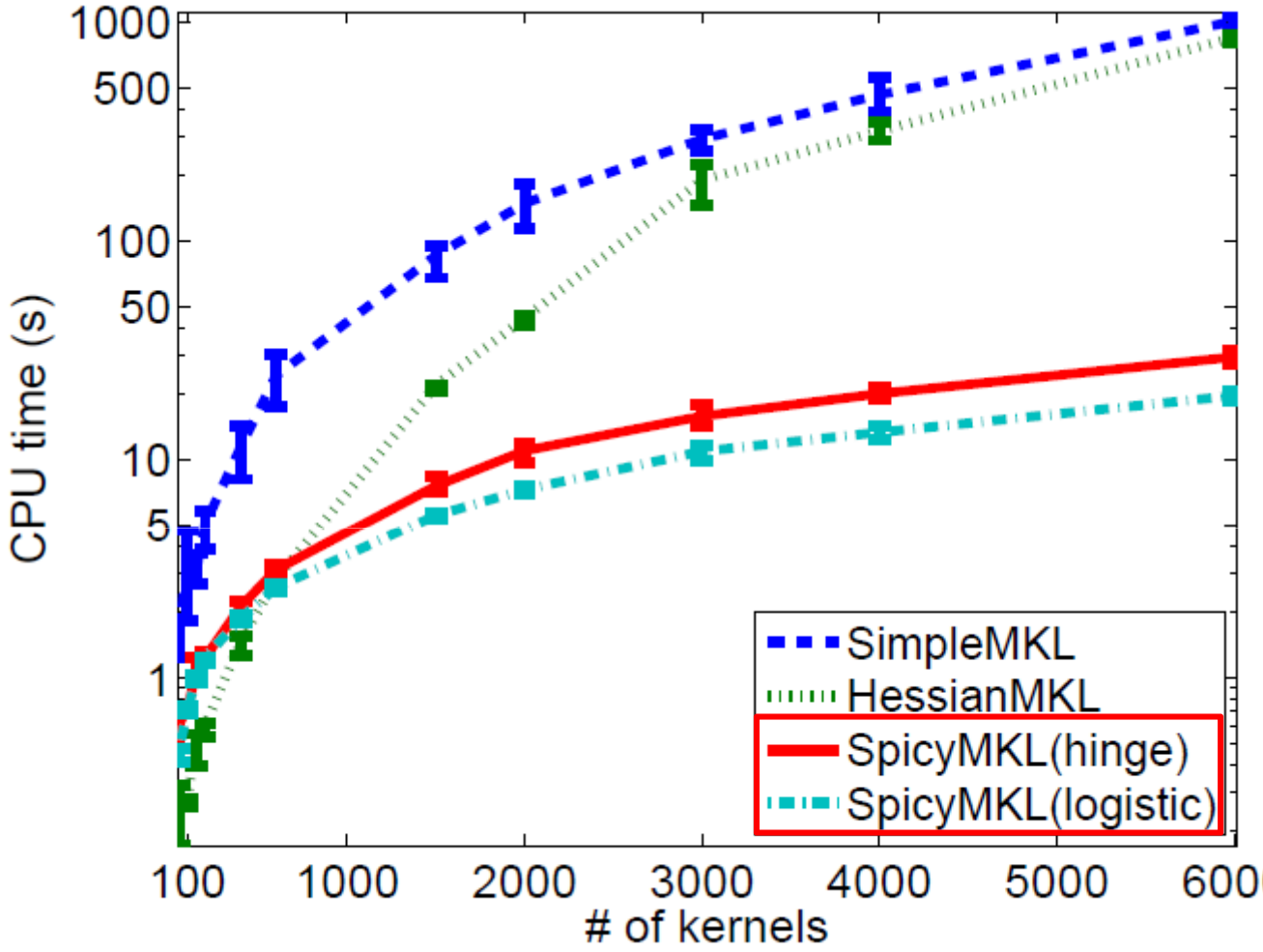
$$\underset{\alpha_m \in \mathrm{R}^N}{\text{minimize}} \quad \sum_{i=1}^{N} \ell\left(y_i, \sum_{m=1}^{M} (K_m \alpha_m)_i\right) + C \sum_{m=1}^{M} \|\alpha_m\|_m$$

Finite dimensional problem

$(K_m)_{i,j} := k_m(x_i, x_j)$: Gram matrix of mth kernel

$\|\alpha_m\|_m := \sqrt{\alpha_m^T K_m \alpha_m}$

SpicyMKL



Proposed: SpicyMKL

- Scales well against # of kernels.
- Formulated in a general framework.

Outline

- Introduction
- Details of Multiple Kernel Learning
- Our method
  - Proximal minimization
  - Skipping inactive kernels
- Numerical Experiments
  - Bench-mark datasets
  - Sparse or dense

Details of Multiple Kernel Learning (MKL)

Generalized Formulation

MKL

$$\underset{\alpha_m \in \mathrm{R}^N}{\text{minimize}} \quad \sum_{i=1}^{N} \ell\left(y_i, \sum_{m=1}^{M} (K_m \alpha_m)_i\right) + C \sum_{m=1}^{M} \|\alpha_m\|_m$$

$$(K_m)_{i,j} := k_m(x_i, x_j) \qquad \|\alpha_m\|_m := \sqrt{\alpha_m^T K_m \alpha_m}$$

$$\underset{\alpha_m \in \mathrm{R}^N}{\text{minimize}} \quad f_\ell\left(K\alpha\right) + \sum_{m=1}^{M} g(\|\alpha_m\|_m)$$

