Consistency of group lasso and multiple kernel learning

Francis Bach
INRIA - Ecole Normale Supérieure
Willow project

November 2007
Summary

- Machine learning and regularization
- Group Lasso
  - Consistent estimation of groups?
- Multiple kernel learning as non parametric group Lasso
- Extension to trace norm minimization
Supervised learning and regularization

- Data: \( x_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = 1, \ldots, n \)

- Minimize with respect to function \( f \in \mathcal{F} \):

\[
\sum_{i=1}^{n} \ell(y_i, f(x_i)) + \frac{\lambda}{2} \| f \|^2
\]

Error on data + Regularization

Loss & function space \( ? \) Norm \( ? \)

- Two issues:
  - Loss
  - Function space / norm
Usual losses

- **Regression**: $y \in \mathbb{R}$, prediction $\hat{y} = f(x)$, quadratic cost $\ell(y, f) = \frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - f)^2$

- **Classification**: $y \in \{-1, 1\}$ prediction $\hat{y} = \text{sign}(f(x))$
  - loss of the form $\ell(y, f) = \ell(yf)$
  - “True” cost: $\ell(yf) = 1_{yf<0}$
  - Usual convex costs:
Regularizations

• Main goal: control the “capacity” of the learning problem

• Two main lines of work

  1. Use Hilbertian (RKHS) norms
     – Non parametric supervised learning and kernel methods
     – Well developed theory
  2. Use “sparsity inducing” norms
     – main example: $\ell_1$ norm
     – Perform model selection as well as regularization
     – Often used heuristically

• Group lasso / MKL : two types of regularizations
Group lasso - linear predictors

- Assume $x_i, w \in \mathbb{R}^p$ where $p = p_1 + \cdots + p_m$, i.e., $m$ groups
  
  $$x_i = (x_{i1}, \ldots, x_{im}) \quad w = (w_1, \ldots, w_m)$$

- Goal: achieve sparsity at the levels of groups: $J(w) = \{i, \ w_i \neq 0\}$

- Main application:
  - Group selection vs. variable selection (Zhao et al., 2006)
  - Multi-task learning (Argyriou et al., 2006, Obozinsky et al., 2007)

- Regularization by block $\ell_1$-norm (Yuan & Lin, 2006, Zhao et al., 2006, Bach et al., 2004):
Group lasso - Main questions

\[ \min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \lambda \sum d_j \| w_j \| \]

1. Analysis of sparsity inducing property:
   - where do \( \hat{w} \) and \( J(\hat{w}) = \{ i, \ \hat{w}_i \neq 0 \} \) converge to?
   - letting the problem grow
     - sizes of the groups \( p_i, i = 1, \ldots, m \Rightarrow \text{“kernelization”} \)
     - number of groups \( m \Rightarrow ? \)
   - Influence of the weights \( d_j \)

2. Algorithms
   - very efficient and elegant for the Lasso (Efron et al., 2004)
Group lasso - Asymptotic analysis
Groups of finite sizes - Square loss

• Assumptions:

1. Data \((X_i, Y_i)\) sampled \text{i.i.d.}
2. \(w \in \mathbb{R}^p\) denotes the (unique) minimizer of \(\mathbb{E}(Y - X^\top w)^2\) (best linear predictor). Assume \(\mathbb{E}((Y - w^\top X)^2 | X) \geq \sigma^2_{\min} > 0\) a.s.
3. Finite fourth order moments: \(\mathbb{E}\|X\|^4 < \infty\) and \(\mathbb{E}\|Y\|^4 < \infty\).
4. Invertible covariance: \(\Sigma_{XX} = \mathbb{E}XX^\top \in \mathbb{R}^{p \times p}\) is invertible.

