Computing regularization paths for learning multiple kernels

Francis Bach   Romain Thibaux   Michael Jordan

Computer Science, UC Berkeley

December, 2004

Code available at www.cs.berkeley.edu/~fbach

Computing regularization paths for learning multiple kernels

- Kernel methods for supervised learning:
  - Predict y from x as $f(x) = w^\top \Phi(x)$
  - Learning from data (x_i, y_i), $i = 1, \ldots, n$
  - Optimization problem:

$$\text{minimize} \quad \sum_i \ell(y_i, w^\top \Phi(x_i)) \quad + \quad \lambda ||w||^2$$

"training error" + "regularization"

- Two major issues:
  - Choosing $\Phi(x)$, i.e., the kernel $k(x, y) = \Phi(x)^\top \Phi(y)$
  - Choosing the regularization parameter λ

Learning multiple kernels and regularization paths

- Search over **conic combinations** $k(x,y) = \sum_j \eta_j k_j(x,y)$, $\eta_j \geqslant 0$

- Equivalent to using $\Phi(x) = (\Phi_1(x), \dots, \Phi_m(x))$, $w = (w_1, \dots, w_m)$ and a block 1-norm:

$$\text{minimize} \quad \sum_i \ell(y_i, \sum_j w_j^\top \Phi_j(x_i)) \quad + \quad \lambda \sum_j ||w_j||$$

- Assume $\Phi_j(x)$, $j = 1, \dots, m$, known, and **solve for all** λ
  $\Rightarrow$ compute the regularization path: $w^*(\lambda)$, $\lambda \in \mathbb{R}_+$

- Potential gains:
  - Theoretical: understand block 1-norm regularization better
  - Practical: get the entire path at the cost of one point

"Classical" kernel learning (2-norm regularization)

Primal problem $\min_w \left(\sum_i \varphi_i(w^\top \Phi(x_i)) + \frac{\lambda}{2}||w||^2\right)$

Dual problem $\max_{\alpha \in \mathbb{R}^n} \left(-\sum_i \psi_i(\lambda \alpha_i) - \frac{\lambda}{2}\alpha^\top K \alpha\right)$

Optimality conditions $\forall i, (K\alpha)_i + \psi_i'(\lambda\alpha_i) = 0$

- Assumptions on "loss" φ_i:
  - $\varphi_i(u)$ strictly convex twice differentiable
  - $\psi_i(v)$ Fenchel conjugate of $\varphi_i(u)$, i.e., $\psi_i(v) = \max_{u \in \mathbb{R}}(vu - \varphi_i(u))$

	$\varphi_i(u)$	$\psi_i(v)$
Least-squares regression	$\frac{1}{2}(y_i - u)^2$	$\frac{1}{2}v^2 + vy_i$
Logistic regression	$\log(1 + \exp(-y_i u_i))$	$(1+vy_i)\log(1+vy_i)$ $-vy_i \log(-vy_i)$

Block 1-norm regularization

- m feature spaces $\mathcal{F}_j$ and feature maps $\Phi_j(x)$:
- Primal problem:

$$\min_{w \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_m} \sum_i \varphi_i \left(\sum_j w_j^\top \Phi(x_l) \right) + \lambda \sum_j d_j ||w_j||.$$

- Convex non differentiable : reformulation using conic constraints
- Dual problem:

$$\max_\alpha - \sum_i \psi_i(\lambda \alpha_i) \text{ such that } \forall j, \alpha^\top K_j \alpha \leqslant d_j^2$$

Block 1-norm regularization

- Optimality conditions:

$$\forall i,\ \left(\sum_j \eta_j K_j \alpha\right)_i + \psi_i'(\lambda \alpha_i) = 0$$
$$\forall j, \alpha^\top K_j \alpha \leqslant d_j^2, \eta_j \geqslant 0, \eta_j (d_i^2 - \alpha^\top K_j \alpha) = 0.$$