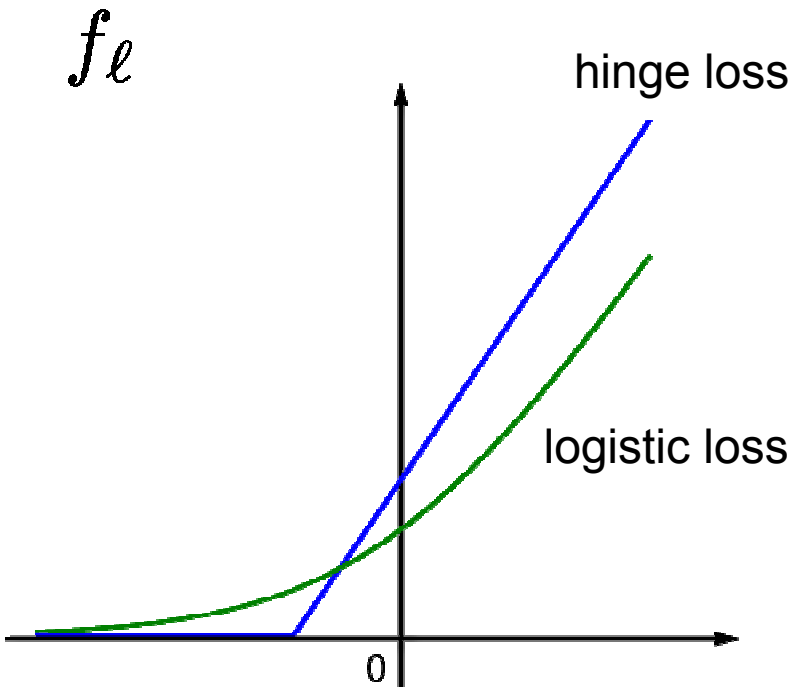
$K = (K_1, \ldots, K_m)$

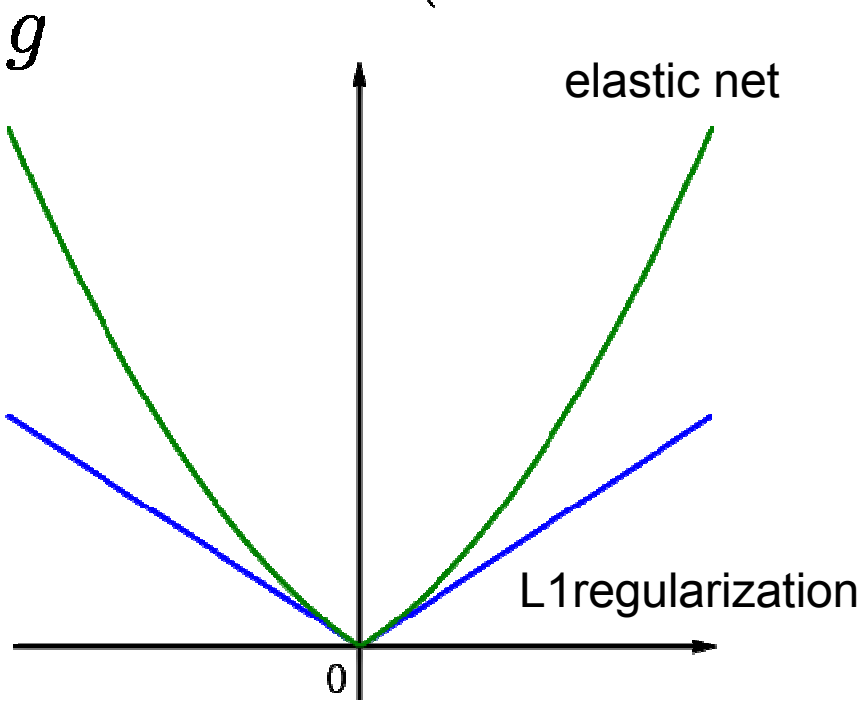
$\alpha^\top = (\alpha_1^\top, \ldots, \alpha_m^\top)$

g : sparse regularization

$$\underset{\alpha_m \in \mathrm{R}^N}{\text{minimize}} \quad f_\ell \left(\sum_{m=1}^{M} K_m \alpha_m \right) + \sum_{m=1}^{M} g(\|\alpha_m\|_m)$$

$$\left(\|\alpha_m\|_m := \sqrt{\alpha_m^T K_m \alpha_m} \right)$$





hinge: $\ell(y, f) = \max(1 - yf, 0)$

logistic: $\ell(y, f) = \log(1 + \exp(-yf))$

elastic net: $g(x) = C((1-\lambda)|x| + \frac{\lambda}{2}x^2)$

L1regularization: $g(x) = C|x|$

Rank 1 decomposition

If g is monotonically increasing and f_ℓ is diff'ble at the optimum, the solution of MKL is rank 1:

$\exists d_m \in \mathbb{R}, \exists \beta \in \mathbb{R}^N$ such that

$$\alpha_m^* = d_m \beta \quad (\forall m) \qquad \sum_{m=1}^{M} d_m = 1, \quad d_m \geq 0$$

$$f^*(x) = \sum_{i=1}^{N} \beta_i \left(\sum_{m=1}^{M} d_m k_m(x_i, x) \right)$$

convex combination of kernels

Proof Derivative w.r.t. α_m

$$K_m \nabla f_\ell(K\alpha^*) + \partial g(\|\alpha_m^*\|_m) \frac{K_m}{\|\alpha_m^*\|_m} \alpha_m^* = 0$$

$$\Rightarrow \quad \alpha_m^* = \frac{\|\alpha_m^*\|_m}{\partial g(\|\alpha_m^*\|_m)} (-\nabla f_\ell(K\alpha^*))$$

Existing approach 1

Constraints based formulation:

[Lanckriet et al. 2004, JMLR] [Bach, Lanckriet & Jordan 2004, ICML]

primal

$$\min_{\alpha} \; f_\ell(K\alpha) + \frac{1}{2}\left(\sum_{m=1}^{M} \|\alpha_m\|_m\right)^2$$

Dual

$$\max_{\rho} \; -f_\ell^*(-\rho) + \frac{1}{2}\left(\max_{m} \|\rho\|_m\right)^2$$

$$\max_{\rho, r} \; -f_\ell^*(-\rho) + \frac{1}{2}r^2$$
$$\text{s.t.} \quad \|\rho\|_m \le r \;\; (\forall m)$$

- Lanckriet et al. 2004: SDP
- Bach, Lanckriet & Jordan 2004: SOCP

Existing approach 2

Upper bound based formulation:

[Sonnenburg et al. 2006, JMLR] [Rakotomamonjy et al. 2008, JMLR]
[Chapelle & Rakotomamonjy 2008, NIPS workshop]

primal

$$\min_{\alpha} \; f_\ell(K\alpha) + \frac{1}{2}\left(\sum_{m=1}^{M} \|\alpha_m\|_m\right)^2$$

$$\min_{\alpha,d} \; f_\ell(K\alpha) + \frac{1}{2}\sum_{m=1}^{M} \frac{\alpha_m^\top K_m \alpha_m}{d_m} \qquad \text{(Jensen's inequality)}$$

$$\text{s.t.} \quad \sum_{m=1}^{M} d_m = 1, \; d_m \geq 0$$

$$\min_{d} \; \min_{\beta} \left\{ f_\ell(K(d)\beta) + \frac{1}{2}\beta^\top K(d)\beta \right\} \rightarrow \boxed{\theta(d)}$$

$$\text{s.t.} \quad \sum_{m=1}^{M} d_m = 1, \; d_m \geq 0, \; K(d) = \sum_{m=1}^{M} d_m K_m.$$

- Sonnenburg et al. 2006: Cutting plane (SILP)
- Rakotomamonjy et al. 2008: Gradient descent (SimpleMKL)
- Chapelle & Rakotomamonjy 2008: Newton method (HessianMKL)

Problem of existing methods

- Do not make use of sparsity
  during the optimization.
  → Do not perform efficiently when
  # of kernels is large.

Outline

- Introduction
- Details of Multiple Kernel Learning
- Our method
  - Proximal minimization
  - Skipping inactive kernels
- Numerical Experiments
  - Bench-mark datasets
  - Sparse or dense

Our method
SpicyMKL

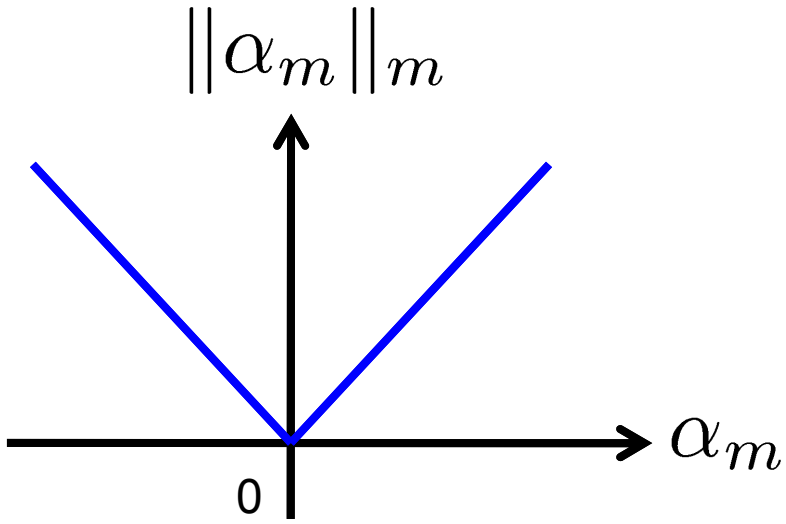
- DAL (Tomioka & Sugiyama 09)
  - Dual Augmented Lagrangian
  - Lasso, group Lasso, trace norm regularization

- SpicyMKL = DAL + MKL
  - Kernelization of DAL

Difficulty of MKL(Lasso)

$\|\alpha_m\|_m$

is non-diff'ble at 0.



