Learning with Clustered Penalties

Vincent Roulet

PhD Advisor: Alexandre d'Aspremont
Collaborators: Francis Bach, Fajwel Fogel
Original paper : Learning with Clustered Penalties, ArXiv:1506.04908

Project FACTORY, Toulouse, 05/2017

A practical problem

- Predict rating of a movie from its review
- Information: histogram of the occurrence of words
- Can be compressed: group synonyms for the task and predict influence of each group
- Problem: Find best groups for the task

Theoretical Motivation

- Alternative to sparse optimization
  - Sparse: Select variables
  - Here: Group variables
- **Same idea:**

  Constrain optimization to get
  compressed information for the task

- Other applications:
  $\rightarrow$ Find group of genomes that explain some phenotype
  $\rightarrow$ Select band of frequencies of a signal and not isolated frequencies (Long term goal...)

Plan

Classical regression task

$$\min_{w} \quad \frac{1}{n}\sum_{i=1}^{n} l(w; x_i, y_i) + \lambda\|w\|_2^2 = L(w)$$

- $X = (x_1, \ldots, x_n)^T$ data points in $\mathbb{R}^d$
- $y = (y_1, \ldots, y_n)$ corresponding labels in $\mathbb{R}$
- $w \in \mathbb{R}^d$ is the prediction vector
- l is a loss that measures quality of the prediction
- $\lambda\|w\|_2^2$ is a regularization term (potentially zero)

The analysis focuses on least squares $l(w; x, y) = \frac{1}{2}(y - w^T x)^2$ s.t.

$$L(w) = \frac{1}{2n}\|y - Xw\|_2^2 + \lambda\|w\|_2^2$$

Modelization of the constraint

- Desired constraint
  - Partition d features in (at most) Q groups
  - Assign one weight per group
- Tools
  - Assignment matrix $Z \in \{0,1\}^{d \times Q}$ s.t.
    - $Z_{iq} = 1$ if variable i is in group q,
    - one variable is in exactly one group, i.e. $Z\mathbf{1} = \mathbf{1}$.
  - Vector of weights $v \in \mathbb{R}^Q$
- Constraint formulation on prediction vector w

$$w = Zv, \quad Z \in \{0,1\}^{d \times Q}, \quad Z\mathbf{1} = \mathbf{1}, \quad v \in \mathbb{R}^Q$$

Problem formulation

$$\min_{w,Z,v} \quad L(w)$$
$$\text{s.t.} \quad w = Zv, \quad Z \in \{0,1\}^{d \times Q}, \quad Z\mathbf{1} = \mathbf{1}$$

- Non-convex: $w = Zv$ and $Z \in \{0,1\}^{d \times Q}$
- Proposed approaches:
  - Convex relaxation of the constraints
  - Non-convex projected gradient with statistical analysis
  - Convex penalization ?

Plan

Modelization

Proposed Resolution

- Convex relaxation
- Projected gradient with statistical analysis
- Convex penalization ?

Empirical Results

Extensions and future directions

Simplification for least squares

- For least square loss, analytical minimization in v possible

$$
\begin{aligned}
&\min_{Z,v} \frac{1}{2n}\|y - XZv\|_2^2 + \lambda\|Zv\|_2^2 \\
=&\min_{Z} \frac{1}{2n} y^T \left(I + \frac{1}{n\lambda} XZ(Z^T Z)^{-1} Z^T X^T\right)^{-1} y \\
=&\min_{M} \phi(M)
\end{aligned}
$$

  where $M = Z(Z^T Z)^{-1} Z^T$ is the normalized equivalence matrix of Z
- M encodes the partition

$$
M_{ij} = \begin{cases} \frac{1}{s_q} \text{if both } i, j \text{ are in group } q \text{ of size } s_q \\ 0 \text{otherwise} \end{cases}
$$

Convex relaxation strategy

- **Setting:**
  - ϕ convex in M
  - But set $\mathcal{M}$ of normalized equivalence matrices not convex (discrete set)
- **Strategy:**
  - Relax problem by optimizing on the convex hull of $\mathcal{M}$
  - Get a feasible solution Z from relaxation solution

Conditional gradient idea

- Classical constraint convex optimization use projection steps
  $\rightarrow$ Potentially costly or not possible
  $\rightarrow$ While linear minimization on the constraint sometimes easy
- Formal setting

$$\begin{array}{ll} \min_x & f(x) \\ \text{s.t.} & x \in Q \end{array}$$

  where f and Q convex
- Access to linear minimization oracle

$$\arg\min_{s \in Q} \langle y, s \rangle \quad \text{for every} \quad y \in Q$$

Conditional gradient algorithm

- Algorithm

$$\begin{aligned} x_0 &\in Q \\ s_t &= \arg\min_{s \in Q} \langle \nabla f(x_t), s \rangle \\ x_{t+1} &= x_t + \alpha_t (s_t - x_t) \end{aligned}$$

  where $\alpha_t \in [0, 1]$ is the stepsize
- Convergence in $O(1/t)$ for f smooth and convex