- Optimal solution α, solution of the 2-norm problem with a conic combination of basis kernels:

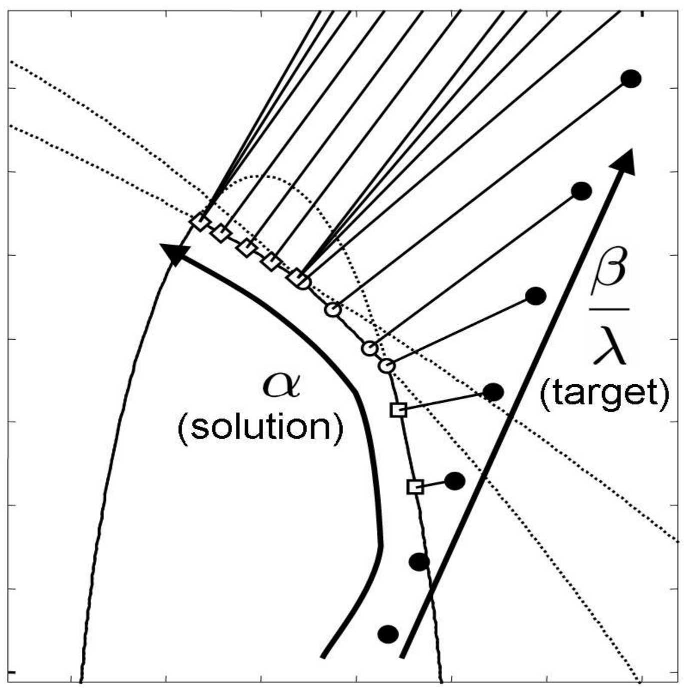
$$K = \sum_j \eta_j K_j$$

Geometric intepretation

- Dual problem:

$$\max_\alpha - \sum_i \psi_i(\lambda \alpha_i) \ \text{ such that } \ \forall j, \alpha^\top K_j \alpha \leqslant d_j^2$$

- "target" : $\beta_i = \arg\max \psi_i(v)$



Active sets

- If $J = \{j, \eta_j > 0\}$ is known, solution (α, η) defined by

$$\forall i,\ \left(\sum_{j \in J} \eta_j K_j \alpha\right)_i + \psi_i'(\lambda \alpha_i) = 0$$
$$\forall j \in J, \alpha^\top K_j \alpha = d_j^2$$

- $n + |J|$ differentiable equations with $n + |J|$ unknowns
  $\Rightarrow$ **smooth path, easy to follow,** but ...

- Valid while $\eta_j \geqslant 0$, $j \in J$, and $\alpha^\top K_j \alpha \leqslant d_j^2$, $j \notin J$.

- Change of active sets
  $\Rightarrow$ **piecewise smooth path, hard to follow**

- NB: with one kernel, path is piecewise linear (Hastie et at., 2004)

Log-barrier regularization

- Dual problem:

$$\max_\alpha -\sum_i \psi_i(\lambda\alpha_i) \text{ such that } \forall j, \alpha^\top K_j \alpha \leqslant d_j^2$$