Relax by

Proximal Minimization

(Rockafellar, 1976)

Our algorithm

SpicyMKL: Proximal Minimization (Rockafellar, 1976)

1. Set some initial solution $\alpha^{(0)}$. Choose increasing sequence $\gamma_1 \leq \gamma_2 \leq \gamma_3 \cdots$.
2. Iterate the following update rule until convergence:

$$\alpha^{(t+1)} \leftarrow \underset{\alpha \in \mathrm{R}^{MN}}{\operatorname{argmin}} \quad f_\ell(K\alpha) + \sum_{m=1}^{M} g(\|\alpha_m\|_m) + \frac{1}{2\gamma_t} \sum_{m=1}^{M} \|\alpha_m - \alpha_m^{(t)}\|_m^2$$

Regularization toward the last update.

It can be shown that

- $f_\ell\left(K\alpha^{(t+1)}\right) + \sum_{m=1}^{M} g(\|\alpha_m^{(t+1)}\|_m) < f_\ell\left(K\alpha^{(t)}\right) + \sum_{m=1}^{M} g(\|\alpha_m^{(t)}\|_m)$
- $\alpha^{(t)} \to \alpha^*$ **super linearly !**

Taking its **dual**, the update step can be **solved efficiently**

Fenchel's duality theorem

$$\min_x \{f(Ax) + h(x)\} = \max_y \{-f^*(-y) - h^*(A^\top y)\}$$

$$\min_{\alpha \in \mathrm{R}^{NM}} \quad f_\ell(K\alpha) + \sum_{m=1}^{M} g(\|\alpha_m\|_m) + \frac{1}{2\gamma_t} \sum_{m=1}^{M} \|\alpha_m - \alpha_m^{(t)}\|_m^2$$

$f_\ell(K\alpha)$ $\qquad$ $\phi_t(\alpha)$

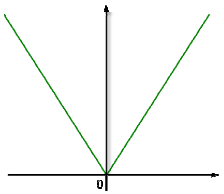
Split into two parts

$$\max_{\rho \in \mathrm{R}^N} \quad -f_\ell^*(-\rho) - \phi_t^*(K^\top \rho)$$

$$\max_{\rho \in \mathrm{R}^N} \; - f_\ell^*(-\rho) - \phi_t^*(K^\top \rho)$$

$$\phi_t(\alpha) = \sum_{m=1}^{M} g(\|\alpha_m\|_m) + \frac{1}{2\gamma_t}\|\alpha_m - \alpha_m^{(t)}\|_m^2$$

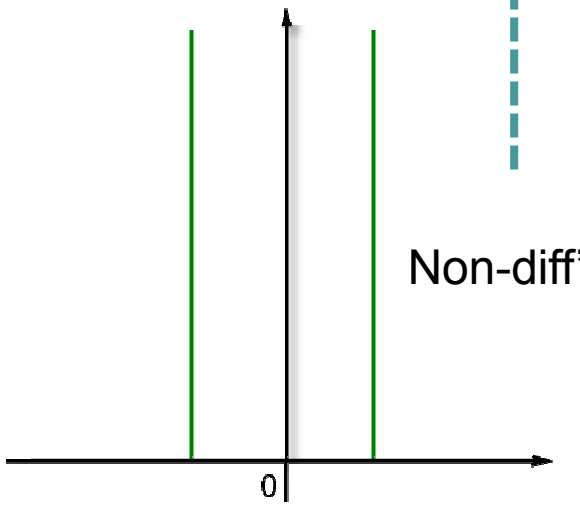
Non-diff'ble



g

dual

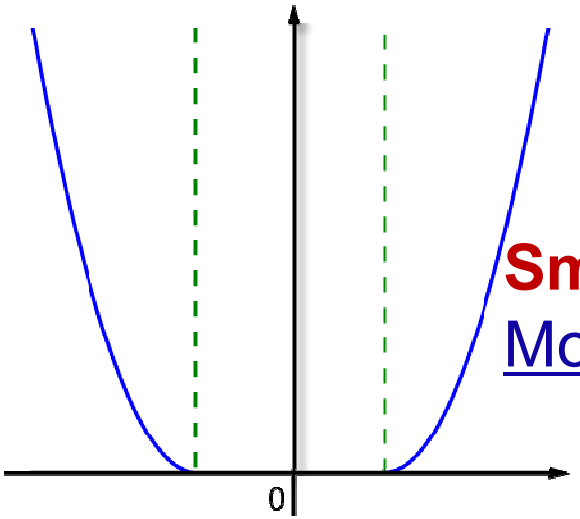
g^*



Non-diff'ble

dual

$\mathcal{M}g^*$



Smooth !

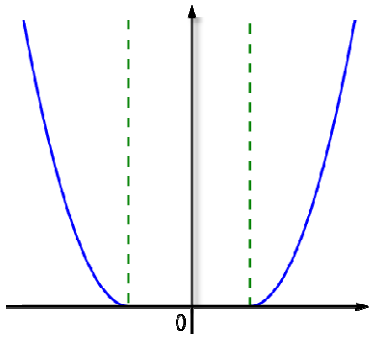
Moreau envelope

Inner optimization

- Newton method is applicable.

$$\max_{\rho \in \mathrm{R}^N} \quad -f_\ell^*(-\rho) - \phi_t^*(K^\top \rho)$$

Twice Differentiable Twice Differentiable (a.e.)



Even if f_ℓ^* is not differentiable,
we can apply almost the same algorithm.

Rigorous Derivation of ϕ_t^*

$$\max_{\rho \in \mathrm{R}^N} \quad -f_\ell^*(-\rho) - \phi_t^*(K^\top \rho)$$

Convolution and duality $(h_1 \square h_2)(z) := \inf_{z=x+y} \{h_1(x) + h_2(y)\}$

$$(f+g)^* = f^* \square g^*$$

$$\phi_t^*(K^\top \rho) = \sum_{m=1}^{K} (g(\cdot) + \frac{1}{2\gamma_t} \| \cdot - \alpha_m^{(t)} \|_m^2)^*(K_m \rho)$$

$$= \sum_{m=1}^{M} \min_{u_m \in \mathrm{R}^N} \left(g^*(\|u_m\|_m) + \frac{\gamma_t}{2} \left\| u_m - \rho - \frac{\alpha_m^{(t)}}{\gamma_t} \right\|_m^2 - \frac{1}{2\gamma_t} \|\alpha_m^{(t)}\|_m^2 \right)$$

$$= \sum_{m=1}^{M} \gamma_t \min_{u_m \in \mathrm{R}^N} \left(\frac{g^*(\|u_m\|_m)}{\gamma_t} + \frac{1}{2} \left\| u_m - \rho - \frac{\alpha_m^{(t)}}{\gamma_t} \right\|_m^2 \right) + (\text{const.})$$

Moreau envelope

$$\phi_t^* \Rightarrow \min_{u_m \in \mathrm{R}^N} \left(\frac{g^*(\|u_m\|_m)}{\gamma_t} + \frac{1}{2} \left\| u_m - \rho - \frac{\alpha_m^{(t)}}{\gamma_t} \right\|_m^2 \right)$$

$$= \min_{q_m \in \mathrm{R}} \left(\frac{g^*(q_m)}{\gamma_t} + \frac{1}{2}(q_m - \|\rho + \frac{\alpha_m^{(t)}}{\gamma_t}\|_m)^2 \right)$$

$$u_m = q_m \frac{\rho + \frac{\alpha_m^{(t)}}{\gamma_t}}{\|\rho + \frac{\alpha_m^{(t)}}{\gamma_t}\|_m}$$

Moreau envelope (Moreau 65)

