TO, AS: arXiv:1902.10375

Cross validation in sparse linear regression with piecewise continuous nonconvex penalties and its acceleration

Tomoyuki Obuchi[1] and Ayaka Sakata[2]

Dept. of Math. and Comp. Sci., Tokyo Tech.[1]
The Institute of Statistical Mathematics[2]

Linear Regression

- Penalized linear regression

$$\hat{\boldsymbol{x}}(\eta) = \arg\min_{\boldsymbol{x}} \left\{ \frac{1}{2} ||\boldsymbol{y} - A\boldsymbol{x}||_2^2 + \underbrace{J(\boldsymbol{x};\eta)}_{\text{Penalty}} \right\}$$

- Representative Penalty

  - ℓ_p norm

$$J(\boldsymbol{x}; \eta = \lambda) = \lambda ||\boldsymbol{x}||_p^p$$

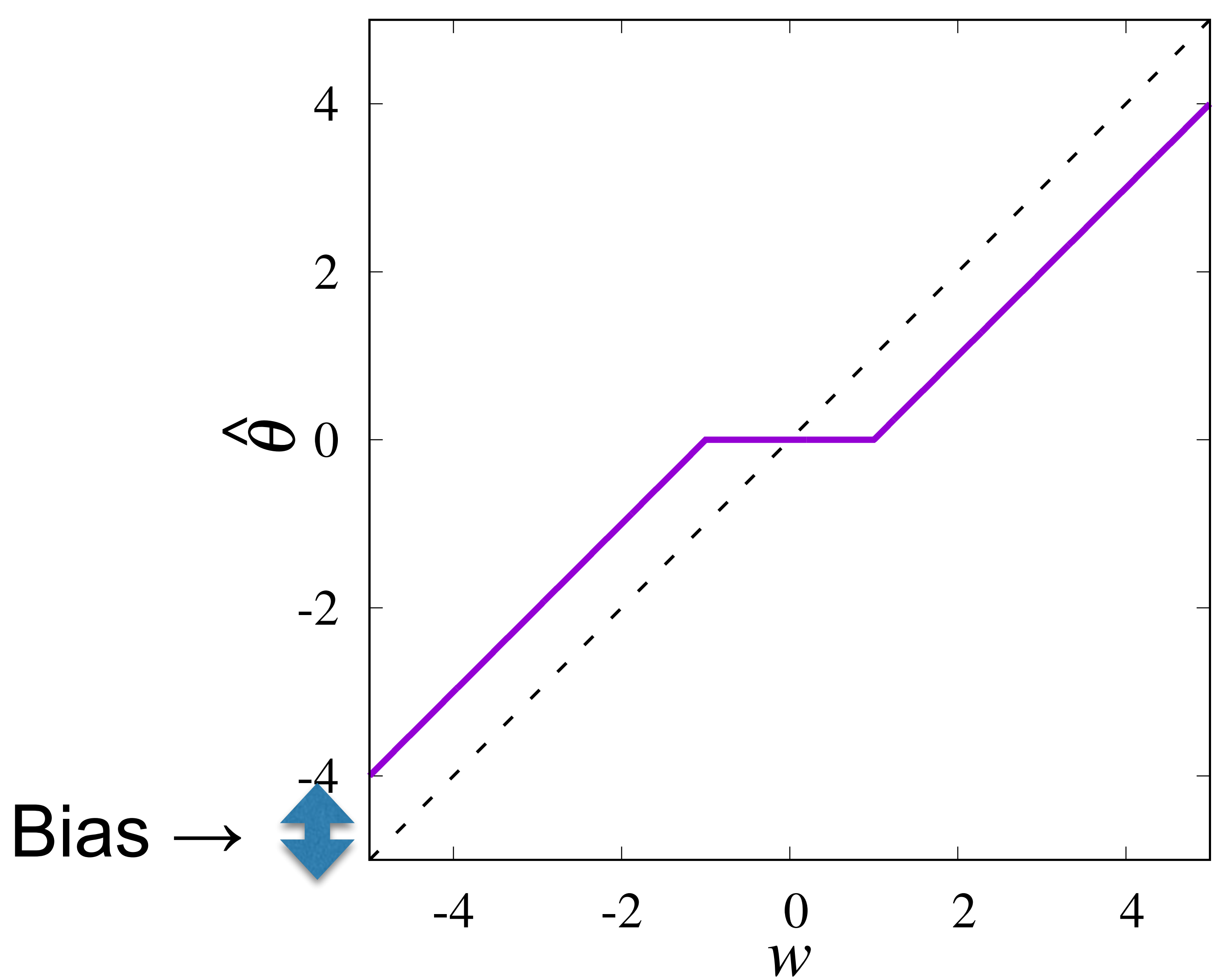
    - $p \geq 1$: convex
    - $p \leq 1$: sparsity-inducing

→ p=1 is nice for variable selection (LASSO)

Statistical bias in LASSO

- LASSO for 1-dimensional estimation

$$\hat{\theta} = \arg\min_{\theta} \left\{ \frac{1}{2\sigma^2}(\theta - w)^2 + \lambda|\theta| \right\}$$



- p<1 can reduce bias but…
  - Nonconvex → possible local minima
  - Noncontinuity → algorithmic instability

↓

Piecewise continuous nonconvex penalty (PCNP)

- Nonconvex, but estimator is continuous

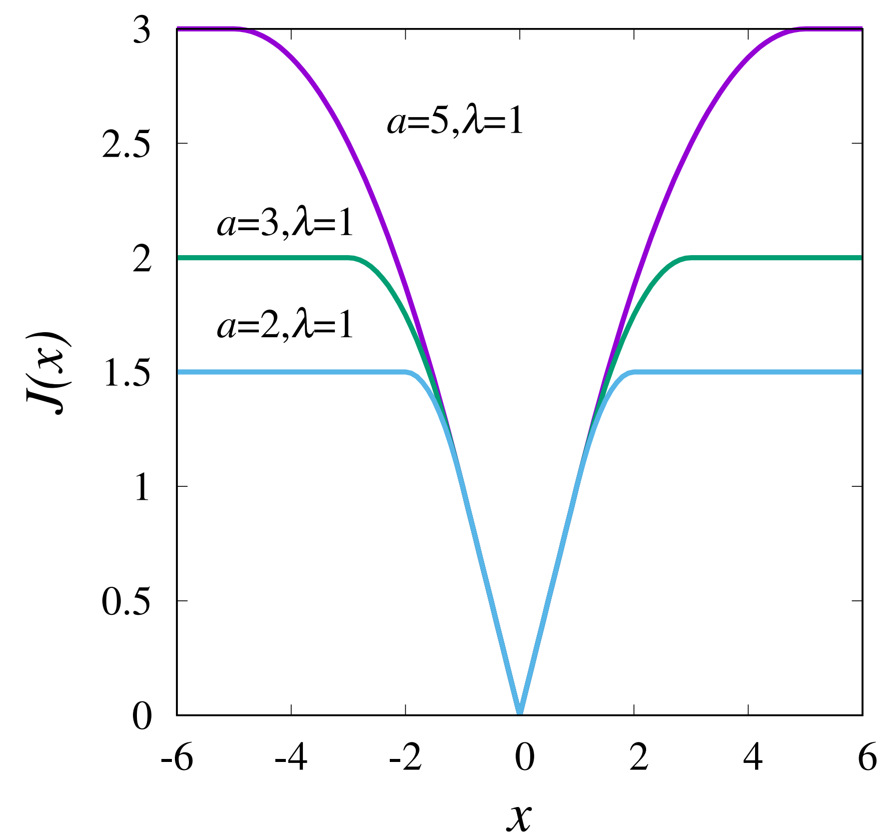
- Two representatives of PCNP
  - Smoothly Clipped Absolute Deviation (SCAD) penalty
  - Minimax Concave Penalty (MCP)