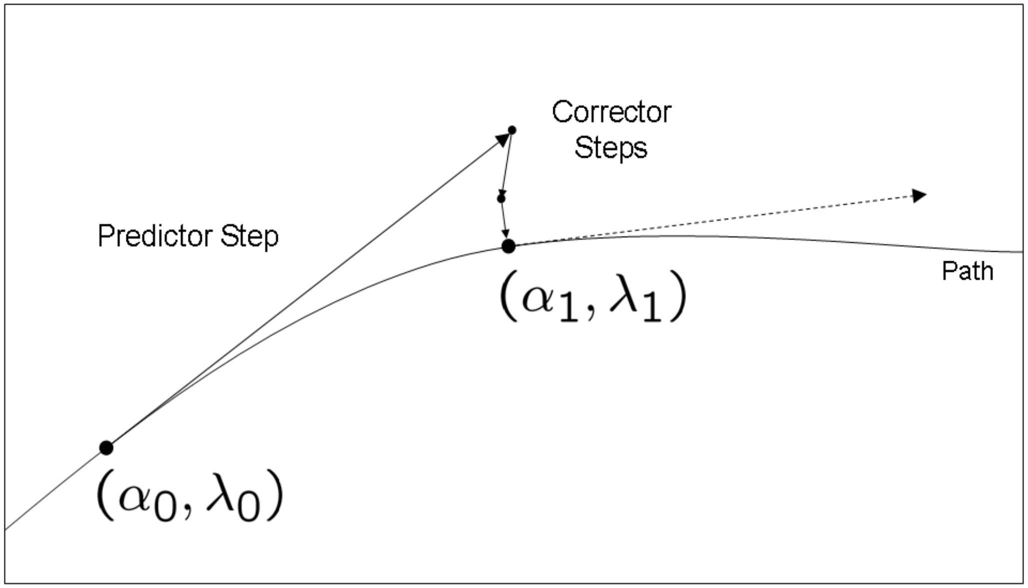
- Regularized dual problem:

$$\max_\alpha -\sum_i \psi_i(\lambda\alpha_i) + \mu \sum_j \log(d_j^2 - \alpha^\top K_j \alpha)$$

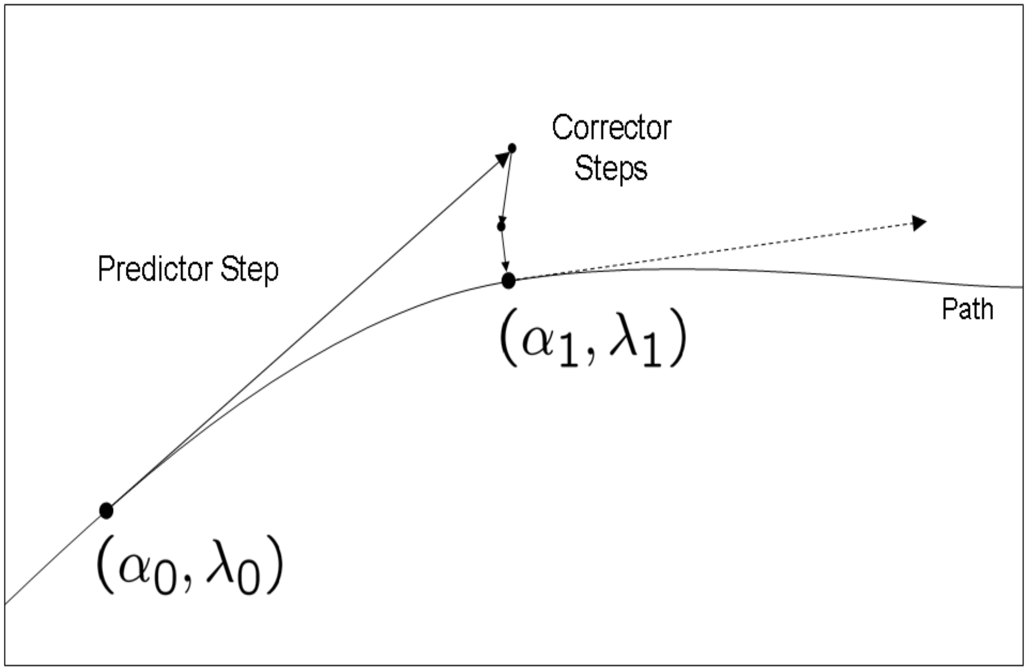
- Properties:
  - Unconstrained concave maximization
  - η function of α
  - α is unique
  - $\alpha(\lambda)$ differentiable function, easy to follow

Predictor-corrector method

- Follow solution of $F(\alpha, \lambda) = 0$
- Predictor steps
  - First order approximation using $\frac{d\alpha}{d\lambda} = -\left(\frac{\partial F}{\partial \alpha}\right)^{-1} \frac{\partial F}{\partial \lambda}$
- Corrector steps
  - Newton's method to converge back to solution