Application to convex relaxation

- Here $\mathcal{M}$ forms the extreme points of hull($\mathcal{M}$), so for a given $M \in \mathcal{M}$

$$\arg \min_{N \in \text{hull}(\mathcal{M})} \langle N, \nabla \phi(M) \rangle = \arg \min_{N \in \mathcal{M}} \langle N, \nabla \phi(M) \rangle$$

- Using that $\nabla \phi(M) \succeq 0$, this is k-means in one dimension (solved exactly by dynamic programming)

$\rightarrow$ Conditional gradient can be applied !

- Projection on feasible Z is also given by a k-means in one dimension
- **Problem** : Computation of $\nabla \phi(M)$ is very costly...

Plan

Modelization

Proposed Resolution

Convex relaxation

Projected gradient with statistical analysis

Convex penalization ?

Empirical Results

Extensions and future directions

Projection on set of constraints

- Projection problem for a given $w \in \mathbb{R}^d$

$$\begin{aligned} \min_{Z,v} \quad & \|w - Zv\|_2^2 \\ \text{s.t.} \quad & Z \in \{0,1\}^{d \times Q}, Z\mathbf{1} = \mathbf{1} \end{aligned}$$

- A closer look

$$\min_{v,\mathcal{P}} \sum_{q=1}^{Q} \sum_{i \in \mathcal{P}_q} (w_i - v_q)^2$$

where $\mathcal{P} = \mathcal{P}_1, \ldots \mathcal{P}_Q$ is a partition of d elements in Q groups

- We recognize k-means in one dimension
- Dynamic program solves it exactly in $O(d \log(d))$ computations

Projected Gradient descent

- Scheme

$$w_0 = 0$$
$$w_{t+1} = P_Q(wt - \gamma \nabla L(w_t))$$

  where P_Q is the projection on the set of constraints, i.e. k-means in one dimension into Q groups.

- Problem non-convex
  $\rightarrow$ no guarantee of convergence to a global optimum.
- Similar to Iterative Hard Thresholding used in sparse optimization
  $\rightarrow$ Potential statistical analysis

Statistical analyis approach

Assume

- $y = Xw_* + \eta$ with η Gaussian noise
- w_* satisfies constraints
- observations $x_1, \ldots, x_n$ were randomly chosen (subgaussian vectors)

Show that

- the algorithm converges to w_*
- need less samples than number of features
  $\rightarrow$ imposed constraint is able to capture the compressed information

Statistical analysis results

Proposition

Projected gradient descent (with $\gamma = 1$) converges then to w_* up to statistical precision if

$$n = \Omega(D) \quad \text{and} \quad n = \Omega(\log(N))$$

where

- D is the compressed dimension
- N is the complexity of the underlying combinatorial problem

Here $D = Q$ and we assumed $Q \ll d$
However $N \geq Q^{d-Q}$, so we still need

$$n = \Omega(d)$$

In comparison for sparse vectors $N \approx d^k$ such that $n \approx k \log(d)$ is sufficient.

Plan

Modelization

Proposed Resolution

- Convex relaxation
- Projected gradient with statistical analysis
- Convex penalization ?

Empirical Results

Extensions and future directions

Convex penalization ?

- **Idea:** Transform combinatorial problem into a convex penalty
- Define

$$\begin{aligned} F : w \to \quad & \mathbf{Card}(G \subset [\![1, d]\!] : \forall i, j \in G,\ w^{(i)} = w^{(j)}) \\ = \quad & \text{number of group of identical features of } w \end{aligned}$$

- Compute norm associated to F by taking the lower convex homogeneous envelope of

$$F(w) + \frac{1}{2}\|w\|_2^2$$

- **Problem:** Resulting norm is not computable neither is its proximal operator

Plan

Synthetic experiments setting

- $y = Xw_* + \eta$ with $\eta \sim \mathcal{N}(0, \sigma^2)$
- w_* composed of $Q = 5$ group of identical features among $d = 100$
- **Goal:**
  - Test robustness of our method with number of samples n and level of noise σ
  - Measure $\|w_* - \hat{w}\|_2$ with $\hat{w}$ estimated vector

Synthetic experiments setting

- Compare our model optimized with
  - Convex relaxation (CG)
  - Projected gradient on non-convex problem (PG)
  - Convex relaxation followed by non-convex refinement (CGPG)

  to basic models:
  - Least-squares (LS)
  - Least-squares followed by a k-means (LSK)
  - OSCAR penalty (enforces cluster in some way) (OS)

  and oracle given the partition
  - Least square solution given the initial clusters of variable (Oracle)