We hereafter focus only on SCAD

SCAD estimator

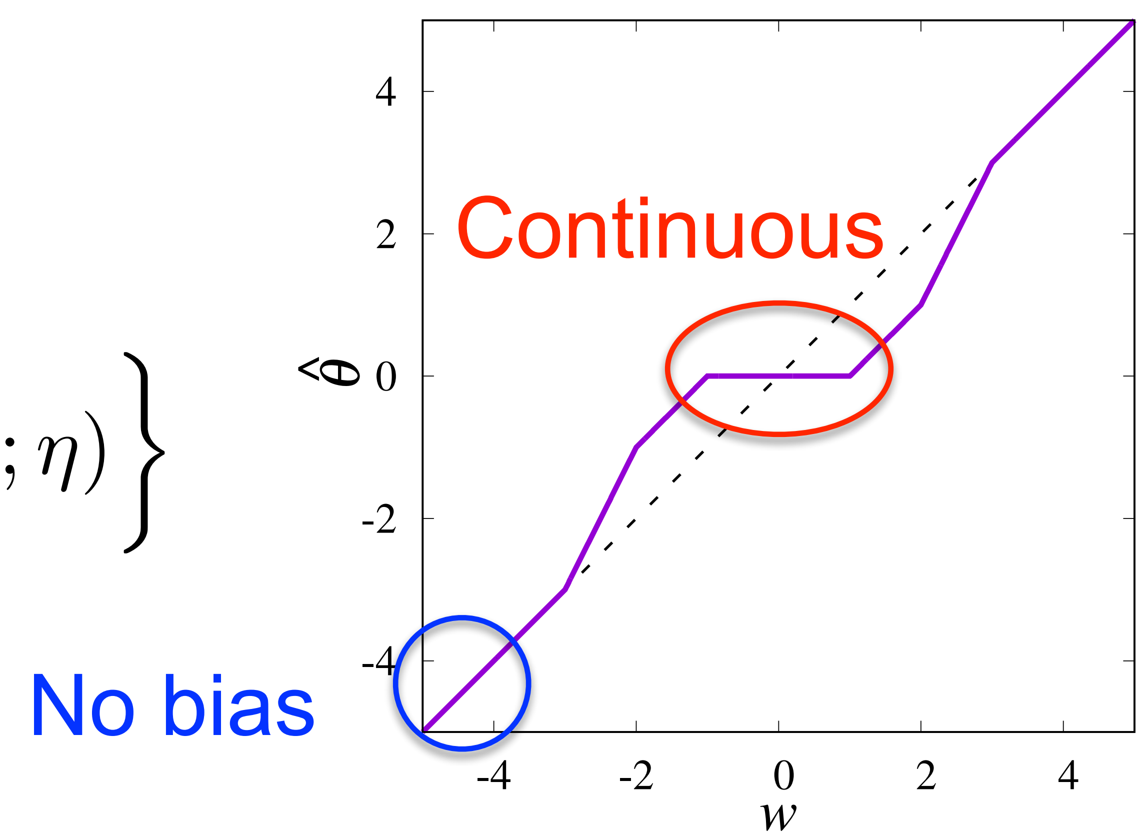
- SCAD penalty ($\eta = \{a, \lambda\}$)

$$J(\theta;\eta) = \begin{cases} \lambda|\theta| & (|\theta| \le \lambda) \\ -\dfrac{\theta^2 - 2a\lambda|\theta| + \lambda^2}{2(a-1)} & (\lambda < |\theta| \le a\lambda) \\ \dfrac{(a+1)\lambda^2}{2} & (|\theta| > a\lambda) \end{cases}$$



- SCAD estimator
  - E.g. 1D estimator

$$\hat{\theta} = \arg\min_{\theta} \left\{ \frac{1}{2\sigma^2}(\theta - w)^2 + J(\theta;\eta) \right\}$$



Our Contributions

- Clarifying the emergence region of local minima
  - Phase transition (w. replica symmetry breaking)
- Quantitative analysis of reconstruction performance
  - SCAD outperforms LASSO in weak noise region
- Developing an approximate CV formula
  - Fast CV becomes possible
  - A method to avoid unstable parameter region

Contents

1. Analytical performance analysis in simulated dataset

2. Approximate CV formula

3. Numerical experiments

Contents

1. Analytical performance analysis in simulated dataset

2. Approximate CV formula

3. Numerical experiments

Problem Setting

- Generative process

crucial assumptions for analysis

$$\boldsymbol{y} = A\boldsymbol{x}_0 + \boldsymbol{\Delta}$$

$$\boxed{A_{\mu i} \sim \mathcal{N}(0, N^{-1})} \quad \Delta_i \sim \mathcal{N}(0, \sigma_\Delta^2)$$

$$x_{0i} \sim (1-\rho_0)\delta(x_{0i}) + \rho_0 \mathcal{N}(0, \sigma_x^2)$$

all i.i.d.

- Quantities of interest

$$\epsilon_y = \frac{1}{2M}||\boldsymbol{y} - \hat{\boldsymbol{y}}||_2^2$$: Output MSE $\quad \hat{\boldsymbol{y}} = A\hat{\boldsymbol{x}}$

$$\epsilon_x = \frac{1}{2N}||\boldsymbol{x}_0 - \hat{\boldsymbol{x}}||_2^2$$: Input MSE

TP, FP : True and False positive rates of support $S = \{i | x_{0i} \neq 0\}$

Investigate typical values of these in high-dimensional limit

$$N \to \infty, (\alpha = M/N = O(1))$$

Stat. Mech. Formulation

- Hamiltonian, Boltzmann distribution, Partition function

$$\mathcal{H}(\boldsymbol{x}) = \frac{1}{2}||\boldsymbol{y} - A\boldsymbol{x}||_2^2 + J(\boldsymbol{x};\eta)$$

$$P(\boldsymbol{x}) = \frac{1}{Z}e^{-\beta\mathcal{H}(\boldsymbol{x})} \to \delta(\boldsymbol{x} - \hat{\boldsymbol{x}}(\eta|\boldsymbol{y}, A)), \ (\beta \to \infty)$$

$$Z = \int d\boldsymbol{x} e^{-\beta\mathcal{H}(\boldsymbol{x})}$$

Solution of the original problem

- Computing “free energy” or moment-generating function *f*(*β*)

$$-\beta f(\beta) = \frac{1}{N}\left[\log Z\right]_{\boldsymbol{y}, A}$$

Average w.r.t. *y* and *A*

- Any quantity of interest can be computed from *f*(*β*)

However, the average w.r.t. y and A is unperformable…

←*Replica Method (with replica symmetric assumption)*

Equations to be solved

- Replica symmetric (RS) free energy

$$f(\beta \to \infty) = \underset{\Omega,\tilde{\Omega}}{\mathrm{Extr}} \left\{ \frac{Q - 2m + \rho_0 \sigma_x^2 + \alpha \sigma_\Delta^2}{2(1+\chi/\alpha)} + m\tilde{m} - \frac{\tilde{Q}Q - \tilde{\chi}\chi}{2} + \frac{\overline{\xi(\sigma;\tilde{Q})}}{2} \right\}$$

$$\Omega = \{Q, \chi, m\} \quad \tilde{\Omega} = \{\tilde{Q}, \tilde{\chi}, \tilde{m}\}$$

$$\xi(\sigma;\tilde{Q}) \equiv 2 \int Dz\ L(\sigma z;\tilde{Q}),$$

$$x^*(h;\tilde{Q}^{-1}) = \underset{x}{\arg\min} \left\{ \frac{\tilde{Q}}{2} x^2 - hx + J(x;\eta) \right\}.$$

$$L(h;\tilde{Q}) \equiv \min_x \left\{ \frac{\tilde{Q}}{2} x^2 - hx + J(x;\eta) \right\}.$$

$$TP = \int Dz\ \left| x^*(\sigma_+ z;\tilde{Q}^{-1}) \right|_0$$

$$FP = \int Dz\ \left| x^*(\sigma_- z;\tilde{Q}^{-1}) \right|_0$$

$$\epsilon_y = \frac{1}{2}\tilde{\chi}.$$

$$\epsilon_x = \frac{1}{2}\left(\rho_0 \sigma_x^2 - 2m + Q\right),$$

$$\int Dz(\cdots) \equiv \int_{-\infty}^{\infty} \frac{dz}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)(\cdots),$$

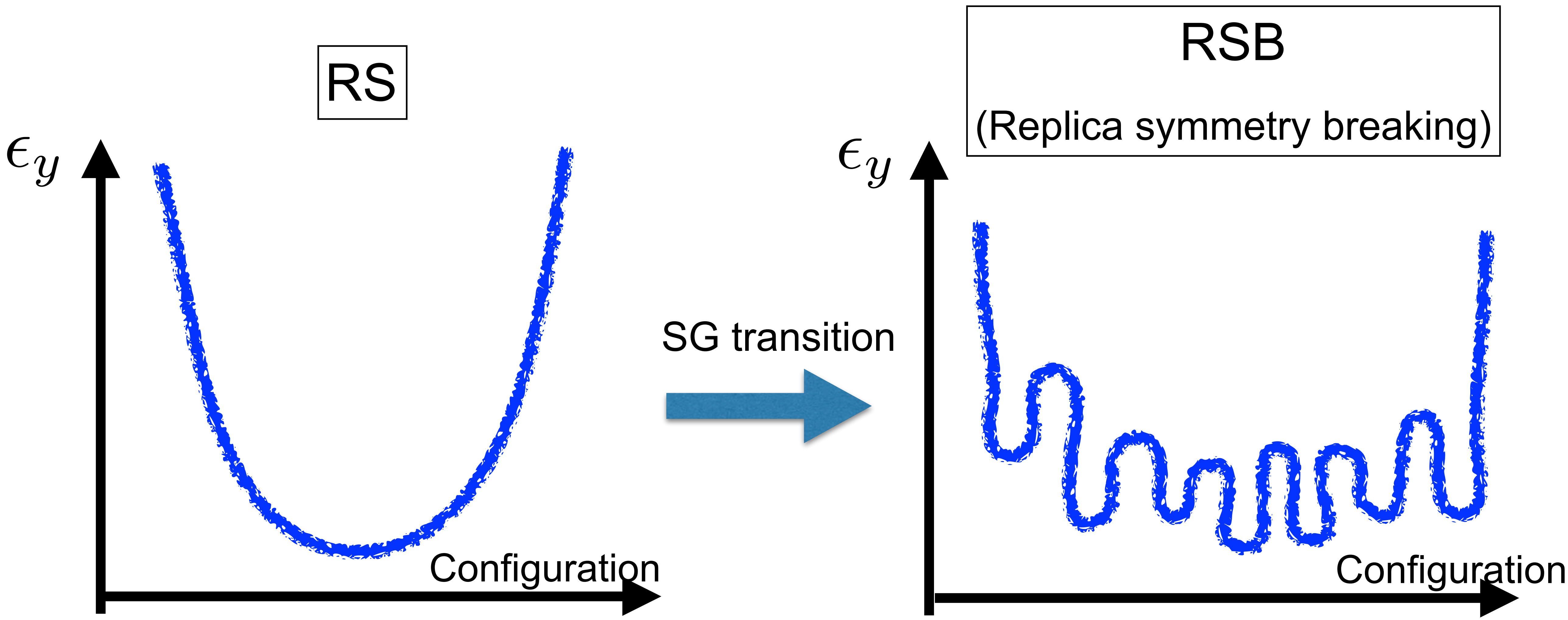
$$\sigma_- = \sqrt{\tilde{\chi}}, \quad \sigma_+ = \sqrt{\tilde{\chi} + \tilde{m}^2 \sigma_x^2}.$$