• Denote \(J = \{j, w_j \neq 0\}\) the sparsity pattern of \(w\)

• Goal: estimate consistently both \(w\) and \(J\) when \(n\) tends to infinity
  
  – \(\forall \epsilon > 0, \ P(\|\hat{w} - w\| > \epsilon)\) tends to zero
  – \(P(\{j, \hat{w}_j \neq 0\} \neq J)\) tends to zero
  – Rates of convergence
Group lasso - Consistency conditions

• Strict condition:

$$\max_{i \in J^c} \frac{1}{d_i} \left\| \Sigma_{X_iX_J} \Sigma_{X_JX_J}^{-1} \text{Diag}(d_j/\|w_j\|)w_J \right\| < 1$$

• Weak condition:

$$\max_{i \in J^c} \frac{1}{d_i} \left\| \Sigma_{X_iX_J} \Sigma_{X_JX_J}^{-1} \text{Diag}(d_j/\|w_j\|)w_J \right\| \leq 1$$

• **Theorem 1:** Strict condition is **sufficient** for joint regular and sparsity consistency of the group lasso ($\lambda_n \to 0$ and $\lambda_n n^{1/2} \to +\infty$)

• **Theorem 2:** Weak condition is **necessary** for joint regular and sparsity consistency of the group lasso (for any $\lambda_n$).
Group lasso - Consistency conditions

• Condition:

$$\max_{i \in J_c} \frac{1}{d_i} \left\| \Sigma_{X_i X_J} \Sigma_{X_J X_J}^{-1} \text{Diag}(d_j / \|w_j\|) w_j \right\| < \text{ or } \leq 1$$


• Additional questions:
  – Is strict condition necessary (as in the Lasso case)?
  – Estimate of probability of correct sparsity estimation
  – Loading independent condition
  – Other losses
  – Negative or positive result?
Group lasso - Strict condition necessary?

- Strict condition necessary for the Lasso (Zou, 2006, Zhao and Yu, 2006)

- Strict condition not necessary for the group Lasso

- If weak condition is satisfied and for all \( i \in J^c \) such that
  \[
  \frac{1}{d_i} \left\| \Sigma X_i X_J \Sigma^{-1} X_J X_J \text{Diag}(d_j/\|w_j\|) w_J \right\| = 1,
  \]
  we have

  \[
  \Delta^T \Sigma X_J X_i \Sigma X_i X_J \Sigma^{-1} X_J X_J \text{Diag} \left[ d_j/\|w_j\| \left( I_{p_j} - \frac{w_j w_j^T}{w_j^T w_j} \right) \right] \Delta > 0,
  \]

  with
  \[
  \Delta = -\Sigma^{-1} X_J \text{Diag}(d_j/\|w_j\|) w_J,
  \]

  then the group lasso estimate leads to joint regular and sparsity consistency \((\lambda_n \to 0 \text{ and } \lambda_n n^{1/4} \to +\infty)\)
Loading independent sufficient condition

- Condition on $\Sigma$ and $J$:

$$\max_{w_J} \max_{i \in J^c} \frac{1}{d_i} \left\| \Sigma X_i X_J \Sigma_{X_J X_J}^{-1} \text{Diag}(d_j/\|w_j\|)w_J \right\| < 1$$

$$\Leftrightarrow \max_{i \in J^c} \frac{1}{d_i} \max_{\|u_j\|=1, \forall j \in J} \left\| \Sigma X_i X_J \Sigma_{X_J X_J}^{-1} \text{Diag}(d_j)u_J \right\| < 1$$

$$\Rightarrow \max_{i \in J^c} \frac{1}{d_i} \sum_{j \in J} d_j \left\| \sum_{k \in J} \Sigma X_i X_k \left( \Sigma_{X_J X_J}^{-1} \right)_{k,j} \right\| < 1$$

- Lasso (groups of size 1): all those are equivalent

- Group lasso: stricter sufficient condition (in general)
  - NB: can obtain better one with convex relaxation (see paper)
Probability of correct selection of pattern

- Simple general result when $\lambda_n = \lambda_0 n^{-1/2}$

- Probability equal to

$$
\mathbb{P} \left( \max_{i \in J^c} \left\| \frac{\sigma}{\sqrt{n} \lambda_n d_i} \sum X_i X_J \sum_{X_J}^{-1/2} u - \frac{1}{d_i} \sum X_i X_J \sum_{X_J}^{-1} \right\| \mathbb{I} \leq 1 \right)
$$

where $u$ is normal with mean zero and identity covariance matrix.