Synthetic experiments results for n increasing

	$n = 50$	$n = 75$	$n = 100$	$n = 125$	$n = 150$
Oracle	0.16±0.06	0.14±0.04	0.10±0.04	0.10±0.04	0.09±0.03
LS	61.94±17.63	51.94±16.01	21.41±9.40	1.02±0.18	0.70±0.09
LSK	62.93±18.05	57.78±17.03	10.18±14.96	0.31±0.19	0.19±0.12
PG	63.31±18.24	52.72±16.51	5.52±14.33	**0.14**±0.09	**0.09**±0.04
CG	61.81±17.78	52.59±16.58	17.24±13.87	1.20±1.38	1.05±1.37
CGPG	62.29±18.15	**50.15**±17.43	**0.64**±2.03	0.15±0.19	0.17±0.53
OS	**61.54**±17.59	52.87±15.90	11.32±7.03	1.25±0.28	0.71±0.10

Table: Measure of $\|w_* - \hat{w}\|_2$, the l_2 norm of the difference between the true vector of weights w_* and the estimated ones $\hat{w}$ along number of samples n.

Synthetic experiments results for σ increasing

	$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.5$	$\sigma = 1$
Oracle	0.86±0.27	1.72±0.54	8.62±2.70	17.19±5.43
LS	7.04±0.92	14.05±1.82	70.39±9.20	140.41±18.20
LSK	1.44±0.46	2.88±0.91	19.10±12.13	48.09±27.46
PG	**0.87**±0.27	**1.74**±0.52	**9.11**±4.00	26.23±18.00
CG	23.91±36.51	122.31±145.77	105.45±136.79	155.98±177.69
CGPG	1.52±3.13	140.83±710.32	17.34±53.31	**24.80**±16.32
OS	14.43±2.45	18.89±3.46	71.00±10.12	140.33±18.83

Table: Measure of $\|w_* - \hat{w}\|_2$, the l_2 norm of the difference between the true vector of weights w_* and the estimated ones $\hat{w}$ along level of noise σ.

Real problem setting

- Predicting ratings of movies from their reviews
- Dataset contains $n = 5006$ documents and vocabulary of $d = 5623$ words

LS	LSK	PG	CG	CGPG	OS
1.51±0.06	1.53±0.06	1.52±0.06	1.58±0.07	1.49±0.08	1.47±0.07

Table: 100 $\times$ mean square errors for predicting movie ratings associated with reviews.

Plan

Modelization

Proposed Resolution

- Convex relaxation
- Projected gradient with statistical analysis
- Convex penalization ?

Empirical Results

Extensions and future directions

Extensions and future directions

- **Mix sparsity and clustering:**
  - Done by modifying dynamic programming of K-means in one dimension
- **Use formulation for other problems:**
  - Supervised clustering of samples
  - Clustered multitask
- **Future directions:**
  - Impose size of clusters to alleviate underlying combinatorial problem

Iterative Hard Thesholding (IHT)

- Least square regression with sparsity constraints

$$\min_{w} \quad \frac{1}{2n}\|y - Xw\|_2^2$$
$$\text{s.t.} \quad \|w\|_0 \leq k$$

  where $\|w\|_0 = \text{Card}(i : w^{(i)} = \neq 0)$
  Remark that
  $\|w\|_0 \leq k \quad \Leftrightarrow \quad w = Zv, \quad Z \in \{0,1\}^{d \times k} \quad Z^T \mathbf{1} = \mathbf{1}$
- Projecting on the constraint set is taking k largest absolute coordinates
- Corresponding projected gradient descent is IHT

Statistical analysis sketch

- Constraint set is a union of spaces
  $U_Z = \{w : w = Zv, v \in \mathbb{R}^Q\}$ with Z an assignment matrix
- Projected gradient descent is then a point-fix kind of algorithm, precisely the iterates satisfy

$$\|w_t - w_*\|_2 \leq \rho^t \|w_*\|_2 + \frac{1-\rho^t}{1-\rho}\nu\|\eta\|_2$$

  where

$$\rho = 2 \max_{U \in \mathcal{E}} \|I - \frac{1}{n}\Pi_U^T X^T X \Pi_U\|_2 \quad \text{and} \quad \nu = \frac{2}{n} \max_{U \in \mathcal{E}} \|X\Pi_U\|_2$$

  Π_U is any orthonormal basis of the subspace U
  and $\mathcal{E} = \{U_{Z_1} + U_{Z_2} + U_{Z_3} : Z_i \quad \text{assignement matrix}\}$
- Study of the largest and smallest singular values of X on subspaces $U \in \mathcal{E}$ for X composed of subgausssian vectors