For a convex function ϕ, we define its **Moreau envelope** as

$$\mathcal{M}\phi(z) := \min_{x \in \mathrm{R}} \left(\frac{1}{2}(x - z)^2 + \phi(x) \right)$$

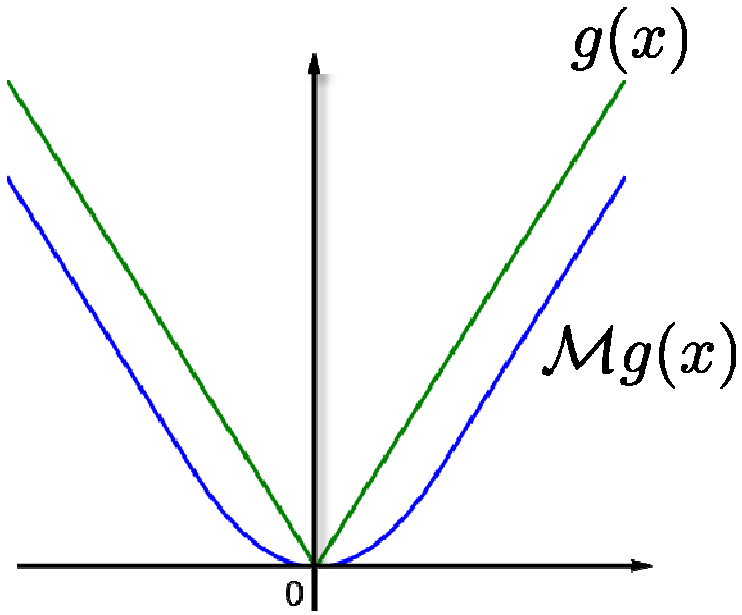
$$= \mathcal{M} \frac{g^*}{\gamma_t} \left(\|\rho + \frac{\alpha_m^{(t)}}{\gamma_t}\|_m \right)$$

Moreau envelope

$$\mathcal{M}g(z) := \min_{x \in \mathrm{R}} \left(\frac{1}{2}(x-z)^2 + g(x) \right)$$

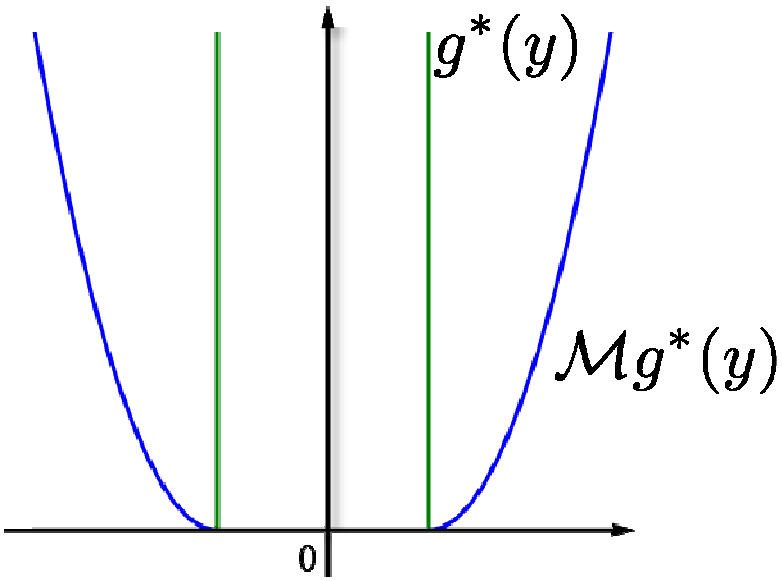
L_1-penalty

$$g(x) = |x|$$



dual

$$g^*(y) = \begin{cases} 0 & (|y| \le 1) \\ \infty & (\text{otherwise}) \end{cases}$$



Smooth !

What we have observed

We arrived at …

Dual of update rule in proximal minimization

$$\max_{\rho \in \mathrm{R}^N} \quad -f_\ell^*(-\rho) - \phi_t^*(K^\top \rho)$$

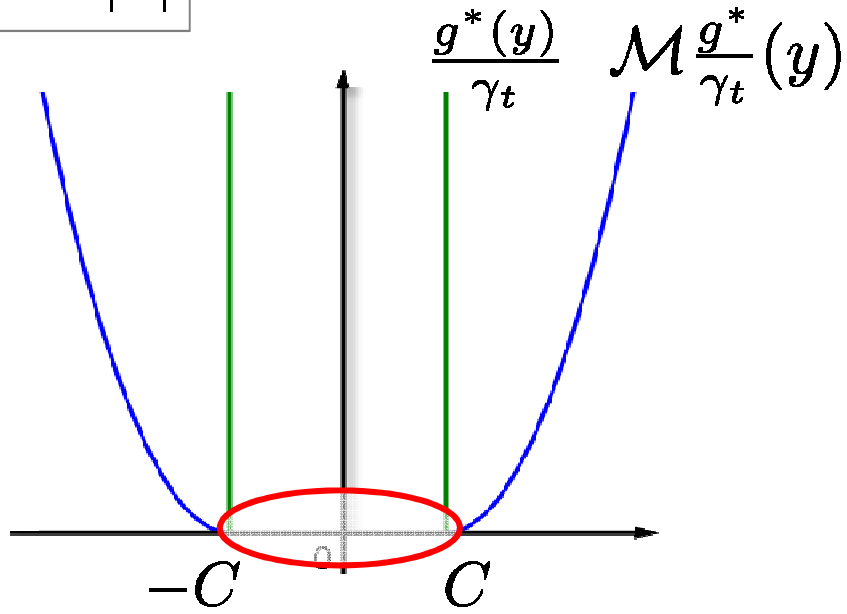
$$\max_{\rho \in \mathrm{R}^N} -f_\ell^*(-\rho) - \gamma_t \sum_{m=1}^{M} \mathcal{M}_{\frac{g^*}{\gamma_t}}\left(\left\|\rho + \frac{\alpha_m^{(t)}}{\gamma_t}\right\|_m\right)$$

- Differentiable
- Only active kernels need to be calculated.
  (next slide)

Skipping inactive kernels

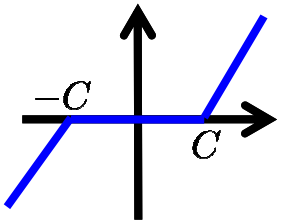
$$\max_{\rho \in \mathbb{R}^N} -f_\ell^*(-\rho) - \gamma_t \sum_{m=1}^{M} \mathcal{M}_{\frac{g^*}{\gamma_t}}\left(\|\rho + \frac{\alpha_m^{(t)}}{\gamma_t}\|_m\right)$$

$g(x) = C|x|$

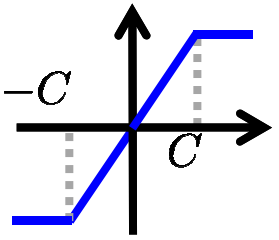


Soft threshold

$\mathrm{prox}_g(x)$



$\mathrm{prox}_{g^*}(x)$



Hard threshold

If $\|\rho + \frac{\alpha_m^{(t)}}{\gamma_t}\|_m \leq C$ (inactive), $\mathcal{M}_{\frac{g^*}{\gamma_t}}\left(\|\rho + \frac{\alpha_m^{(t)}}{\gamma_t}\|_m\right) = 0$,
and its gradient also vanishes.