- With additional conditions, valid when $\lambda_n n^{1/2}$ not too far from constant $\Rightarrow$ exponential rate of convergence if strict condition is satisfied

- Dependence on $\sigma$ and $n$
Positive or negative result?

- “Disappointing” result for Lasso/group Lasso
  - Does not always do what heuristic justification suggests!

- Can we make it always consistent?
  - Data dependent weights \(\Rightarrow\) adaptive Lasso/group Lasso

- Do we care about exact sparsity consistency?
  - Recent results by Meinshausen and Yu (2007)
Relationship with multiple kernel learning (MKL) (Bach, Lanckriet, Jordan, 2004)

- Alternative equivalent formulation:

\[
\min_{w \in \mathbb{R}^p} \frac{1}{2n} \| \bar{Y} - \bar{X} w \|^2 + \frac{1}{2} \mu_n \left( \sum_{j=1}^m d_j \| w_j \| \right)^2
\]

- Dual optimization problem (using conic programming):

\[
\max_{\alpha \in \mathbb{R}^n} \left\{ -\frac{1}{2n} \| \bar{Y} - n \mu_n \alpha \|^2 - \frac{1}{2 \mu_n} \max_{i=1, \ldots, m} \frac{\alpha^\top K_i \alpha}{d_i^2} \right\}
\]
Relationship with multiple kernel learning (MKL) (Bach, Lanckriet, Jordan, 2004)

- Alternative equivalent formulation:

\[
\min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{2n} \| \bar{\mathbf{Y}} - \bar{\mathbf{X}} \mathbf{w} \|^2 + \frac{1}{2} \mu_n \left( \sum_{j=1}^{m} d_j \| \mathbf{w}_j \| \right)^2
\]

- Dual optimization problem (using conic programming):

\[
\max_{\alpha \in \mathbb{R}^n} \left\{ -\frac{1}{2n} \| \bar{\mathbf{Y}} - n \mu_n \alpha \|^2 - \frac{1}{2 \mu_n} \max_{i=1,\ldots,m} \frac{\alpha^\top K_i \alpha}{d_i^2} \right\}
\]

\[
\Leftrightarrow \max_{\alpha \in \mathbb{R}^n} \min_{\eta \geq 0, \sum_{j=1}^{m} \eta_j d_j^2 = 1} \left\{ -\frac{1}{2n} \| \bar{\mathbf{Y}} - n \mu_n \alpha \|^2 - \frac{1}{2 \mu_n} \alpha^\top \left( \sum_{j=1}^{m} \eta_j K_j \right) \alpha \right\}
\]
Relationship with multiple kernel learning (MKL) (Bach, Lanckriet, Jordan, 2004)

• Alternative equivalent formulation:

\[
\min_{w \in \mathbb{R}^p} \frac{1}{2n} \| \bar{Y} - \bar{X} w \|^2 + \frac{1}{2\mu_n} \left( \sum_{j=1}^m d_j \| w_j \| \right)^2
\]

• Dual optimization problem:

\[
\max_{\alpha \in \mathbb{R}^n} \left\{ -\frac{1}{2n} \| \bar{Y} - n\mu_n \alpha \|^2 - \frac{1}{2\mu_n} \max_{i=1,\ldots,m} \frac{\alpha^\top K_i \alpha}{d_i^2} \right\}
\]

\[
\Leftrightarrow \min_{\eta \geq 0, \sum_{j=1}^m \eta_j d_j^2 = 1} \max_{\alpha \in \mathbb{R}^n} \left\{ -\frac{1}{2n} \| \bar{Y} - n\mu_n \alpha \|^2 - \frac{1}{2\mu_n} \alpha^\top \left( \sum_{j=1}^m \eta_j K_i \right) \alpha \right\}
\]
Relationship with multiple kernel learning (MKL)

\[
\min_{\eta \geq 0, \sum_{j=1}^{m} \eta_j d_j^2 = 1} \max_{\alpha \in \mathbb{R}^n} \left\{ -\frac{1}{2n} \| \bar{Y} - n \mu n \alpha \|^2 - \frac{1}{2\mu n} \alpha^\top \left( \sum_{j=1}^{m} \eta_i K_i \right) \alpha \right\}
\]

