Asymptotic Errors for Convex Penalized Linear Regression beyond Gaussian Matrices

C. Gerbelot, A. Abbara and F. Krzakala

Laboratoire de Physique de l'Ecole normale supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, F-75005 Paris, France

June 20, 2020

Position of the problem

Convex penalized linear regression

$$\mathbf{x}^* = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{F}\mathbf{x}\|_2^2 + f(\mathbf{x}) \right\} \tag{1}$$

where $\mathbf{y} = \mathbf{F}\mathbf{x}_0 + \mathbf{w}$

$$\mathbf{w} \sim \mathcal{N}(0, \Delta_0 Id), \quad \mathbf{x}_0 \sim p_{x_0}$$

- ground-truth $\mathbf{x}_0$ pulled from any (well-behaved) distribution
- f is a convex, separable function
- **high dimensional limit** $M, N \to \infty$, fixed ratio $\alpha = M/N$

Examples

Ridge Regression

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^N} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{F}\mathbf{x}\|_2^2 + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2 \right\}$$

Simplest building block, basis of kernel regression, neural net training, ...

LASSO

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^N} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{F}\mathbf{x}\|_2^2 + \lambda_1|\mathbf{x}|_1 \right\}$$

Ubiquitous in statistics, compressed sensing, variable selection

Elastic net

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^N} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{F}\mathbf{x}\|_2^2 + \lambda_1|\mathbf{x}|_1 + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2 \right\}$$

Combined regularization and variable selection, also mainstream

Objective : how good is my regression ?

Asymptotic reconstruction performance

$$E = \lim_{N \to \infty} \frac{1}{N} \|\mathbf{x}^* - \mathbf{x}_0\|_2^2$$

- fundamental building-block of modern statistical learning
- choice of $\mathbf{F}$ and penalty f is crucial
- well-known problem for i.i.d Gaussian matrix :

For ridge regression : closed form solution, random matrix theory

For the LASSO : [BM11] with message-passing algorithms, [TOH15] using Gordon's comparison theorem

The big question

Can we go beyond i.i.d Gaussian F ?

For any convex regularization f ?

Our answer

Can we go beyond i.i.d Gaussian F ? YES

Rotationally invariant matrix

$\mathbf{F} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, $\mathbf{U}, \mathbf{V}$ Haar distributed, and $\mathbf{D}$ contains singular values with **arbitrary distribution with compact support**.

For any convex regularization f ? YES

Any convex, **separable** f.

Main result : analytical solution

Fixed point equations

$$V = \mathbb{E}\left[\frac{1}{\mathcal{R}_{\mathbf{C}}(-V)}\text{Prox}'_{f/\mathcal{R}_{\mathbf{C}}(-V)}\left(x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}(-V)}\sqrt{(E-\Delta_0 V)\,\mathcal{R}'_{\mathbf{C}}(-V) + \Delta_0\mathcal{R}_{\mathbf{C}}(-V)}\right)\right]$$

$$E = \mathbb{E}\left[\left\{\text{Prox}_{f/\mathcal{R}_{\mathbf{C}}(-V)}\left(x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}(-V)}\sqrt{(E-\Delta_0 V)\,\mathcal{R}'_{\mathbf{C}}(-V) + \Delta_0\mathcal{R}_{\mathbf{C}}(-V)}\right) - x_0\right\}^2\right],$$

where $\mathbf{C} = \mathbf{F}^T\mathbf{F}$, $\mathcal{R}_{\mathbf{C}}$ is the R-transform with respect to the spectral distribution of $\mathbf{F}^T\mathbf{F}$, and expectations are over $z \sim \mathcal{N}(0,1)$ and $x_0 \sim p_{x_0}$.

Prox is the **proximal operator** defined as:

$$\forall \gamma \in \mathbb{R}^+, x, y \in \mathbb{R} \quad \text{Prox}_{\gamma f}(y) \equiv \arg\min_x \left\{ f(x) + \frac{1}{2\gamma}(x-y)^2 \right\}.$$

Proving a replica formula

- initially conjectured by [RGF09], [KVC12], [KV14]
- done using the **replica method** from statistical physics
- replicas are typically proven using interpolation methods [Guerra-Toninelli02], [Talagrand03], [BDMK16]
- for i.i.d. Gaussian matrices !