$$P(\sigma) = (1-\rho)\delta(\sigma - \sigma_-) + \rho\delta(\sigma - \sigma_+)$$

$$\overline{(\cdots)} = \sum_\sigma (\cdots) P(\sigma)$$

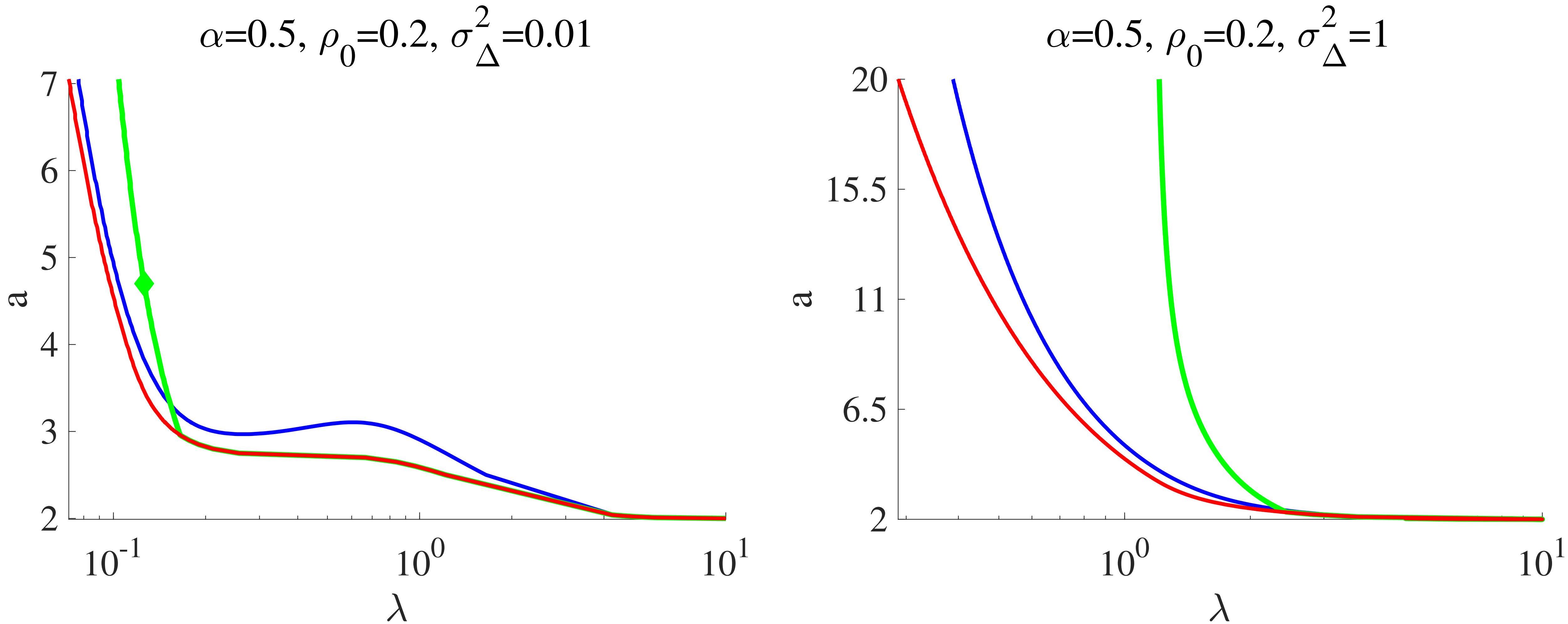
Stability and Multiple solutions

- RS solution is sometimes unstable
  - The instability can be signaled by a formula (not shown here)
    - Spin-glass transition or Almeida-Thouless (AT) instability



Exponentially many (w.r.t. *N*) local minima exist.

Phase diagrams



Green line: Minimum of input MSE for each λ

Green dot: Minimum of input MSE along the green line

Blue line: AT line (Above the line, our analysis is stable)

- (λ, a) that gives minimum input MSE is in stable region.
- For large noise, LASSO is sufficient.

ROC curve

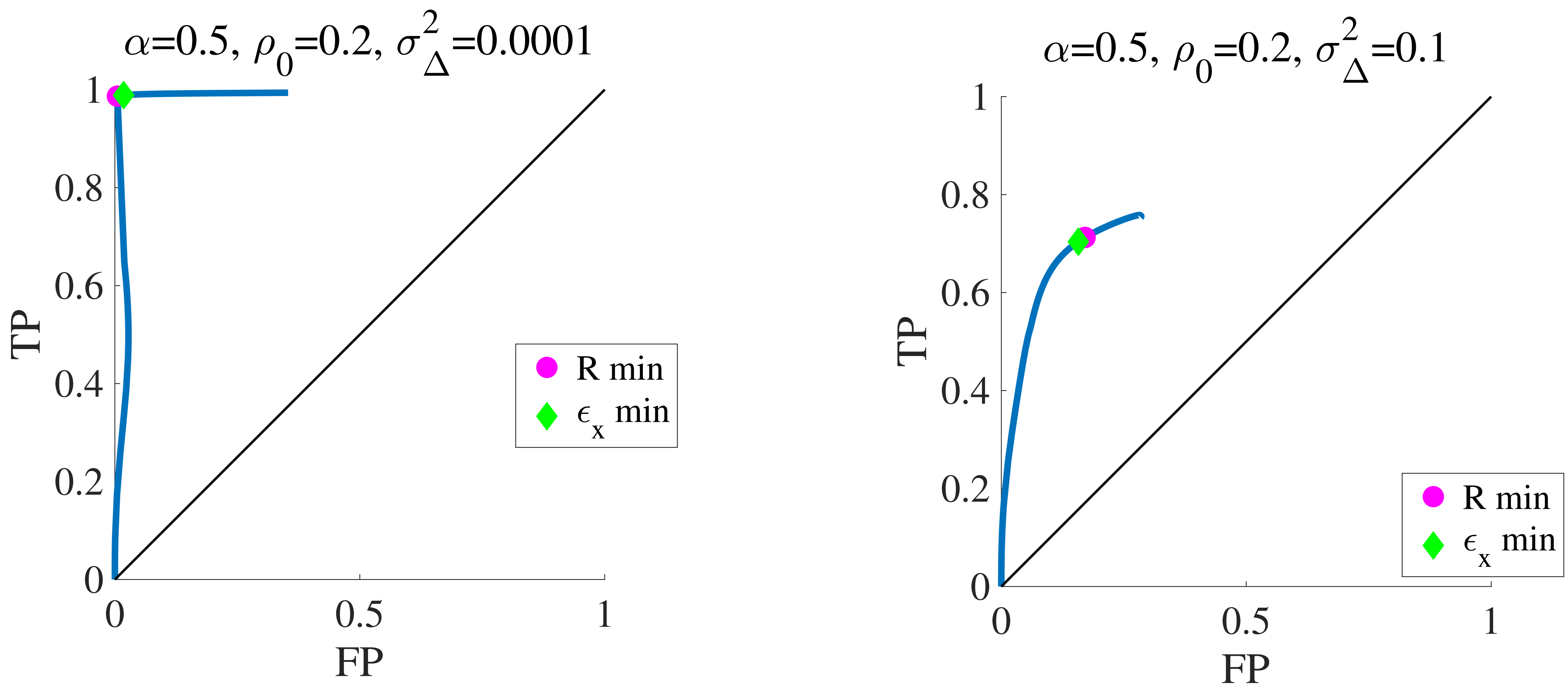
- Receiver operating characteristic (ROC) curve
  - Plot of TP against FP
- A criterion:
  "Optimal point" on ROC curve is the minimum of $R(\eta)$

$$R(\eta) = (TP(\eta) - 1)^2 + (FP(\eta) - 0)^2, \qquad \eta = \{\lambda, a\}$$

At the optimal point, the support recovery error is expected to be minimized.

- Here, we identify the optimal value of λ at a fixed value of a, and compare the value with that gives minimum of input MSE.

