DISCML 2013

Novel Sparse Modeling by L2 + L0 Regularization

Hidekazu Oiwa, Issei Sato, Hiroshi Nakagawa
The University of Tokyo

Problem

- Regularized Empirical Risk Minimization

$$F(\mathbf{w}) = \min_{\mathbf{w}} \sum_{\gamma=1}^{t} \underbrace{\ell_\gamma(\mathbf{w})}_{\text{loss function}} + \underbrace{r(\mathbf{w})}_{\text{regularization}}$$

- What type of regularization should we use?
  - L1, L0, elastic net, or other structured ones?
  - Sparsity-inducing effect or Grouping effect?

L0 Elastic Net

$$r(\mathbf{w}) = \lambda(\pi\|\mathbf{w}\|_2^2 + (1-\pi)\|\mathbf{w}\|_0)$$

L2 + L0 regularization

- Pros
  - Strong sparsity-inducing ability (by L0)
  - Generalization ability (by L2)
- Cons
  - Non-convex problem due to L0 norm
  - NP-hard (from viewpoints of discrete optimization)

Our contributions

$$r(\mathbf{w}) = \lambda(\pi\|\mathbf{w}\|_2^2 + (1-\pi)\|\mathbf{w}\|_0)$$

- We develop
  - convex solver for L0 elastic net
  - theoretical guarantee as simultaneous optimization of feature selection and parameter optimization
- Experiment guarantees that L0 elastic net
  - produces more compact and predictive model than conventional ones

Why L0 elastic net?

OSCAR

Elastic net

Lasso

L_0 elastic net

L_0

Correlation-oriented ⟷ Sparse-oriented

capture all effective features
optimization easier
redundant predictive model

construct minimal feature set
optimization harder
compact predictive model

We will show…

L0 elastic net can produce more compact predictive model than L1
L0 elastic net can be optimized more easily than L0

Comparison of Regularizations

	Our method	COR [10, 12]	Lasso [1]	Group [13]	L_0
Grouping effect		√		√	
Noise reduction	√	√	√	√	
Redundancy reduction	√		√		√
No prior knowledge	√	√	√		√
One-step optimality	√				√

- L0 elastic net
  - has strong noise and redundancy reduction effect
  - does not need prior knowledge
  - has one-step optimality

Our contribution!

Dual Decomposition Solver

- Decompose original problem into two sub-problems

$$\min_{\mathbf{u},\mathbf{v}} \sum_{\gamma=1}^{t} \ell_\gamma(\mathbf{u}) + r(\mathbf{v}) \quad \text{where} \quad \mathbf{u} = \mathbf{v}$$

- Lagrange relaxation

$$L(\mathbf{z}) = \min_{\mathbf{u},\mathbf{v}} \sum_{\gamma=1}^{t} \ell_\gamma(\mathbf{u}) + r(\mathbf{v}) + \mathbf{z}^T(\mathbf{u} - \mathbf{v})$$

$$= \min_{\mathbf{u}} \left(\sum_{\gamma=1}^{t} \ell_\gamma(\mathbf{u}) + \mathbf{z}^T\mathbf{u} \right) + \min_{\mathbf{v}} \left(r(\mathbf{v}) - \mathbf{z}^T\mathbf{v} \right)$$

each subproblem can be solved efficiently!

Sub-problems

- Loss part is similar to risk minimization problem

$$\min_{\mathbf{u}} \left(\sum_{\gamma=1}^{t} \ell_\gamma(\mathbf{u}) + \mathbf{z}^T \mathbf{u} \right)$$

If loss functions are convex,
it is solvable by GD, SGD, lbfgs…

- Regularization part is…

$$\mathbf{v}_t = \operatorname*{argmin}_{\mathbf{v}} \left(\lambda_2 \|\mathbf{v}\|_2^2 + \lambda_0 \|\mathbf{v}\|_0 - \mathbf{z}_t^T \mathbf{v} \right)$$

$$\lambda_0 = \lambda(1-\pi)$$

$$\lambda_2 = \lambda\pi$$

Fortunately, this problem is solvable by closed form!