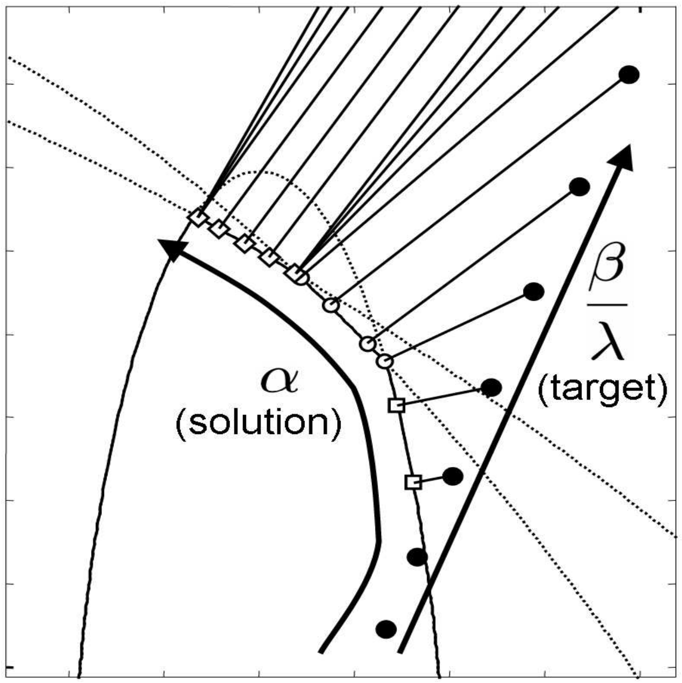


Predictor-corrector method: implementation issues



- Step-size selection for predictor step: $\delta\sigma$
  - adaptive selection
- Second order approximation

Initialization



- if $\left(\frac{\beta}{\lambda}\right)^\top K_j \left(\frac{\beta}{\lambda}\right) \leqslant d_j^2$, then $\alpha = \frac{\beta}{\lambda}$ is solution

- Initialize using $\lambda = \max_j (\beta^\top K_j \beta / d_j^2)^{1/2}$ and $\alpha = \beta/\lambda$

Link with interior point methods

- Regularized dual problem:

$$\max_\alpha - \sum_i \psi_i(\lambda \alpha_i) + \mu \sum_j \log(d_j^2 - \alpha^\top K_j \alpha)$$

- Interior point methods:
  - λ fixed, μ followed from large to small

- Regularization path:
  - μ fixed small, λ followed from large to small

Computational complexity

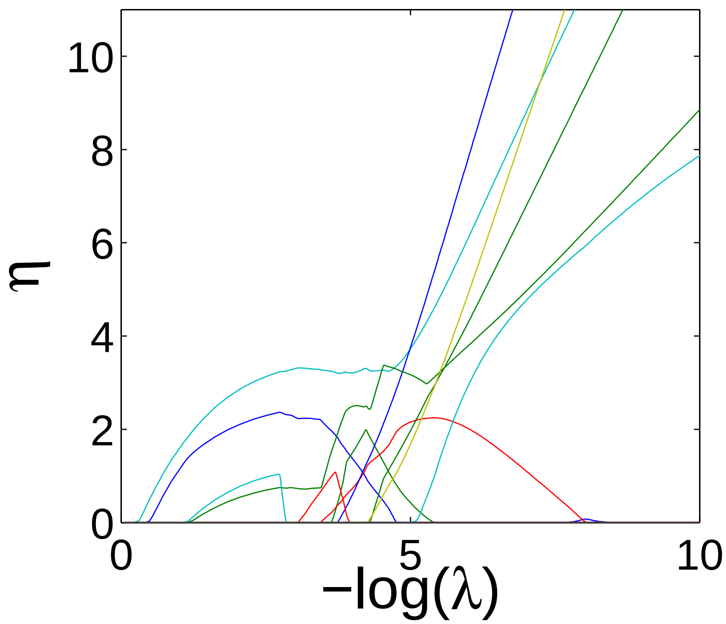
- n number of data points, m number of kernels
- Interior point method to obtain one solution: $O(mn^3)$
- Path following method
  - Each predictor-corrector step: $O(n^3)$
  - Empirically $O(m)$ steps
  - Total complexity $O(mn^3)$