ROC curve



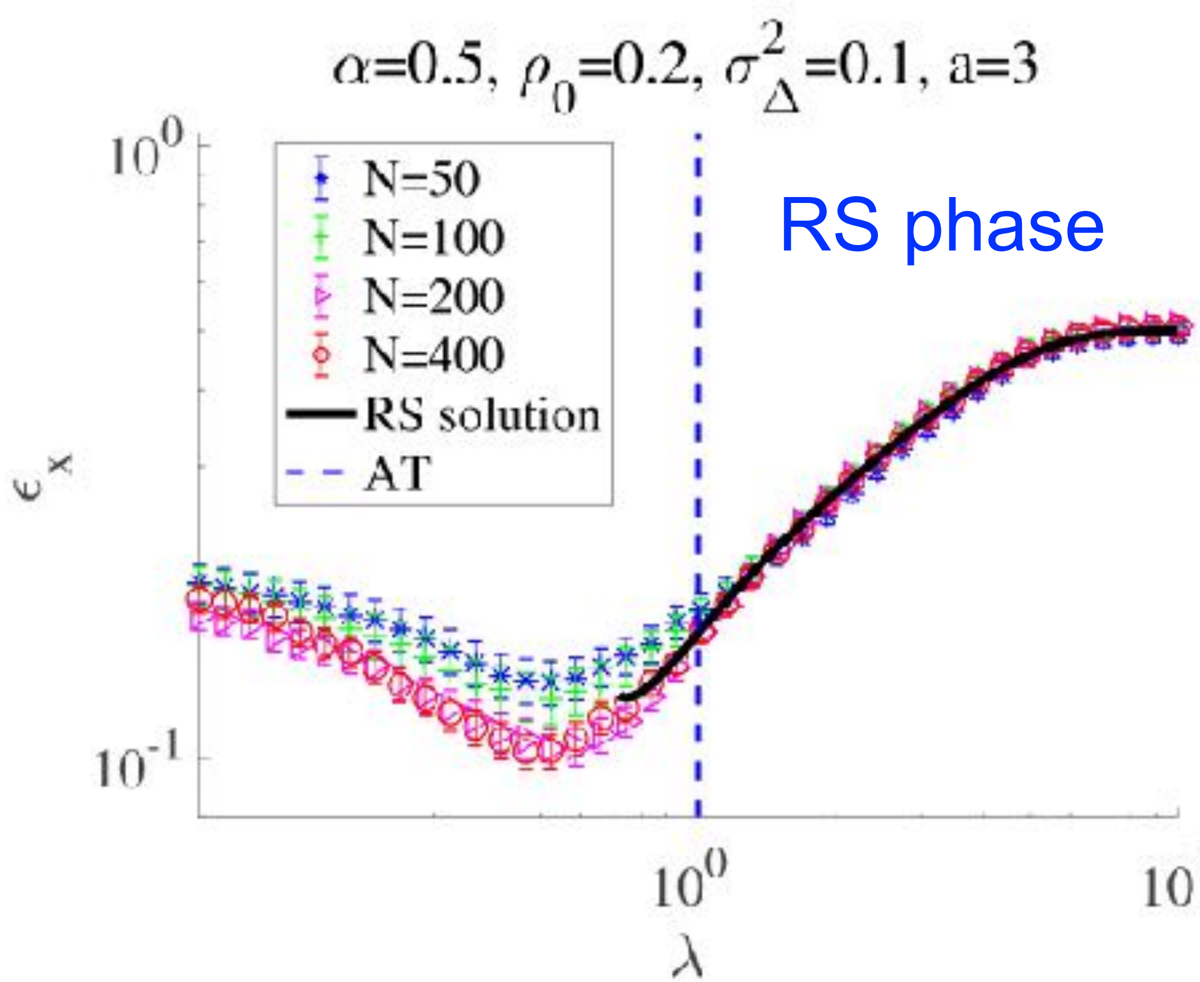
- Minimum locations of input MSE and R are close.

  This property is absent in LASSO [Obuchi and Kabashima, JSTAT (2016)]

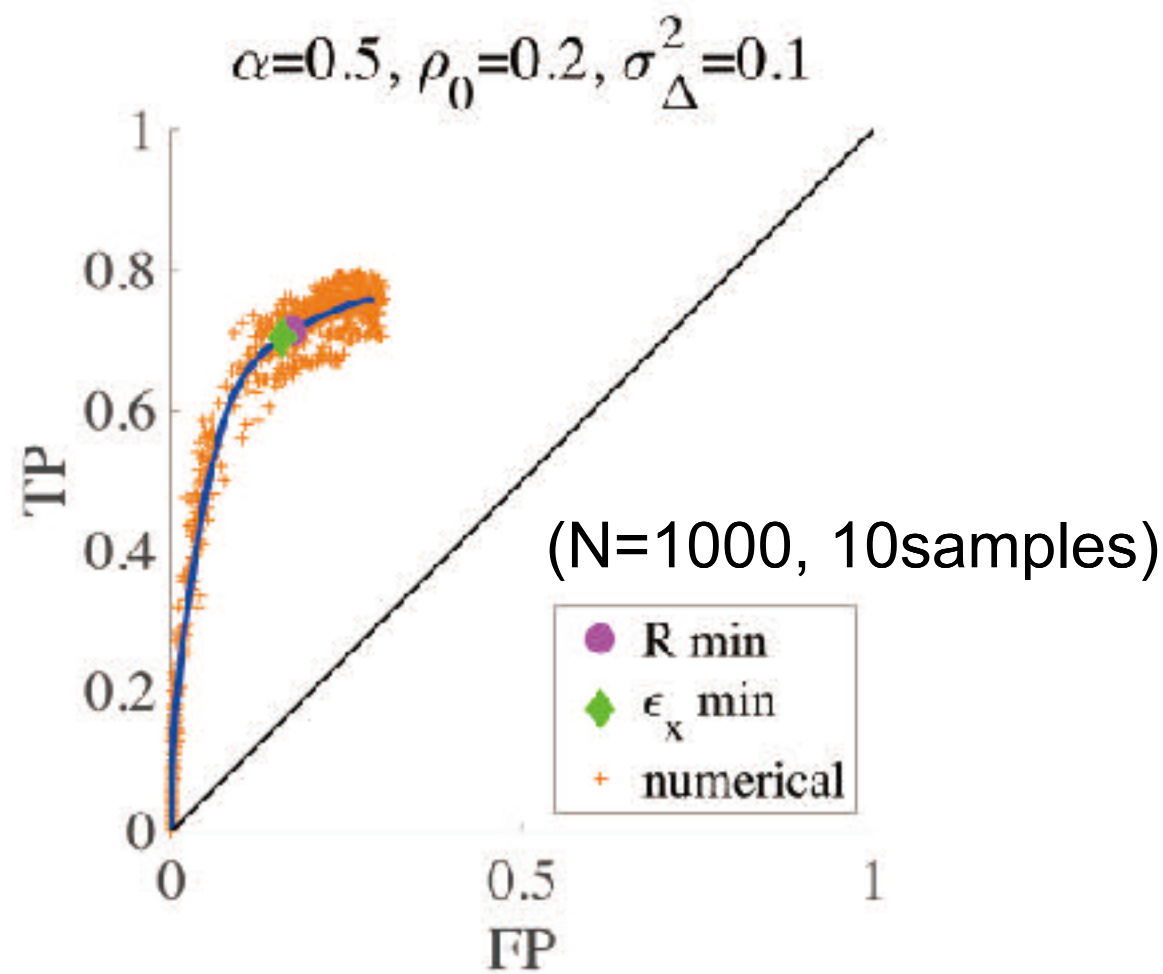
- Input MSE is unknown in general settings, but relates to Cross-validation (CV) error, hence we may minimize CV error to determine optimal support.

Verification of Theoretical Result

Input MSE



ROC curve



Analytically derived lines match to numerical simulation in RS phase.

Contents

LOOCV and Linear Approx.

Define: $\mathscr{H}(\mathbf{x}\,|\,D) \equiv \left\{ \frac{1}{2} \sum_{\mu} \left(y_\mu - \sum_i A_{\mu i} x_i \right)^2 + J(\mathbf{x};\eta) \right\}$

- Leave-one-out CV (LOOCV)

$$\hat{\mathbf{x}}^{\backslash \mu} = \arg\min_{\mathbf{x}} \mathscr{H}(\mathbf{x}\,|\,D^{\backslash \mu})$$

$$\epsilon_{\mathrm{LOO}}(\eta) = \frac{1}{2M} \sum_{\mu} \left(y_\mu - \sum_i A_{\mu i} \hat{x}_i^{\backslash \mu}(\eta) \right)^2$$ ← Large cost!