Derivatives

• **We can apply Newton method for the dual optimization.**

The objective function:

$$\psi(\rho) = -f_\ell^*(-\rho) - \gamma_t \sum_{m=1}^{M} \mathcal{M}\frac{g^*}{\gamma_t}\left(\|\rho + \frac{\alpha_m^{(t)}}{\gamma_t}\|_m\right)$$

Its derivatives:

$$\nabla_\rho \psi(\rho) = -\nabla_\rho^2 f_\ell^*(-\rho) - \gamma_t \sum_{m:\text{active}} q_m \frac{K_m p_m}{\|p_m\|_m}$$

$$\nabla_\rho^2 \psi(\rho) = -\nabla_\rho^2 f_\ell^*(-\rho) - \gamma_t \sum_{m:\text{active}} \left\{ (r_m - q_m \|p_m\|_m) \frac{K_m p_m p_m^\top K_m}{\|p_m\|_m^2} + \gamma_t q_m \frac{K_m}{\|p_m\|_m} \right\}$$

where $p_m := \|\rho + \frac{\alpha_m^{(t)}}{\gamma_t}\|_m$, $q_m := \mathrm{prox}_{\gamma_t g}(\|\alpha_m^{(t)} + \gamma_t \rho\|_m)$, $r_m := \left.\frac{\mathrm{d}\mathrm{prox}_{\gamma_t g}(x)}{\mathrm{d}x}\right|_{x=\|\alpha_m^{(t)}+\gamma_t\rho\|_m}$.

• **Only active kernels contributes their computations.**
→ **Efficient for large number of kernels.**

Convergence result

$$\varphi(\alpha) = f_\ell(K\alpha) + \sum_{m=1}^{M} g(\|\alpha_m\|_m)$$

Thm.

For any proper convex functions f_ℓ and g , and any non-decreasing sequence γ_t ,

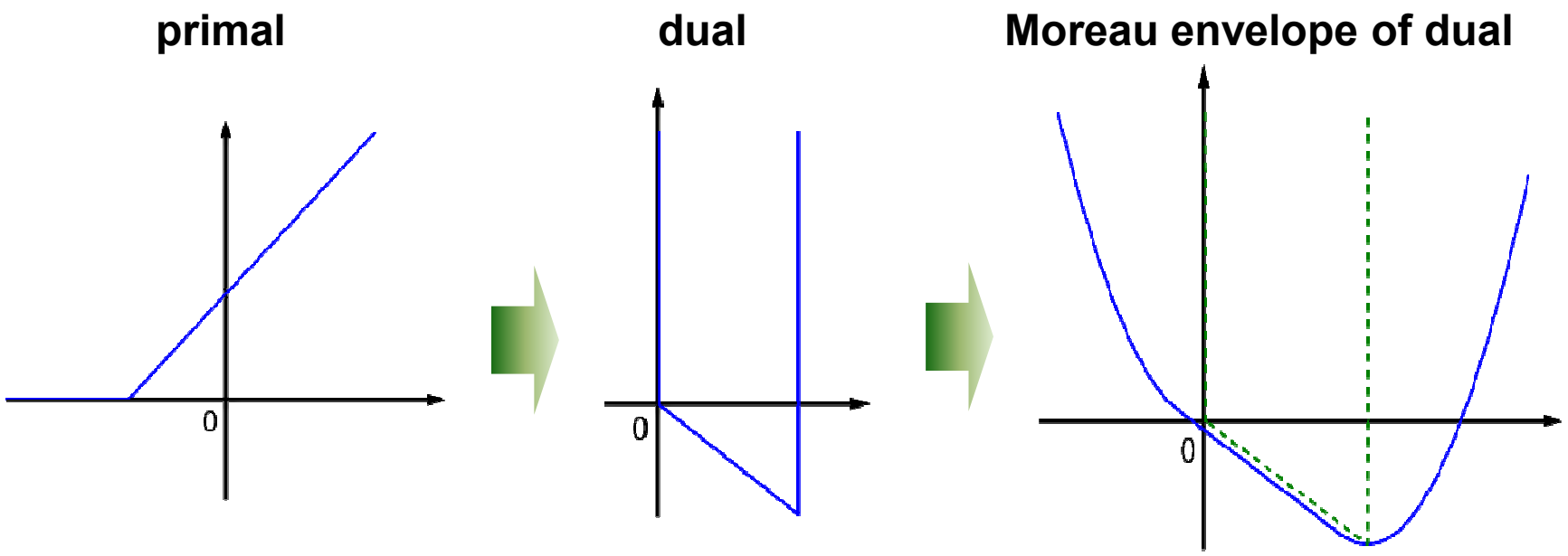
$$\mathrm{dist}\left(\alpha^{(t)}, \operatorname*{argmin}_{\alpha}(\varphi(\alpha))\right) \to 0 \quad (\text{as } t \to \infty).$$

Thm.

If f_ℓ is strongly convex and g is sparse reg., for any increasing sequence γ_t , the convergence rate is **super linear**.

Non-differentiable loss

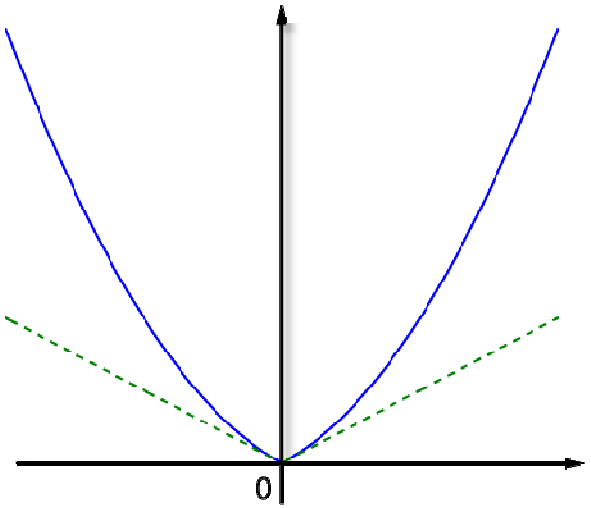
- If loss f_ℓ^* is non-differentiable, almost the same algorithm is available by utilizing Moreau envelope of f_ℓ^*.



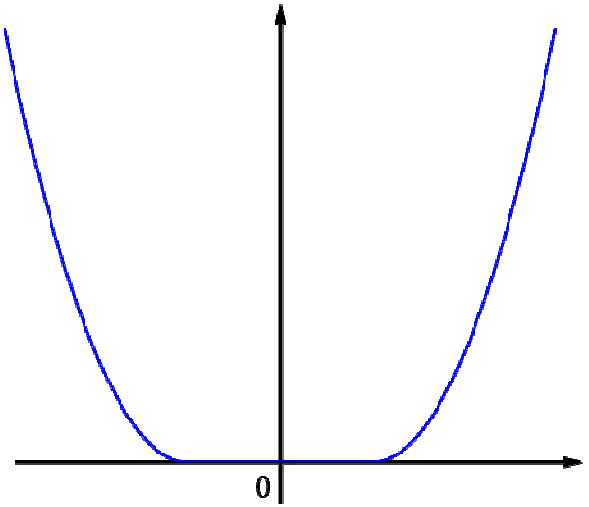
Dual of elastic net

$$g(x) = C((1-\lambda)|x| + \frac{\lambda}{2}x^2) \qquad g^*(x) = \frac{\max(|y| - C(1-\lambda), 0)^2}{2C\lambda}$$

primal



dual



Smooth !

Naïve optimization in the dual is applicable.
But we applied prox. minimization method for small λ to avoid numerical instability.

Why ‘Spicy’ ?

- SpicyMKL = DAL + MKL.

- DAL also means a major Indian cuisine.