Here we propose a proof for matrices with arbitrary bounded spectrum, using **message-passing algorithms**

Experimental verification : LASSO with non-Gaussian data

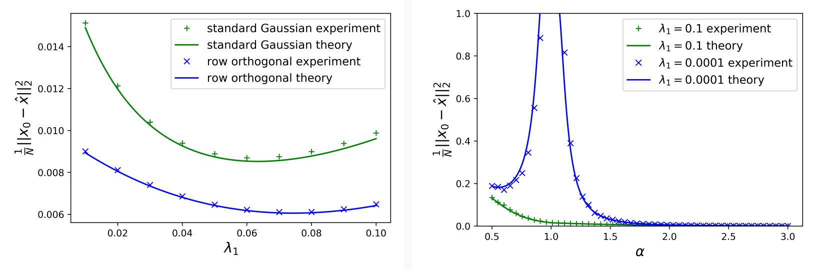

Figure 1: Left: LASSO parameter tuning with row-orthogonal matrix. **Right**: effect of aspect ratio on LASSO with uniformly sampled singular values.

Very accurate at finite sizes ! Here $N = 250, M = \alpha N$

"**double descent**" **depends on singular value distribution**

The proof

Let's look at the sketch of proof

Sketch of proof : key points ...

Key points :

(i) Build a sequence whose fixed point solves problem (1)

(ii) Have asymptotic statistical characterization of the iterates

(iii) Ensure convergence of the sequence

At the fixed point of the sequence, we will have $\mathbf{x}^*$ and its statistical properties.

Sketch of proof : key points ... and how to handle them

Key points :

(i) Use **vector approximate message-passing** [Rangan et. al. 2019]

(ii) Statistical characterization with **state evolution equations**

(iii) Study the **convergence** of VAMP

VAMP has been developed at the crossroads between statistical physics, variational inference and information theory.

Specifically derived to handle rotationally invariant matrices

(i) The sequence : Vector approximate message passing

Choose initial A_{10} *and* $\mathbf{B}_{10}$

$$\hat{\mathbf{x}}_{1k} = \text{Prox}_{\frac{1}{A_{1k}} f}\left(\frac{\mathbf{B}_{1k}}{A_{1k}}\right) \qquad \hat{\mathbf{x}}_{2k} = (\mathbf{F}^T\mathbf{F} + A_{2k} Id)^{-1}(\mathbf{F}^T y + \mathbf{B}_{2k}) \quad (2)$$

$$V_{1k} = \frac{\langle \text{Prox}'_{\frac{1}{A_{1k}} f} \rangle}{A_{1k}} \qquad V_{2k} = \frac{1}{N}\text{Tr}\left[(\mathbf{F}^T\mathbf{F} + A_{2k} Id)^{-1}\right] \quad (3)$$

$$A_{2k} = \frac{1}{V_{1k}} - A_{1k} \qquad A_{1,k+1} = \frac{1}{V_{2k}} - A_{2k} \quad (4)$$

$$\mathbf{B}_{2k} = \frac{\hat{\mathbf{x}}_{1k}}{V_{1k}} - \mathbf{B}_{1k} \qquad \mathbf{B}_{1,k+1} = \frac{\hat{\mathbf{x}}_2^t}{V_{2k}} - \mathbf{B}_{2k} \quad (5)$$

(2): estimation (3),(4) : adaptative parameters (5): update

Adaptative step size proximal descent

(ii) Statistical properties : State Evolution Equations

Estimators are asymptotically Gaussian

$$\mathbf{B}_{1k} - \mathbf{x}_0 \sim \mathcal{N}(0, \tau_{1k} Id) \quad \mathbf{B}_{2k} - \mathbf{x}_0 \sim \mathcal{N}(0, \tau_{2k} Id)$$

Proven in [Rangan et. al. 2019]. Full state evolution:

$$\alpha_{1k} = \mathbb{E}\left[\mathrm{Prox}'_{\frac{1}{A_{1k}}f}(x_0 + P_{1k})\right] \qquad V_{1k} = \frac{\alpha_{1k}}{A_{1k}}$$

$$A_{2k} = \frac{1}{V_{1k}} - A_{1k} \qquad \tau_{2k} = \frac{1}{(1-\alpha_{1k})^2}\left[\mathcal{E}_1(A_{1k}, \tau_{1k}) - \alpha_{1k}^2 \tau_{1k}\right]$$

$$\alpha_{2k} = \mathbb{E}\left[\frac{A_{2k}}{\lambda_{\mathbf{F}^T\mathbf{F}} + A_{2k}}\right] \qquad V_{2k} = \frac{\alpha_{2k}}{A_{2k}}$$

$$A_{1,k+1} = \frac{1}{V_{2k}} - A_{2k} \qquad \tau_{1,k+1} = \frac{1}{(1-\alpha_{2k})^2}\left[\mathcal{E}_2(A_{2k}, \tau_{2k}) - \alpha_{2k}^2 \tau_{2k}\right].$$