Regularization part solution

- L0 elastic net case

$$v_t^{(i)} = \begin{cases} 0 & (z_t^{(i)})^2 \leq 4\lambda_0\lambda_2 \\ \dfrac{z_t^{(i)}}{2\lambda_2} & \text{otherwise} \end{cases}$$

- This technique can be applied to other regularizations
  - L1 regularization, elastic net, etc.
  - Unfortunately, L0 regularization is not solvable :-(

Algorithm Description

- Algorithm
  - Iterative update of primal and dual parameters
  - 1. Update primal parameters $\mathbf{u}_t, \mathbf{v}_t$
  - 2. Update dual parameter $\mathbf{z}_t$ by subgradient method
  - Convergence is guaranteed
- Optimality Condition

Theorem 1 *(Koo et al. [17]) If $\mathbf{u}_k = \mathbf{v}_k$ is satisfied at some k, $\mathbf{u}_k = \mathbf{v}_k = \mathbf{w}_k$ is a solution of formula (2). That is,*

$$L(\mathbf{z}_k) = F(\mathbf{w}_k) = F(\mathbf{w}^*) \,, \tag{8}$$

is satisfied.

Theoretical Guarantee

Theorem 2 *Let us assume that* $\mathbf{w}^*$ *is an optimal weight vector for the problem with* L_0 *elastic net.*

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left\{ F(\mathbf{w}) = \sum_{\tau=1}^{t} \ell_\tau(\mathbf{w}) + r(\mathbf{w}) \right\}$$

$$r(\mathbf{w}) = \lambda \left(\pi \|\mathbf{w}\|_2^2 + (1-\pi)\|\mathbf{w}\|_0 \right) . \quad (12)$$

Let us define S_0 *as the index set where the component value is 0 in* $\mathbf{w}^*$. *In this case,* $\mathbf{w}^*$ *is one of the most compact optimal weight vectors for the following problem.*

$$G(\mathbf{w}) = \sum_{\tau=1}^{t} \ell_\tau(\mathbf{w}) + \lambda\pi \|\mathbf{w}\|_2^2$$

$$s.t. \quad \forall i \in S_0 \quad \hat{w}^{(i)} = 0 . \quad (13)$$

- This Theorem guarantees
  - feature subset is minimal to predict as same accuracy
  - parameters are optimal and no need to re-estimate

Experiments

- Synthetic Regression Data
  - 6 features, 2 groups
  - Feature are highly correlated in same group
  - Best feature subset: x_1, x_4