  → hot, spicy

  → **SpicyMKL**

“Dal”

from Wikipedia

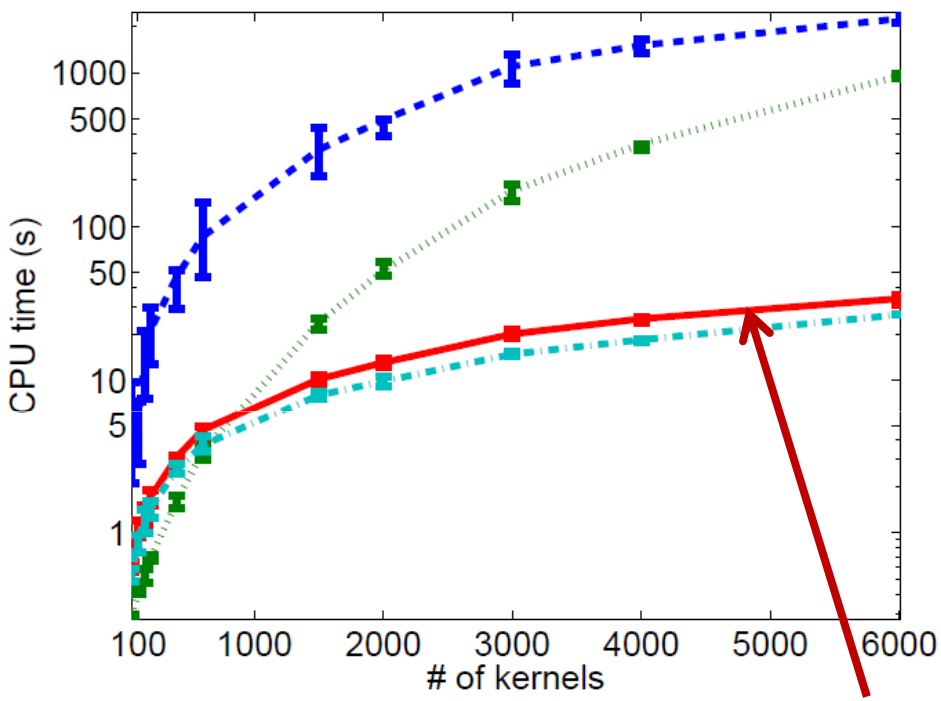
Outline

- Introduction
- Details of Multiple Kernel Learning
- Our method
  - Proximal minimization
  - Skipping inactive kernels
- Numerical Experiments
  - Bench-mark datasets
  - Sparse or dense

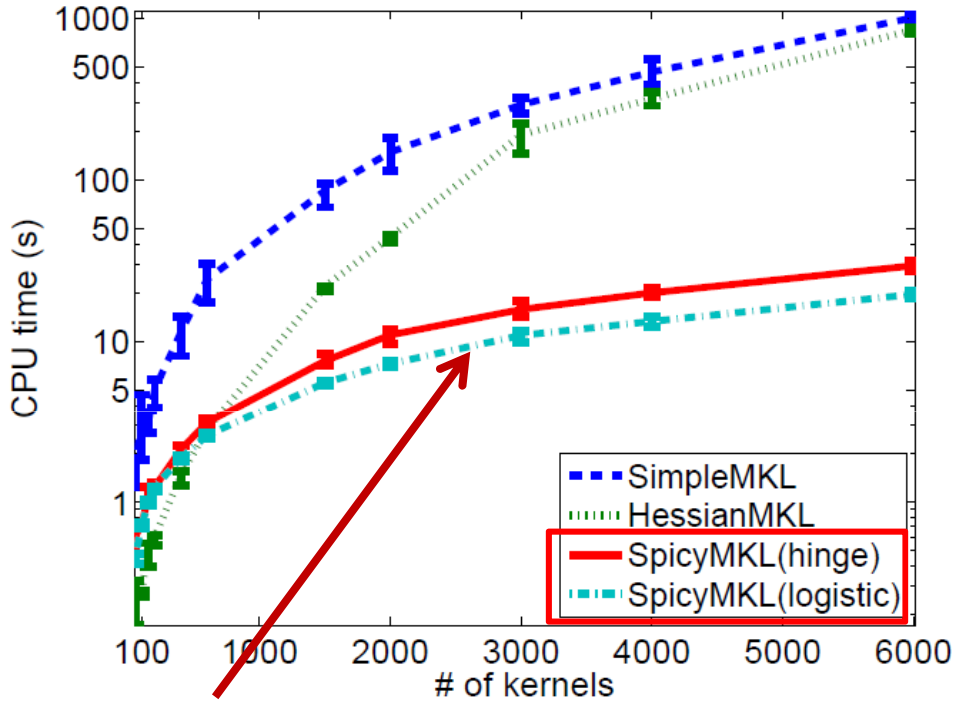
Numerical Experiments

IDA data set, L1 regularization.

CPU time against # of kernels



Splice



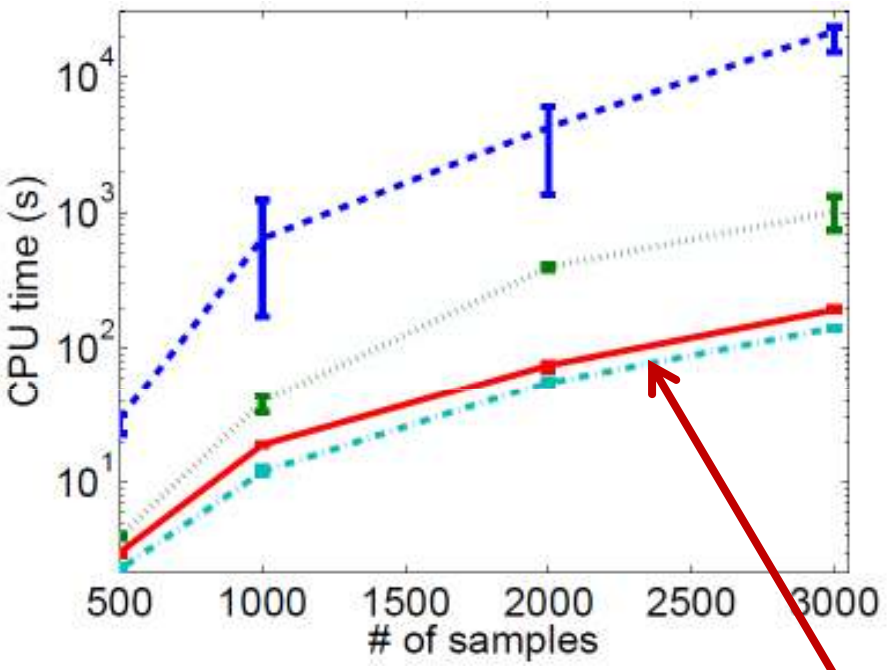
Ringnorm

SpicyMKL

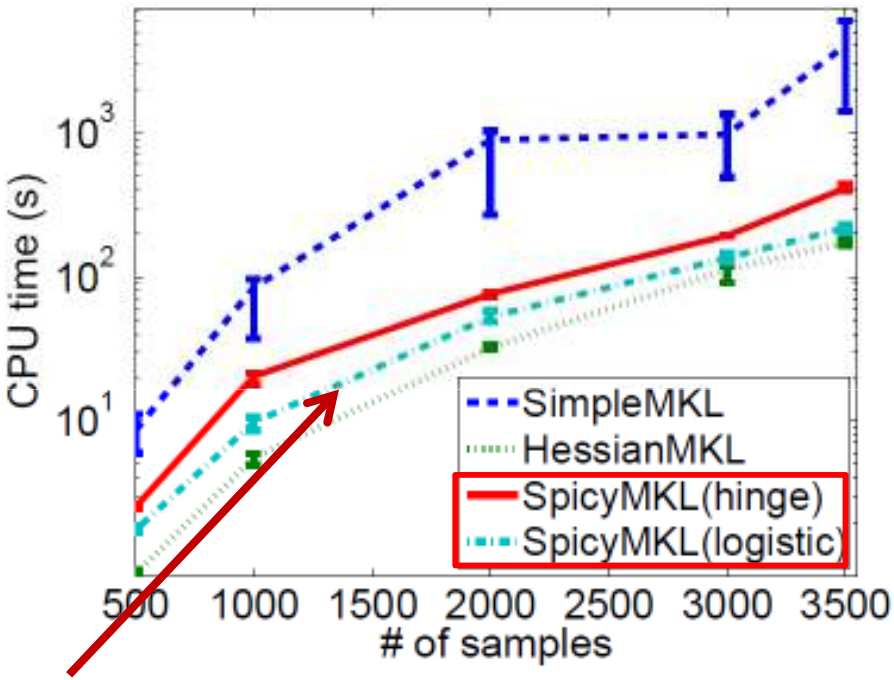
- Scales well against # of kernels.