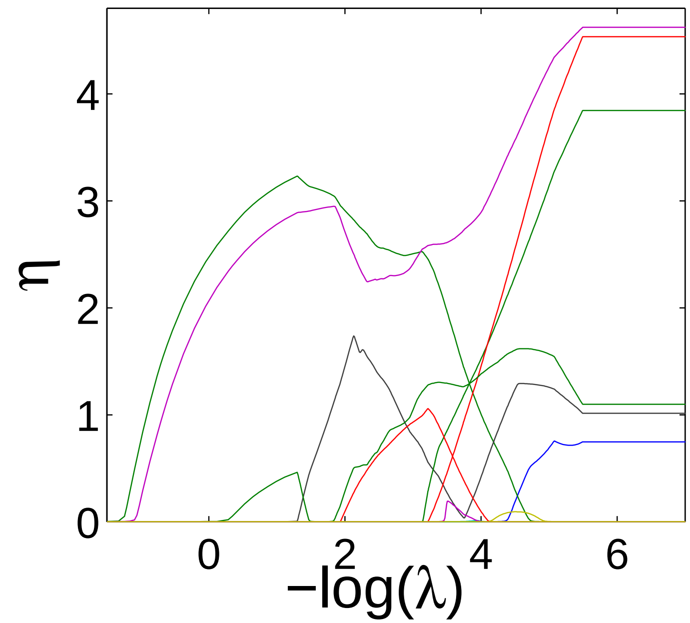
Simulations

- Set up for given supervised learning problem:
  - Build a large number of “classical” kernels
  - Perform path following
  - Compute performance on held out validation data
- Goals:
  - Select best regularization parameter
  - Understand how regularization behaves

Simple example

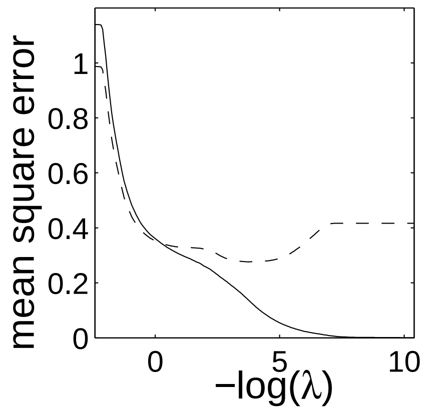
- Left: regression, right: classification

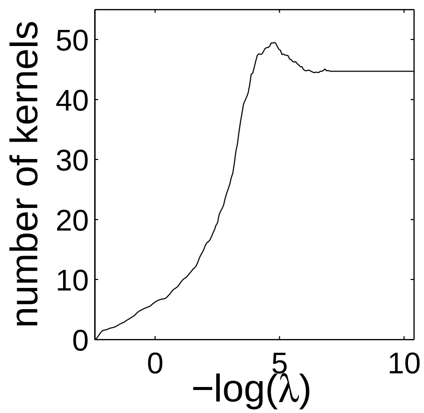




  - η_j is not a monotonic function of λ
  - Canonical behavior for extreme values of λ