- Approximation: Expand $\mathscr{H}$ w.r.t. $\boldsymbol{d} = \hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{\backslash \mu}$

$$\mathscr{H}(\hat{\mathbf{x}}\,|\,D) - \mathscr{H}(\hat{\mathbf{x}}^{\backslash \mu}\,|\,D^{\backslash \mu}) \sim \sum_{\mu} \mathbf{d}^{\mathrm{T}} \mathbf{h}^{\mu}(\hat{\mathbf{x}})$$

$$\hat{\mathbf{x}}^{\backslash \mu} \sim \hat{\mathbf{x}} - \chi^{\backslash \mu} \mathbf{h}^{\mu}(\hat{\mathbf{x}}), \qquad \chi^{\backslash \mu} = \frac{\partial \hat{\mathbf{x}}^{\backslash \mu}}{\partial \mathbf{h}}$$

Approximate CV formula

- Approximate CV formula:

Computable only from $\hat{\boldsymbol{x}}$

$$\epsilon_{\mathrm{LOO}} \approx \frac{1}{2M} \sum_{\mu=1}^{M} \boxed{\Theta_\mu \left(y_\mu - \boldsymbol{a}_\mu^\top \hat{\boldsymbol{x}}\right)^2}$$

S_A: support

$$\Theta_\mu = \left(1 - (\boldsymbol{a}_\mu)_{S_A}^\top \underline{\left(\left(A_{*S_A}\right)^\top A_{*S_A} + \left(\partial^2 J(\hat{\boldsymbol{x}}_{S_A}; \eta)\right)_{S_A S_A}\right)}^{-1} (\boldsymbol{a}_\mu)_{S_A}\right)^{-2}.$$

cost function's Hessian on support

- Delicate points
  - Invariance of support between full and LOO solutions is assumed (approximately (exactly in N→∞) correct)
  - Regularity of cost function Hessian
    - Actually violated in RSB phase
  - Computational cost is $O(|S_A|^3)$

Contents

1. Analytical performance analysis in simulated dataset

2. Approximate CV formula

3. Numerical experiments

Experimental Setting

- Generative process: Identical to theoretical setting

$$\boldsymbol{y} = A\boldsymbol{x}_0 + \boldsymbol{\Delta}$$

$$A_{\mu i} \sim \mathcal{N}(0, N^{-1}) \qquad \Delta_i \sim \mathcal{N}(0, \sigma_\Delta^2)$$

$$x_{0i} \sim (1-\rho_0)\delta(x_{0i}) + \rho_0\mathcal{N}(0, \sigma_x^2)$$

all i.i.d.

- Optimization algorithm: Cyclic Coordinate Descent (CCD)
  - Coordinate-wise update optimizing the cost function
- A technique: λ annealing
  - Pathwise optimization with gradually changing λ
    - Faster convergence
    - Robust solution even in RSB region

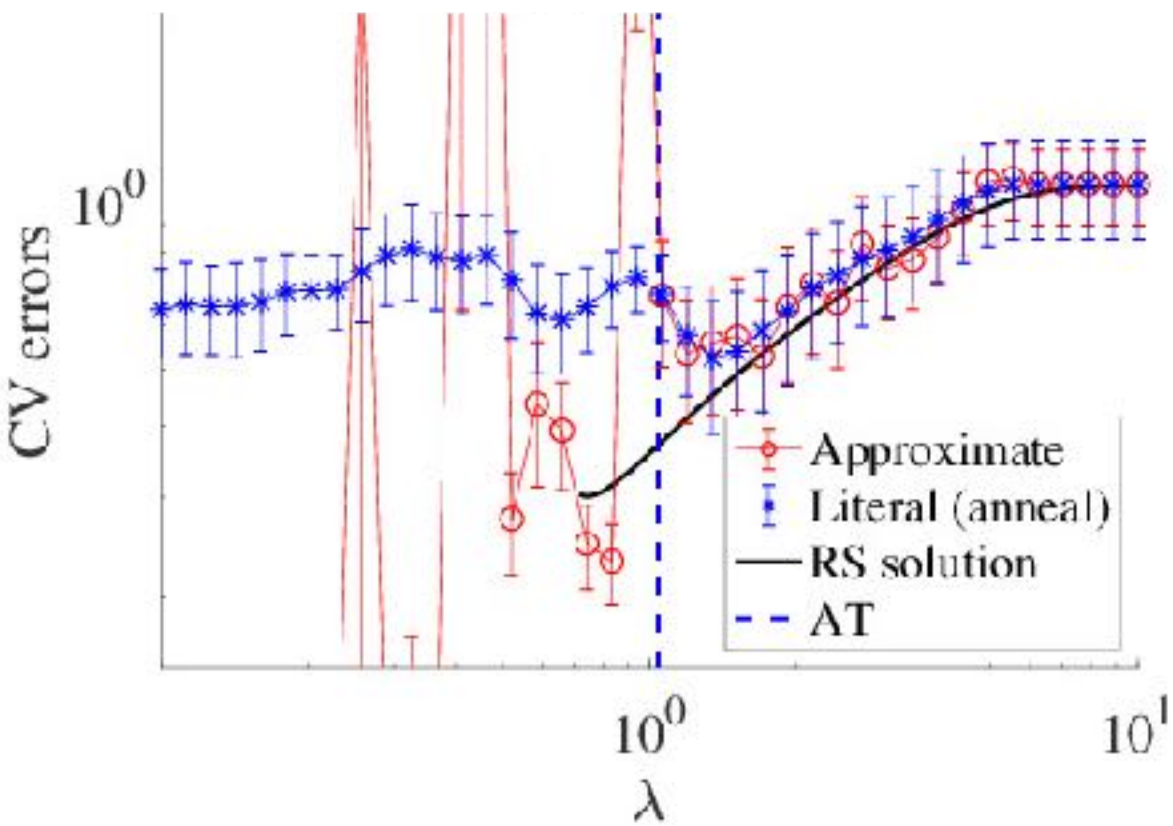
Approx. CV: Sample dependence

SCAD parameter a = 3

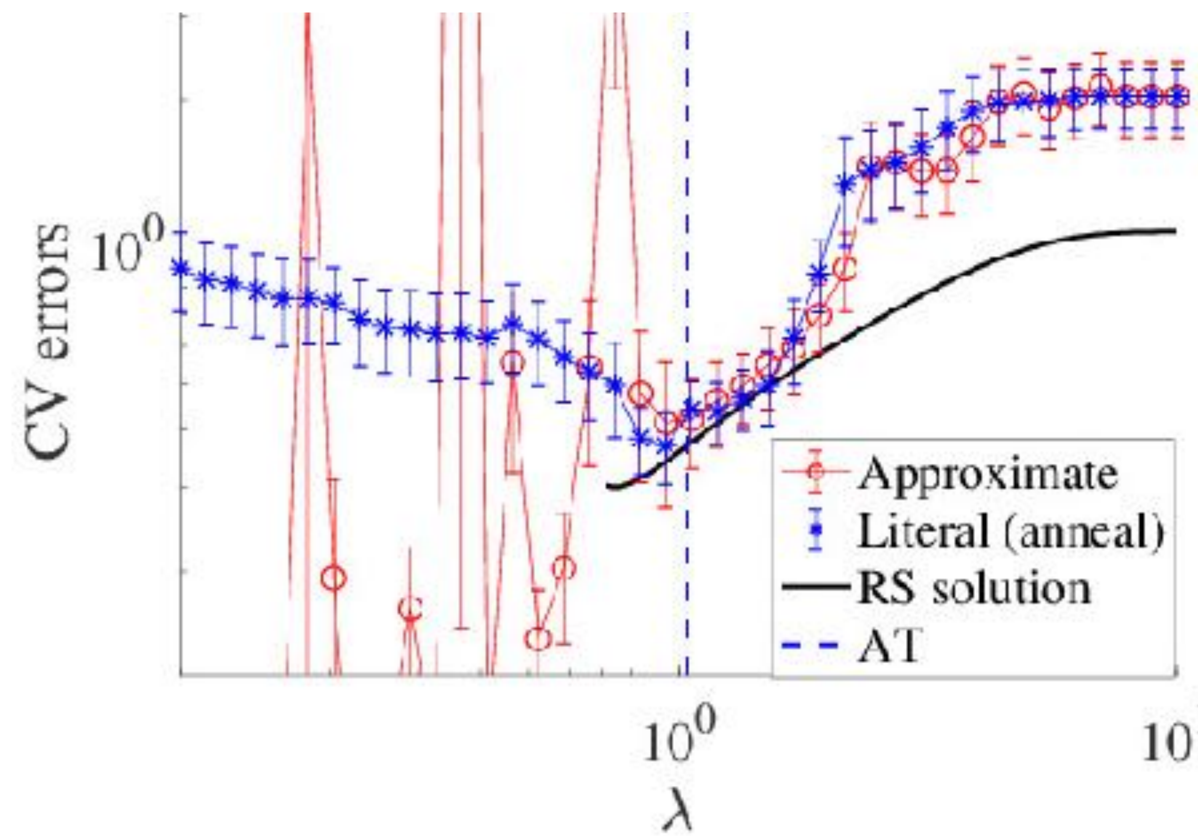
$$\alpha = 0.5,\ \rho_0 = 0.2,\ \sigma_\Delta^2 = 0.1,\ N = 100$$

(Error bar is for components of data.)

Sample No.1



Sample No.4



- CV error fluctuates depending on sample.
- Approximated CV error is valid in RS phase for both samples.