IDA data set, L1 regularization.

CPU time against # of samples



Splice

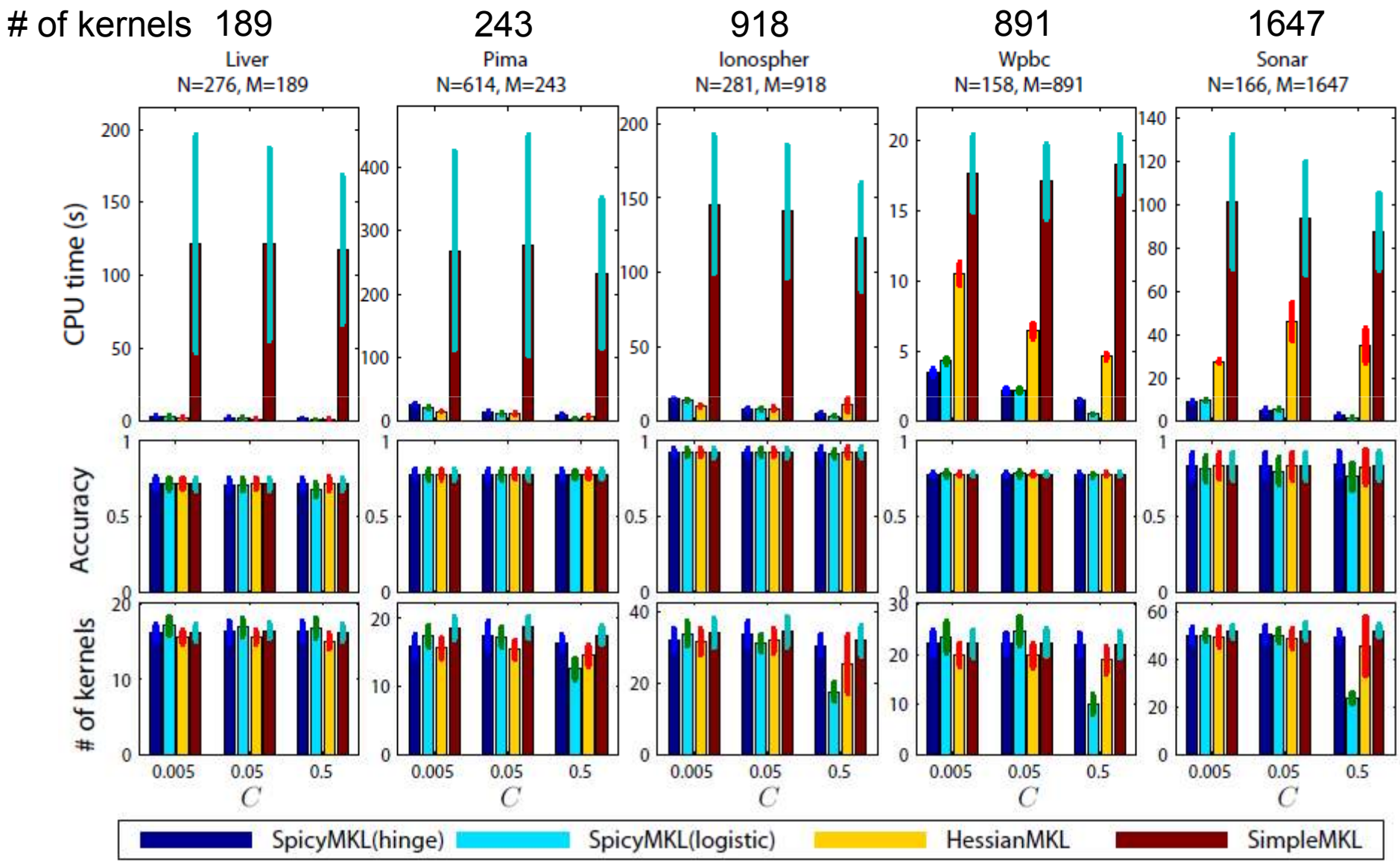


Ringnorm

SpicyMKL

- Scaling against # of samples is almost the same as the existing methods.

UCI dataset


# of kernels 189 243 918 891 1647
Liver
N=276, M=189
Pima
N=614, M=243
Ionospher
N=281, M=918
Wpbc
N=158, M=891
Sonar
N=166, M=1647
CPU time (s)
Accuracy
# of kernels
0.005
0.05
0.5
C
SpicyMKL(hinge)
SpicyMKL(logistic)
HessianMKL
SimpleMKL

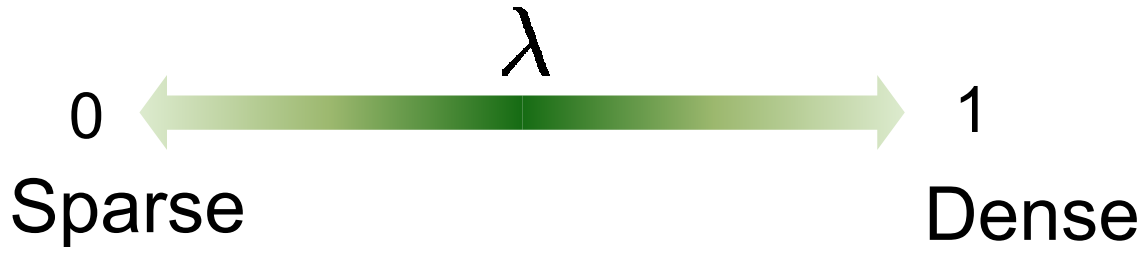
Comparison in elastic net (Sparse v.s. Dense)

Sparse or Dense ?

Elastic Net

$$f_\ell\left(K\alpha\right) + C\sum_{m=1}^{M}\left((1-\lambda)\|\alpha_m\|_m + \lambda\frac{\|\alpha_m\|_m^2}{2}\right)$$

where $\lambda \in [0,1]$.

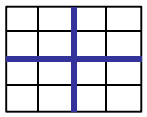


Compare performances of sparse and dense solutions in a computer vision task.