- Optimality conditions: the dual variable \( \alpha \in \mathbb{R}^n \) is optimal if and only if there exists \( \eta \in \mathbb{R}_+^m \) such that \( \sum_{j=1}^{m} \eta_j d_j^2 = 1 \) and \( \alpha \) is optimal for ridge regression problem with kernel matrix \( K = \sum_{j=1}^{m} \eta_j K_j \)

- \( \eta \) can also be obtained as the minimizer of

\[
J(\eta) = \max_{\alpha \in \mathbb{R}^n} \left\{ -\frac{1}{2n} \| \bar{Y} - n \mu n \alpha \|^2 - \frac{1}{2\mu n} \alpha^\top \left( \sum_{j=1}^{m} \eta_j K_j \right) \alpha \right\}
\]

- \( J(\eta) \) is the optimal value of the objective function of the single kernel estimation problem with kernel \( K = \sum_{j=1}^{m} \eta_j K_j \)
Multiple kernel learning (MKL)

- Jointly learn optimal (sparse) combination of kernel ($\eta$) together with the estimate with this kernel ($\alpha$)

- Application
  - Kernel learning
  - Heterogeneous data fusion

- Known issues
  - Algorithms
  - Influence of weights $d_j$ (feature spaces have different sizes)
  - Consistency
Analysis of MKL as non parametric group Lasso

- Assume $m$ Hilbert spaces $\mathcal{F}_i$, $i = 1, \ldots, m$

$$\min_{f_i \in \mathcal{F}_i, i=1,\ldots,m} \frac{1}{2n} \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{m} f_j(x_{ji}) \right)^2 + \frac{\mu n}{2} \left( \sum_{j=1}^{m} d_j \| f_j \| \right)^2.$$

- Sparse generalized additive models (Hastie and Tibshirani, 1990)

- Estimate is obtained through MKL formulation

- Same question: regular and sparsity consistency when the groups are infinite-dimensional Hilbert spaces
Analysis of MKL as non parametric group Lasso (non centered) covariance operators

- Single random variable $X$: $\Sigma_{XX}$ is a bounded linear operator from $\mathcal{F}$ to $\mathcal{F}$ such that for all $(f, g) \in \mathcal{F} \times \mathcal{F}$,

$$\langle f, \Sigma_{XX} g \rangle = \mathbb{E}(f(X)g(X))$$

Under minor assumptions, the operator $\Sigma_{XX}$ is auto-adjoint, non-negative and Hilbert-Schmidt

- Tool of choice for the analysis of least-squares non parametric methods (Fukumizu et al., 2005, 2006, Gretton et al., 2006, etc...)

- Natural empirical estimate $\hat{\Sigma}_{XX} = \frac{1}{n} \sum_{i=1}^{n} k(\cdot, x_i) \otimes k(\cdot, x_i)$ converges in probability to $\Sigma_{XX}$ in HS norm.
Cross-covariance operators

- Several random variables: cross-covariance operators $\Sigma_{X_i X_j}$ from $F_j$ to $F_i$ such that $\forall (f_i, f_j) \in F_i \times F_j$,

\[
\langle f_i, \Sigma_{X_i X_j} f_j \rangle = \mathbb{E}(f_i(X_i)f_j(X_j))
\]

- Similar convergence properties of empirical estimates

- Joint covariance operator $\Sigma_{XX}$ defined by blocks

- We can define the bounded *correlation* operators through

\[
\Sigma_{X_i X_j} = \Sigma_{X_i X_i}^{1/2} C_{X_i X_j} \Sigma_{X_j X_j}^{1/2}
\]

- NB: the joint covariance operator is never invertible, but the correlation operator may be
Analysis of MKL as non parametric group Lasso

• Assumptions

1. \( \forall j, \mathcal{F}_j \) is a separable RKHS associated with kernel \( k_j \), and \( \mathbb{E}k_j(X_j, X_j)^2 < \infty \).