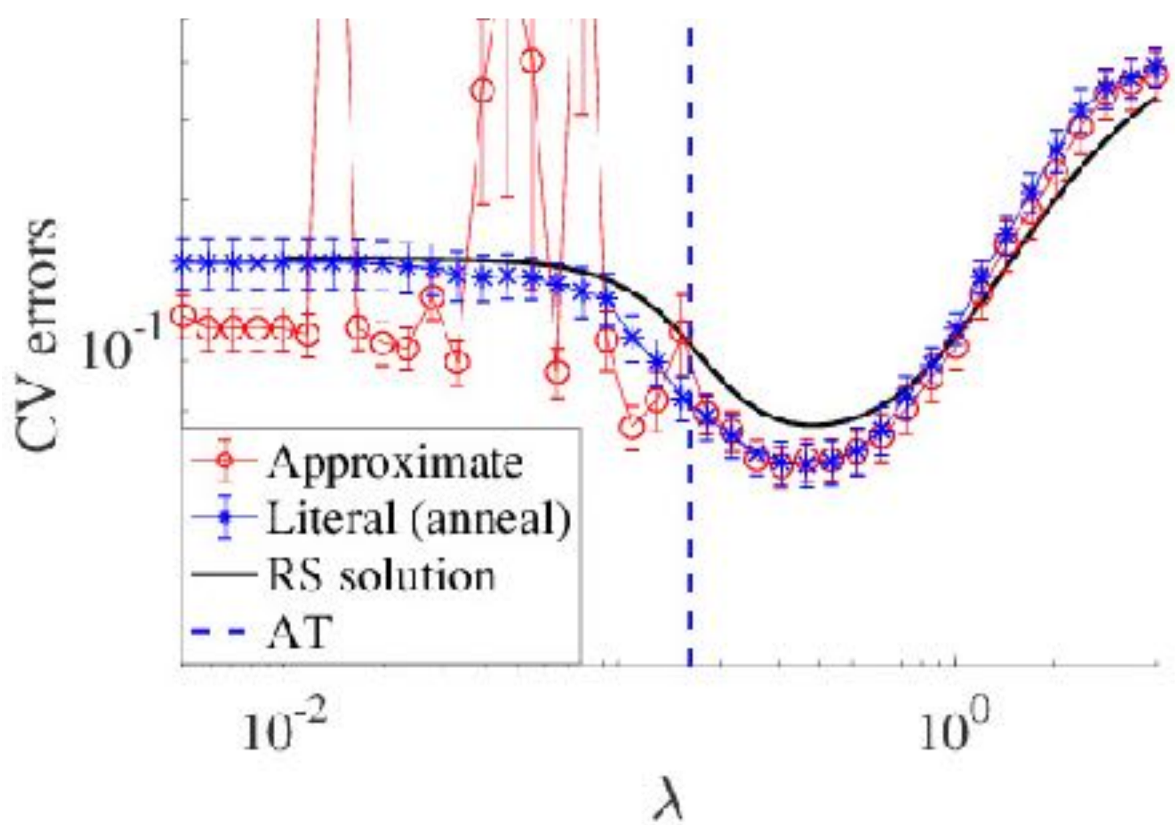
Approx. CV: Sample dependence

SCAD parameter a = 4

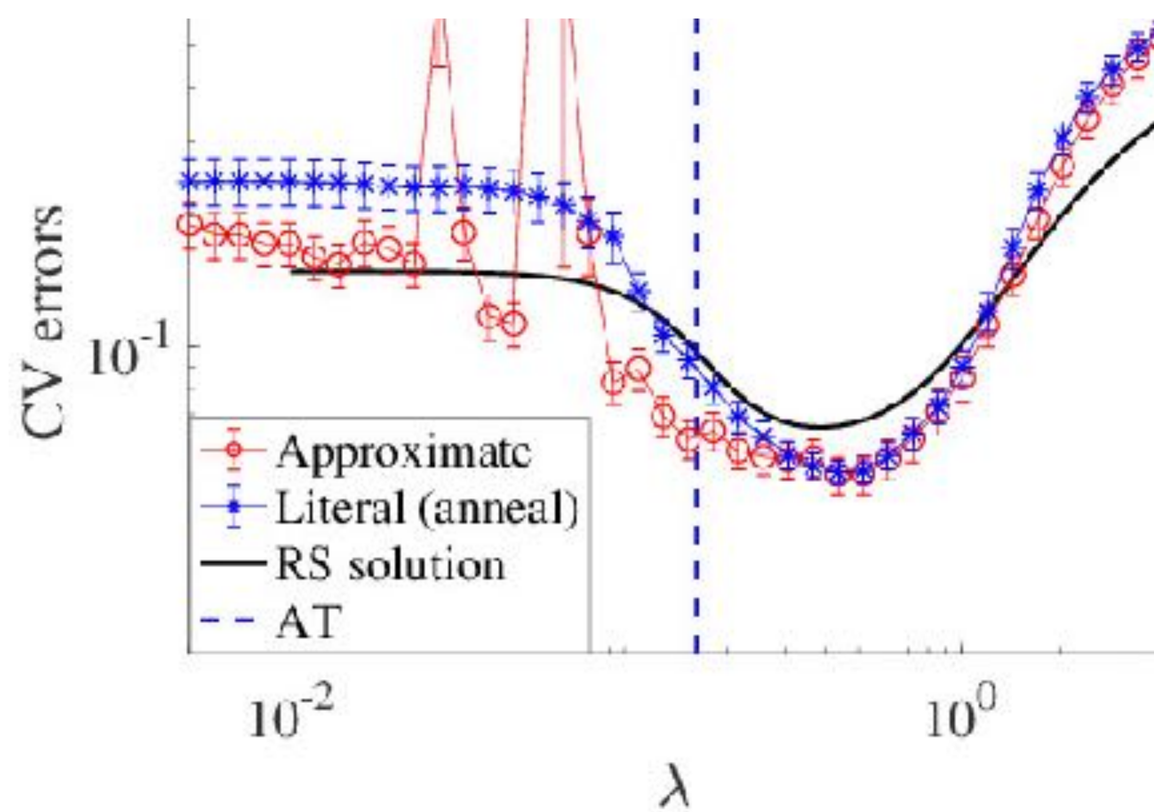
$$\alpha = 0.5,\ \rho_0 = 0.2,\ \sigma_\Delta^2 = 0.1,\ N = 100$$

(Error bar is for components of data.)

Sample No.1



Sample No.4



Sample dependence becomes moderate

as increase SCAD parameter a.

“Phase diagram” for given data

- “Phase” is defined for the infinite set of samples that are distributing according to a probability distribution.

- In practical problems,
  - Appropriate parameter region for a given data is required.
  - In particular for finite size system, sample-dependency is large.

- We propose a method to get “phase diagram” for given data.
- In other words, we identify the parameter region where we should rule out as candidates.

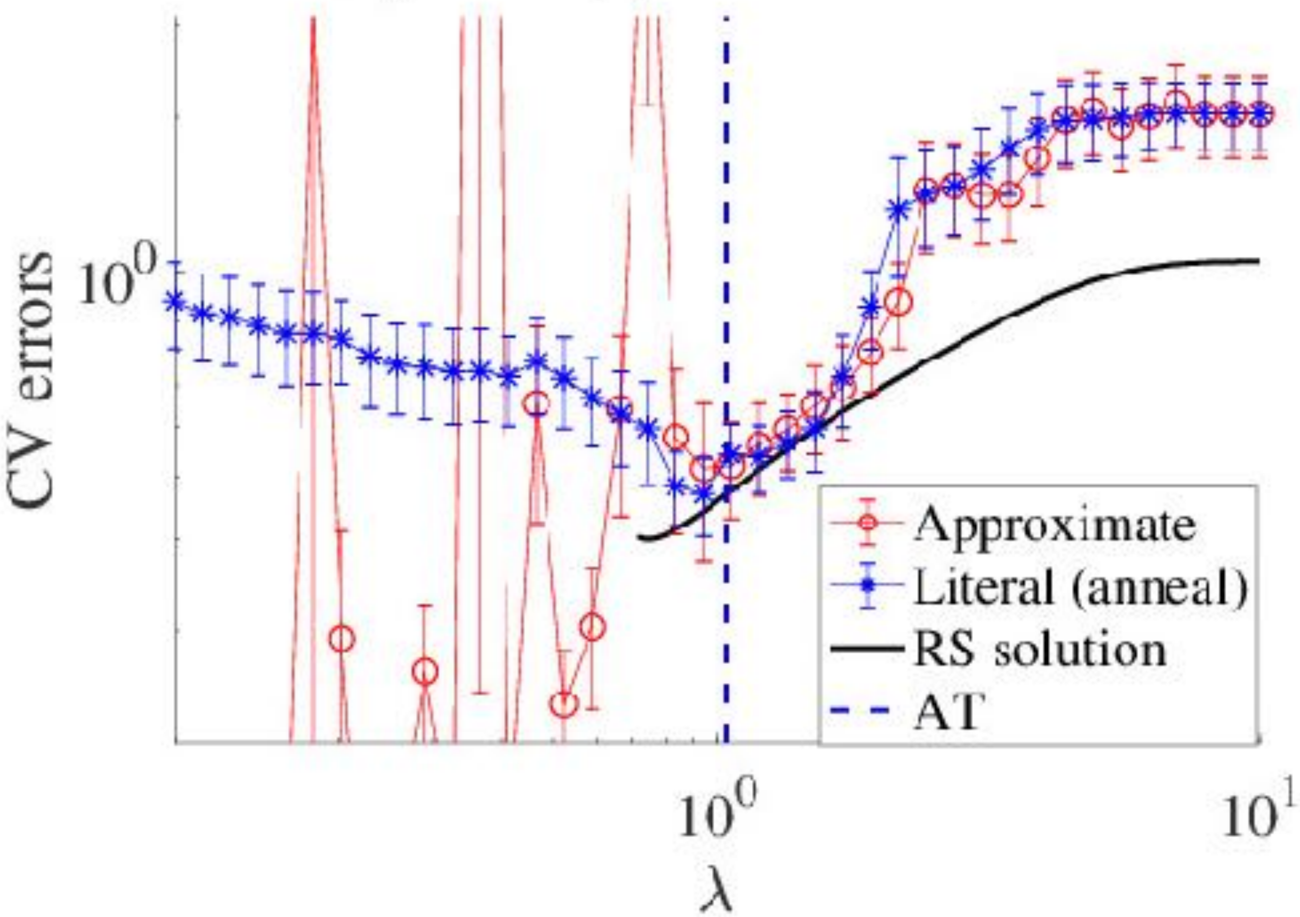
Approx. CV: Instability detection for "phase diagram"

We use our approximate CV formula to detect "RSB" region.