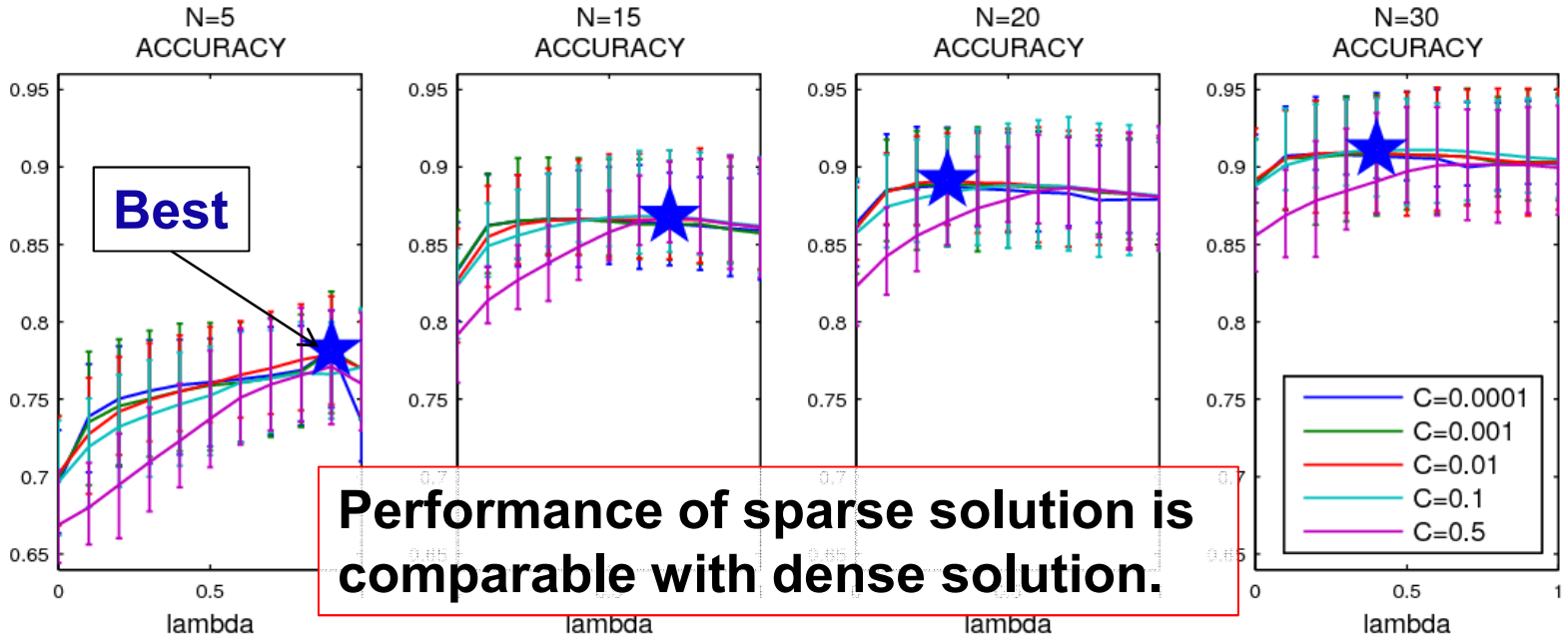
caltech101 dataset (Fei-Fei et al., 2004)

- object recognition task
- anchor, ant, cannon, chair, cup

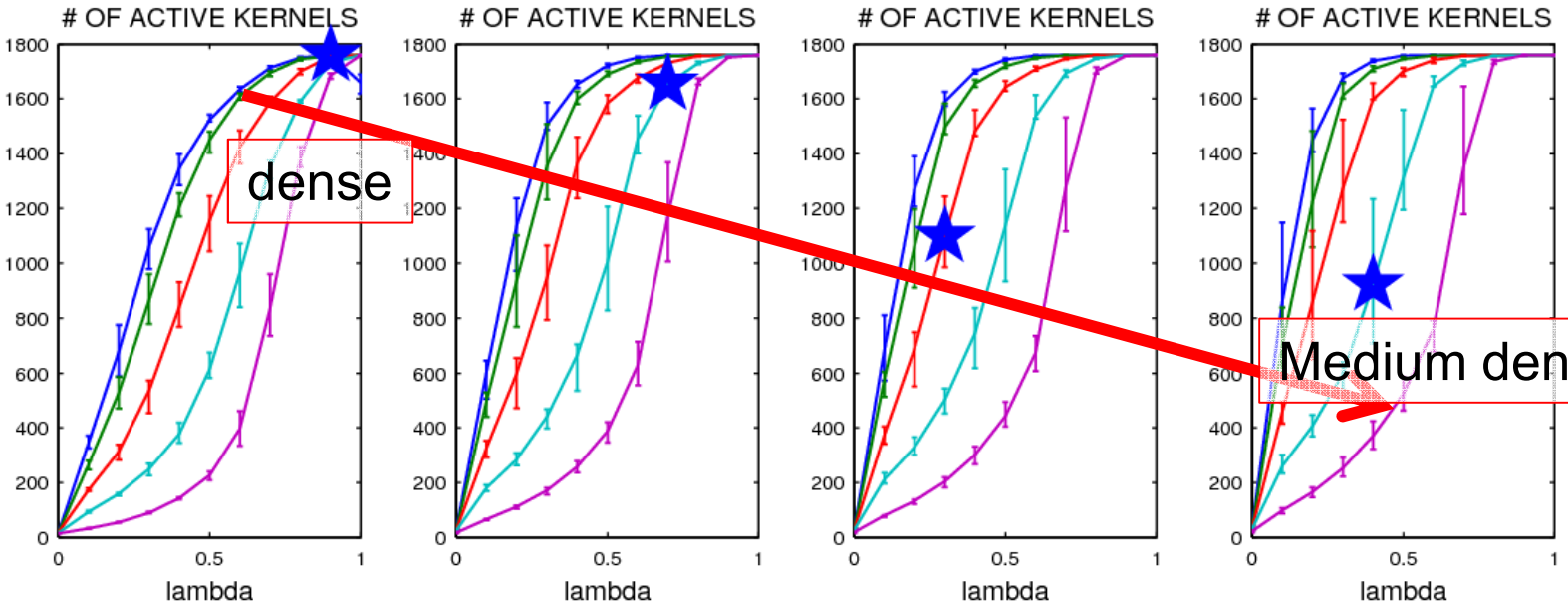
- 10 classification problems (10=5 × (5-1)/2)
- # of kernels: **1760** = 4 (features) × 22 (regions) × 2 (kernel functions) × 10 (kernel parameters)
  - sift features [Lowe99]: colorDescriptor [van de Sande et al.10] hsvsift, sift (scale = 4px, 8px, automatic scale)
  - regions: whole region, 4 division, 16 division, spatial pyramid. (computed histogram of features on each region.)
  - kernel functions: Gaussian kernels, χ^2 kernels with 10 parameters.



Performances are averaged over all classification problems.



Performance of sparse solution is comparable with dense solution.



Conclusion and Future works

Conclusion

- We proposed a new MKL algorithm that is efficient when # of kernels is large.
  - proximal minimization
  - neglect ‘in-active’ kernels
- Medium density showed the best performance, but sparse solution also works well.

Future work

- Second order update

- **Technical report**

  T. Suzuki & R. Tomioka: SpicyMKL.

  arXiv: http://arxiv.org/abs/0909.5026

- **DAL**

  R. Tomioka & M. Sugiyama: Dual Augmented Lagrangian Method for Efficient Sparse Reconstruction. *IEEE Signal Proccesing Letters*, 16 (12) pp. 1067--1070, 2009.

Thank you for your attention!

Some properties

Moreau envelope

$$\mathcal{M}g(z) := \min_x \left(\frac{1}{2}\|x - z\|^2 + g(x) \right)$$

proximal operator

$$\mathrm{prox}_g(z) := \operatorname*{argmin}_x \left(\frac{1}{2}\|x - z\|^2 + g(x) \right)$$

- $\mathcal{M}g(z) + \mathcal{M}g^*(z) = \dfrac{\|z\|^2}{2}$ (Fenchel duality th.)
- $\mathrm{prox}_g(z) + \mathrm{prox}_{g^*}(z) = z$ (Fenchel duality th.)
- $\nabla_z \mathcal{M}g^*(z) = z - \mathrm{prox}_{g^*}(z) = \mathrm{prox}_g(z)$

  Differentiable !

$$\nabla_z \mathcal{M}g^*(x) = \partial_z(\text{obj.}) + \nabla_z \mathrm{prox}_{g^*}(z) \underbrace{\frac{\partial(\text{obj.})}{\partial x}\Big|_{x=\mathrm{prox}_{g^*}(z)}}_{0}$$

$$= z - \mathrm{prox}_{g^*}(z)$$