Match the replica prediction at their fixed point

(iii) Convergence analysis : Oracle-VAMP

Prescribe $A_1, A_2, V_1 = V_2 = V$ from state evolution fixed point

Choose initial $\mathbf{B}_{10}$

$$\hat{\mathbf{x}}_{1k} = \mathrm{Prox}_{\frac{1}{A_1}f}\left(\frac{\mathbf{B}_{1k}}{A_1}\right) \qquad \hat{\mathbf{x}}_{2k} = (\mathbf{F}^T\mathbf{F} + A_2 Id)^{-1}(\mathbf{F}^T y + \mathbf{B}_{2k})$$

$$\mathbf{B}_{2k} = \frac{\hat{\mathbf{x}}_{1k}}{V_1} - \mathbf{B}_{1k} \qquad \mathbf{B}_{1,k+1} = \frac{\hat{\mathbf{x}}_2^t}{V_2} - \mathbf{B}_{2k}$$

Oracle, single update sequence

$$\mathbf{B}_2^{t+1} = \left(\frac{1}{V}\mathrm{Prox}_{\frac{1}{A_1}f}\left(\frac{\cdot}{A_1}\right) - Id\right) \circ \left(\frac{1}{V}\mathrm{Prox}_{\frac{1}{2A_2}||\mathbf{y}-\mathbf{F}\mathbf{x}||_2^2}\left(\frac{\cdot}{A_2}\right) - Id\right)(\mathbf{B}_2^t)$$

(iii) Convergence analysis : Oracle-VAMP

Generate a sequence with the prescription :

$$\mathbf{B}_2^{t+1} = \left(\frac{1}{V} \text{Prox}_{\frac{1}{A_1} f}\left(\frac{\cdot}{A_1}\right) - Id \right) \circ \left(\frac{1}{V} \text{Prox}_{\frac{1}{2A_2} ||\mathbf{y} - \mathbf{F}\mathbf{x}||_2^2}\left(\frac{\cdot}{A_2}\right) - Id \right) (\mathbf{B}_2^t)$$

Upper bound on the Lipschitz constant of the update operator :

$$\max\left(\frac{|A_1 - \lambda_{min}(\mathbf{F}^T\mathbf{F})|}{A_2 + \lambda_{min}(\mathbf{F}^T\mathbf{F})}, \frac{|\lambda_{max}(\mathbf{F}^T\mathbf{F}) - A_1|}{A_2 + \lambda_{max}(\mathbf{F}^T\mathbf{F})} \right) \sqrt{\left(\frac{(A_2^2 - A_1^2)}{(A_1 + \sigma_1)^2} + 1 \right)}$$

where σ_1 is the strong convexity constant of the penalty f.

Could this be a contraction ?

(iii) Forcing the convergence

Imposing strong convexity :

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min}\ \frac{1}{2}\|\mathbf{y} - \mathbf{F}\mathbf{x}\|_2^2 + f(\mathbf{x}) + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2$$

Remember :

$$\max\left(\frac{A_1 - \lambda_{min}(\mathbf{F}^T\mathbf{F})}{A_2 + \lambda_{min}(\mathbf{F}^T\mathbf{F})}, \frac{\lambda_{max}(\mathbf{F}^T\mathbf{F}) - A_1}{A_2 + \lambda_{max}(\mathbf{F}^T\mathbf{F})}\right)\sqrt{\left(\frac{(A_2^2 - A_1^2)}{(A_1 + \sigma_1 + \lambda_2)^2} + 1\right)}$$

Possibility to force convergence for large enough λ_2 due to:

$$\lambda_{min}(\mathbf{F}^T\mathbf{F}) \leqslant A_1 \leqslant \lambda_{max}(\mathbf{F}^T\mathbf{F}) \quad \lambda_{min}(\mathcal{H}_f) + \lambda_2 \leqslant A_2 \leqslant \lambda_{max}(\mathcal{H}_f) + \lambda_2$$

Experimental verification of this fact in the paper.

Final step : analytic continuation

- proof complete for an open subset of λ_2
- dependence in λ_2 is analytical in the replica formulas
- dependence in λ_2 is analytical in the coordinates of $\mathbf{x}^*$
- extend the result for any λ_2 with analytic continuation theorem [KP02]

The proof is complete

Thank you