2. **Model**: There exists functions \( f = (f_1, \ldots, f_m) \in \mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_m \) and a function \( h \) of \( X = (X_1, \ldots, X_m) \) such that

\[
\mathbb{E}(Y|X) = \sum_{j=1}^{m} f_j(X_j) + h(X)
\]

with \( \mathbb{E}h(X)^2 < \infty \) and \( \mathbb{E}h(X)f_j(X_j) = 0 \) for all \( j = 1, \ldots, m \) and \( \mathbb{E}((Y - \sum_{j=1}^{m} f_j(X_j))^2|X) \geq \sigma_{\min}^2 > 0 \text{ a.s.} \)

3. **Compacity and invertibility** : All cross-correlation operators are compact and the joint correlation operator is invertible.

4. **Range condition**: For all \( j \), \( \exists g_j \in \mathcal{F}_j \) such that \( f_j = \Sigma_{X_jX_j}^{1/2}g_j \)
Compacity and invertibility of joint correlation operator

• Sufficient condition for compacity when distributions have densities:

\[ \mathbb{E} \left\{ \frac{p_{X_i X_j}(x_i, x_j)}{p_{X_i}(x_i)p_{X_j}(x_j)} - 1 \right\} < \infty. \]

– Dependence between variables is not too strong

• Sufficient condition for invertibility: no exact correlation using functions in the RKHS.

– Empty concurvity space assumption (Hastie and Tibshirani, 1990)
Range condition

• Technical condition: For all $j$, $\exists g_j \in \mathcal{F}_j$ such that $f_j = \sum_{X_j}^{1/2} X_j g_j$
  – Conditions on the support of $f_j$ with respect to the support of the data
  – Conditions on the smoothness of $f_j$

• Sufficient condition for translation invariant kernels

\[ k(x, x') = q(x - x') \text{ in } \mathbb{R}^d: \]

– $f_j$ is of the form $f_j = q \ast g_j$ where $\int \frac{g_j^2(x_j)}{p_{X_j}(x_j)} dx_j$. 
Group lasso - Consistency conditions

- **Strict condition**

\[
\max_{i \in J^c} \frac{1}{d_i} \left\| \Sigma X_i X_i C_{X_i X_J} C_{X_J X_J}^{-1} \text{Diag}(d_j/\|f_j\|) g_J \right\| < 1
\]

- **Weak condition**

\[
\max_{i \in J^c} \frac{1}{d_i} \left\| \Sigma X_i X_i C_{X_i X_J} C_{X_J X_J}^{-1} \text{Diag}(d_j/\|f_j\|) g_J \right\| \leq 1
\]

- **Theorem 1**: Strict condition is **sufficient** for joint regular and sparsity consistency of the lasso.

- **Theorem 2**: Weak condition is **necessary** for joint regular and sparsity consistency of the lasso.
Adaptive group lasso

- Consistency condition depends on \( w \) or \( f \) and is not always satisfied!

- Empirically, the weights do matter a lot (Bach, Thibaux, Jordan, 2005)
Importance of weights (Bach, Thibaux, Jordan, 2005)

- Canonical behavior as $\lambda$ decreases
  - Training error decreases to zero
  - Testing error 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 $K_j$, e.g.,
    (when matrices are normalized to unit trace):
    $$d_j = \left( \text{number of eigenvalue } \geq \frac{1}{2n} \right)^\gamma$$
Importance of weights (Bach, Thibaux, Jordan, 2005)

• Left: $\gamma = 0$ (unit trace, Lanckriet et al., 2004), right: $\gamma = 1$

• Top: training (bold)/testing (dashed) error 
  bottom: number of kernels

Regression (Boston dataset)  
Classification (Liver dataset)
Adaptive group lasso

- Consistency condition depends on \( w \) or \( f \) and is not always satisfied!