Training/testing error





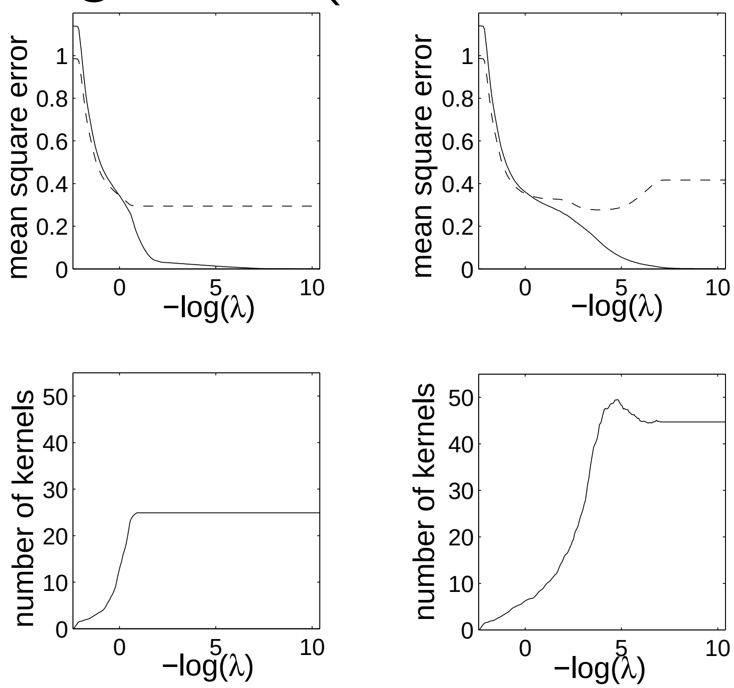
- Canonical behavior as λ decreases
  - Training performance decreases to zero
  - Testing performance decreases, increases, then stabilizes

- Importance of d_j (weight of penalization $= \sum_j d_j ||w_j||$)
  - d_j should be an increasing function of the "rank" of K_j:
$$d_j = \left(\text{number of eigenvalue } \geqslant \frac{1}{2n}\right)^{\gamma}$$
  - γ small $\Rightarrow d_j$ rank independent

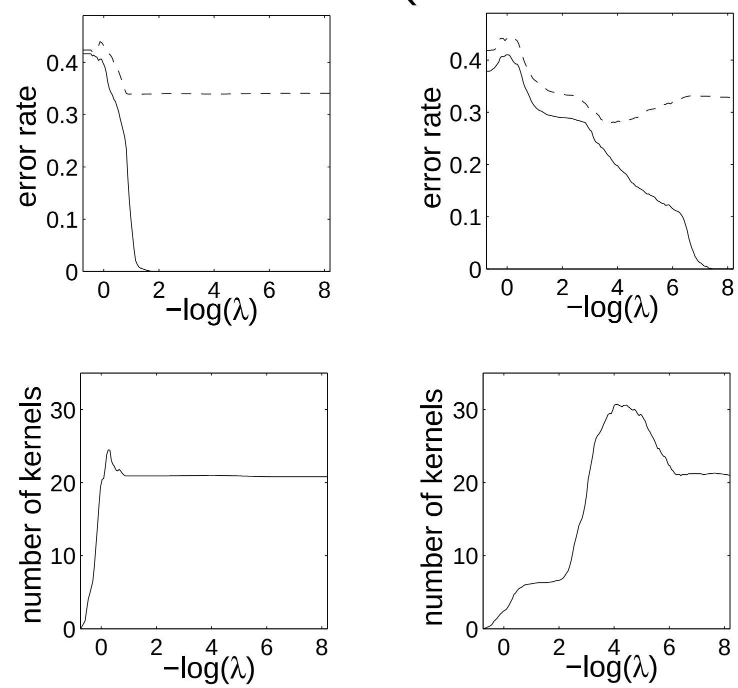
Importance of d_j

- Left: $\gamma = 0$, right: $\gamma = 1$
- Top: training (bold)/testing (dashed) error
  bottom: number of kernels

Regression (Boston dataset)



Classification (Liver dataset)



Conclusion

- Computing regularization paths for multiple kernels
  - Same complexity than solving for one point
  - Theoretical understanding of regularization
  - Practical implications

- Future work:
  - Theoretical complexity results
  - Efficient implementation: from cubic to quadratic in n