SCAD parameter a = 3

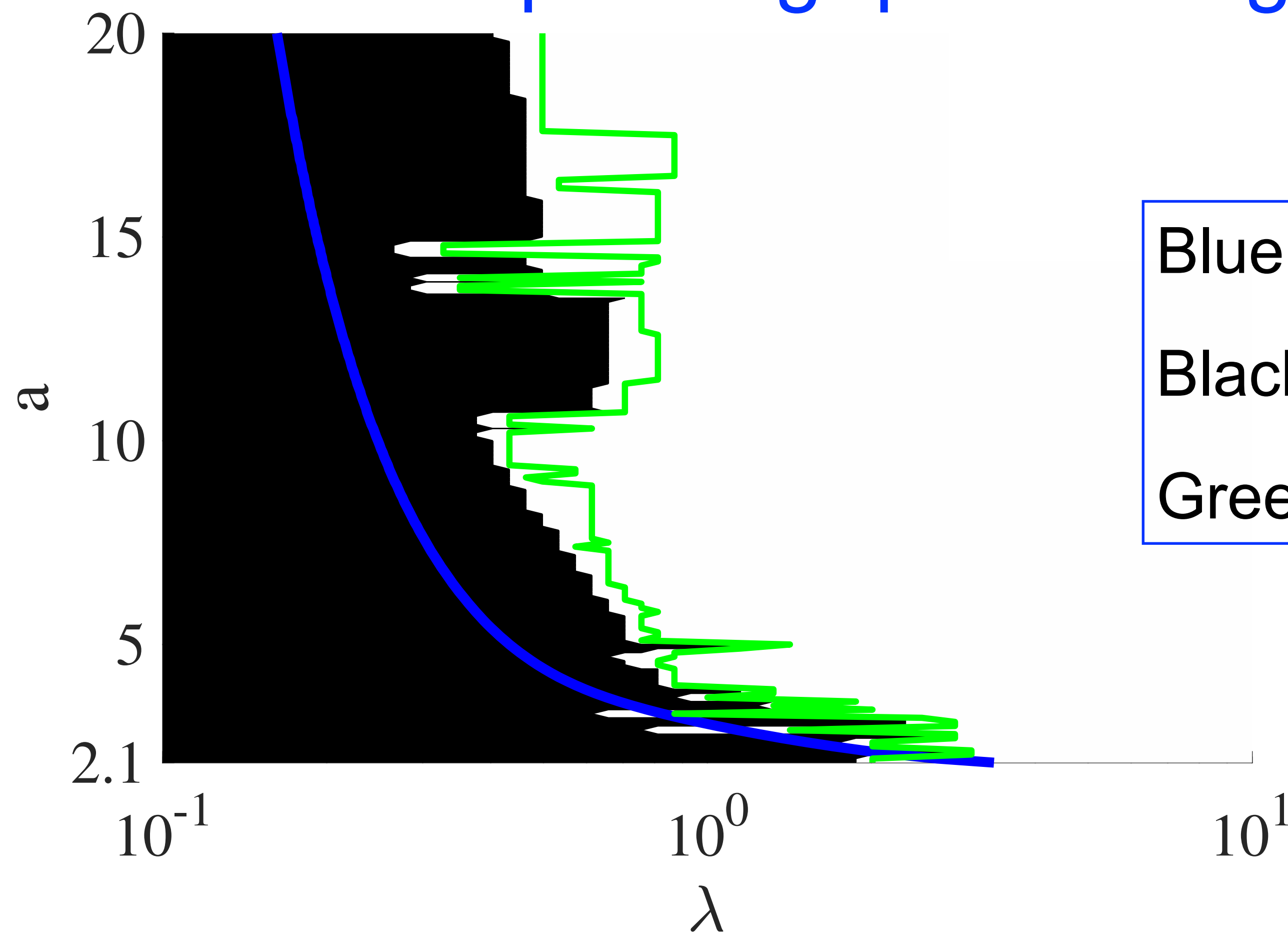
$\alpha = 0.5,\ \rho_0 = 0.2,\ \sigma_\Delta^2 = 0.1,\ N = 100$

Sample No.4



- Detect "irregular" datapoints along the λ path.
- Find the maximum λ value of irregular datapoints.
- λ smaller than the maximum value is inappropriate in the sense that instability appears.

Corresponding "phase diagram" for sample No.4

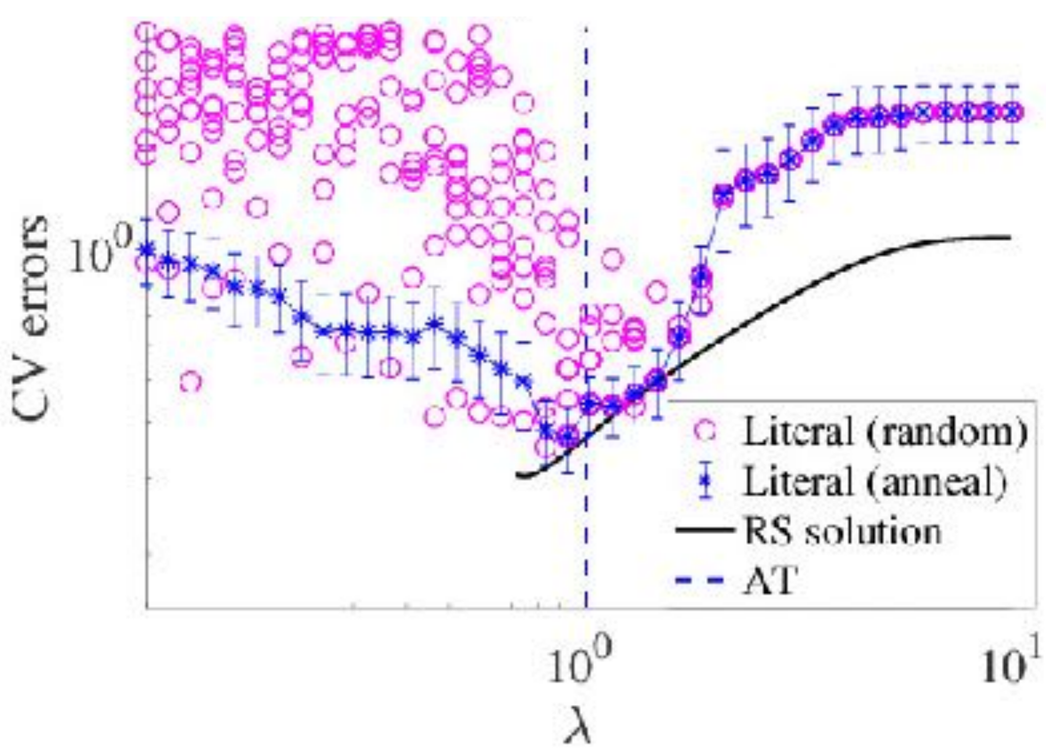


Blue line: AT line (RS-RSB transition)

Black region: "RSB region" for sample No.4

Green line: Minimum of CV error

What happens in "RSB" region (black)?