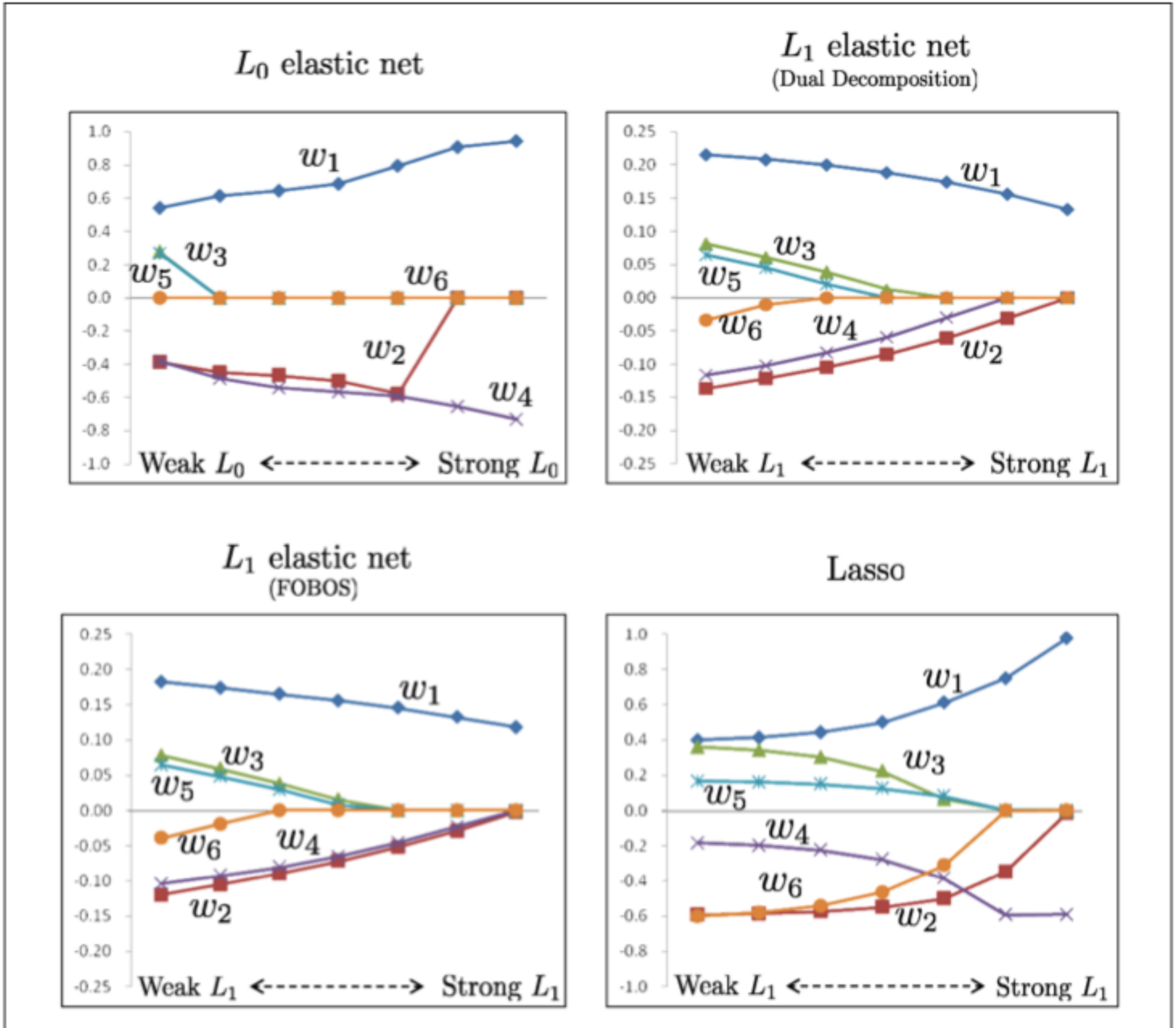
Input

$$Z_1, Z_2 \sim U(0, 20) \qquad \epsilon_i \sim \mathcal{N}(0.0, 0.05)$$

$$x_1 = Z_1 + \epsilon_1, \quad x_2 = -0.7Z_1 + \epsilon_2,$$
$$x_3 = 0.5Z_1 + \epsilon_3, \quad x_4 = Z_2 + \epsilon_4,$$
$$x_5 = -0.7Z_2 + \epsilon_5, \quad x_6 = 0.5Z_2 + \epsilon_6,$$

Output

$$y_i \sim \mathcal{N}(Z_1 - 0.6Z_2, 1.0)$$


L_0 elastic net
L_1 elastic net (Dual Decomposition)
L_1 elastic net (FOBOS)
Lasso
w_1
w_2
w_3
w_4
w_5
w_6
Weak L_0 <-----------> Strong L_0
Weak L_1 <-----------> Strong L_1

Result and Future Work

- Experiment guarantees
  - L0 elastic net outperforms conventional regularization methods in both feature selection and parameter estimation
- Future Work
  - How to determine hyperparameters?
  - More convincing experiments

$$r(\mathbf{w}) = \lambda(\pi\|\mathbf{w}\|_2^2 + (1-\pi)\|\mathbf{w}\|_0)$$