- Empirically, the weights do matter a lot (Bach, Thibaux, Jordan, 2005)

- Extension of the Lasso adaptive version (Yuan & Lin, 2006) using the regularized LS estimate \( \hat{f}^{LS}_{\kappa_n} \) defined as:

\[
\min_{f_i \in F_i, \ i=1,\ldots,m} \frac{1}{2n} \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{m} f_j(x_{ji}) \right)^2 + \frac{\kappa_n}{2} \sum_{j=1}^{m} \| f_j \|^2,
\]
Adaptive group lasso

- **Theorem**: Let \( \hat{f}_{LS}^{n^{-1/3}} \) be the least-square estimate with regularization parameter proportional to \( n^{-1/3} \). Let \( \hat{f} \) denote any minimizer of

\[
\frac{1}{2n} \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{m} f_j(x_{ji}) \right)^2 + \frac{\mu_0 n^{-1/3}}{2} \left( \sum_{j=1}^{m} \| (\hat{f}_{LS}^{\kappa_n})_j \| - \gamma \| f_j \| \right)^2.
\]

For any \( \gamma > 1 \), \( \hat{f} \) converges to \( f \) and \( J(\hat{f}) \) converges to \( J \) in probability.

- Convergence rates with more assumptions (and more work!)
- Practical implications in applications to be determined
Algorithms for group lasso and MKL

- Algorithms for general convex losses
- many different interpretations implies many different algorithms
  - Group Lasso - primal formulation w.r.t. $w$
  - Group Lasso - dual formulation w.r.t. $\alpha$
  - Direct problem involving $\eta$
Algorithms for MKL

• (very) costly optimization with SDP, QCQP ou SOCP (Lanckriet et al., 2004)
  – $n \geq 1,000 - 10,000$, $m \geq 100$ not possible
  – “loose” required precision ⇒ first order methods

• Shooting algorithm (Yuan & Lin, 2006)

• Dual coordinate ascent (SMO) with smoothing (Bach et al., 2004)

• Optimization of $J(\eta)$ by cutting planes (Sönnernenburg et al., 2005)

• Optimization of $J(\eta)$ with steepest descent with smoothing (Rakotomamonjy et al, 2007)

• Regularization path (Bach, Thibaux & Jordan, 2005)
Illustrative toy experiments

- 6 groups of size 2 - $\text{Card}(J) = 3$

- Consistent condition **fullfilled**:

![Graphs showing original, adaptive, and unit trace results](image-url)
Illustrative toy experiments

- 6 groups of size 2 - \( \text{Card}(J) = 3 \)

- Consistent condition **not fullfilled**:

```
Original

Adaptive

Unit trace
```
Applications

- Bioinformatics (Lanckriet et al., 2004)
  - Protein function prediction
  - ...

- Image annotation (Harchaoui & Bach, 2007)
  - Fusing information from different aspects of images
Image annotation

• Corel14: 1400 *natural images* with 14 classes
Performance on Corel14  
(Harchaoui & Bach, 2007)

- Histogram kernels ($H$)
- Walk kernels ($W$)
- Tree-walk kernels ($TW$)
- Weighted tree-walks ($wTW$)
- MKL ($M$)
Extension to trace norm minimization

• Consider learning linear predictor where covariates $X$ are rectangular matrices

• loading matrix $W$, and prediction $\text{tr} \ W^\top X$

• Assumption of low rank loading matrix:
  – Matrix completion (Srebro et al., 2004)
  – collaborative filtering (Srebro et al., 2004, Abernethy et al., 2006)
  – Multi-task learning (Argyriou et al., 2006, Obozinsky et al., 2007)

• Equivalent of the $\ell_1$ norm: \textbf{trace norm} $= \text{sums of singular values}$

• Do we actually get low-rank solutions?
  – Necessary and sufficient consistency conditions (Bach, 2007)
  – Extension of the group Lasso results.
Conclusion

• Analysis of sparsity behavior of the group lasso
  – infinite dimensional groups $\Rightarrow$ MKL
  – Adaptive version to define appropriate weights

• Current work:
  – Analysis for other losses
  – Consider growing number of groups
  – Analysis when consistency condition not satisfied
  – non parametric group lasso: universal consistency?
  – Infinite dimensional extensions of trace norm minimization