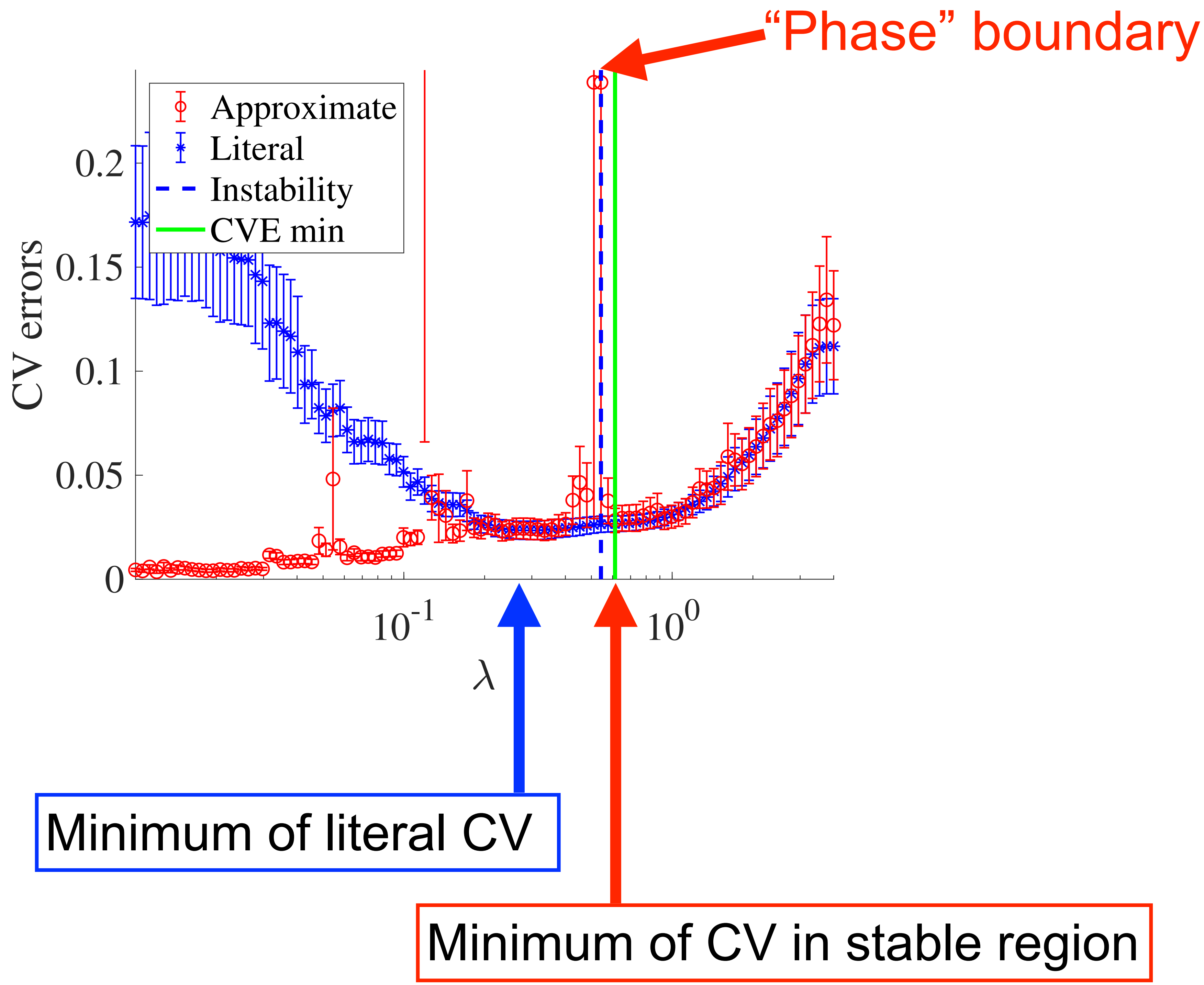


Starting from 10 different initial condition without annealing, literal CV's value fluctuates in the "RSB" region.

Application to SuperNovae data analysis

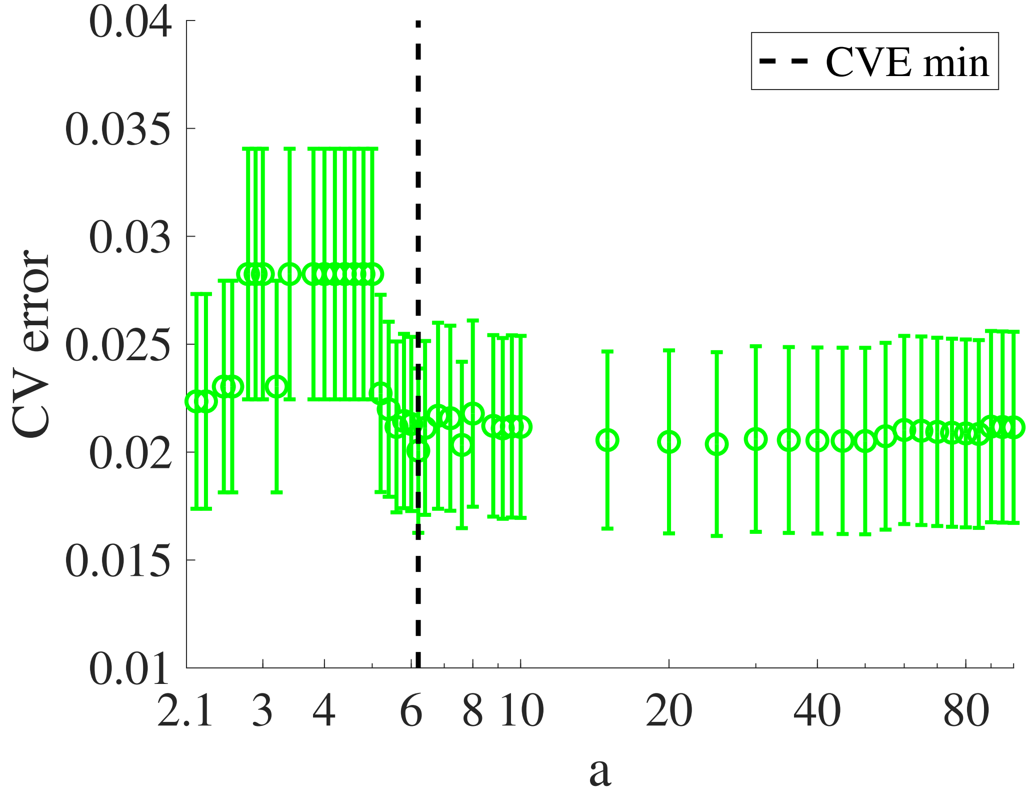
http://heracles.astro.berkeley.edu/sndb/

CV error for a=4

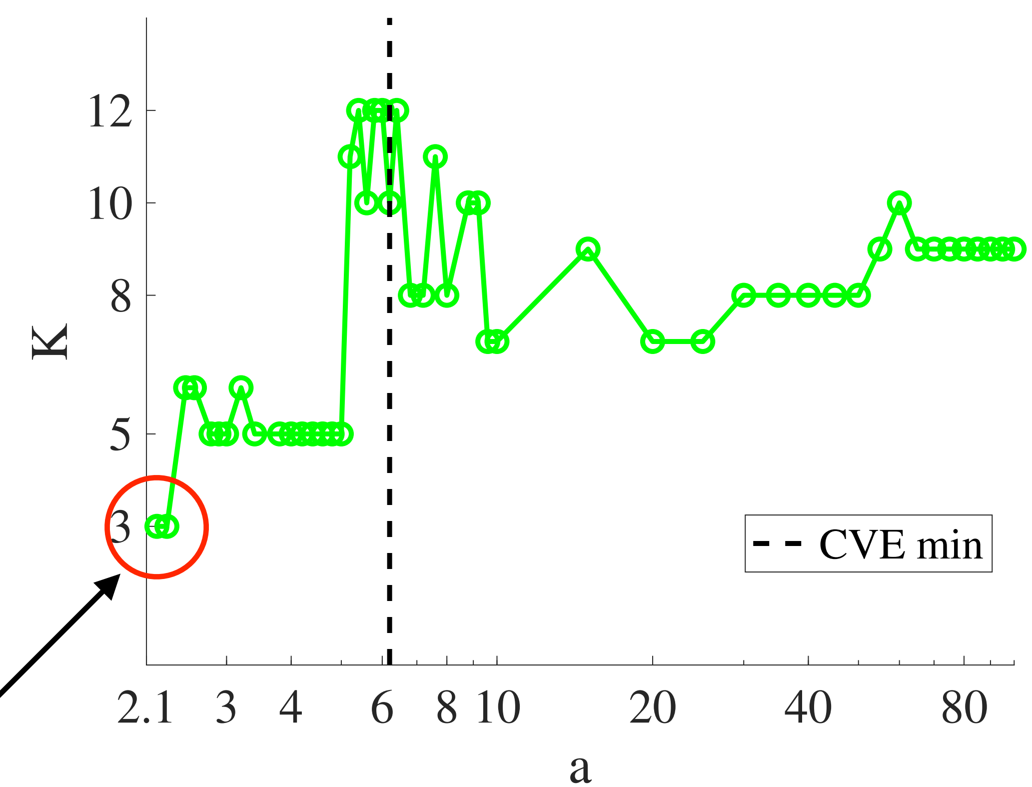


Application to SuperNovae data analysis

CV error



Number of parameters in model



Sparsest within one-sigma rule: K=3

←Identical solution to a Monte-Carlo method solving L0 problem

(TO et al, 2016,2018)

Our method is consistent with L0 result

even though SCAD is more computationally reasonable

Summary

TO, AS, arXiv:1902.10375

- Theoretical analysis of SCAD estimator in linear regression
  - Emergence of local minima = Phase transition w. RSB
  - Analytical evidence of outperformance of SCAD to LASSO
- Invention of an approximate CV formula
  - The scaling is O(N^3) but still practical in a wide range of N
  - Approximate CV instability <-> Local minima or RSB
    - Instability detection in CV formula also signals RSB
- Numerical results fully support the theoretical result
- A MATLAB Package of approx. CV formula + CCD algorithm: https://github.com/T-Obuchi/SLRpackage_AcceleratedCV_matlab

- Future work
  - Characterization of the λ annealed solution path
  - Applications, different models (non-L2 cost function)