

The Annals of Statistics
2002, Vol. 30, No. 3, 631–682

THE STATISTICAL WORK OF LUCIEN LE CAM

BY AAD VAN DER VAART

Free University, Amsterdam

We give an overview and appraisal of the scientific work in theoretical statistics, and its impact, by Lucien Le Cam. The references to Le Cam's papers refer to the Le Cam bibliography. The reference is the first paper for the given year if not stated.

Le Cam Theory

BAYESIAN LINEAR REGRESSION WITH SPARSE PRIORS

BY ISMAËL CASTILLO^{1,*}, JOHANNES SCHMIDT-HIEBER^{2,†}
AND AAD VAN DER VAART^{2,†}

CNRS Paris and Leiden University†*

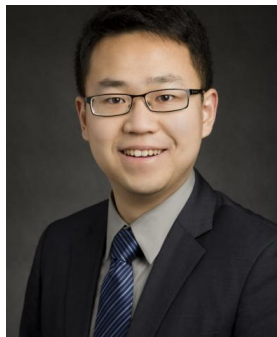
We study full Bayesian procedures for high-dimensional linear regression under sparsity constraints. The prior is a mixture of point masses at zero and continuous distributions. Under compatibility conditions on the design matrix, the posterior distribution is shown to contract at the optimal rate for recovery of the unknown sparse vector, and to give optimal prediction of the response vector. It is also shown to select the correct sparse model, or at least the coefficients that are significantly different from zero. The asymptotic shape of the posterior distribution is characterized and employed to the construction and study of credible sets for uncertainty quantification.

Frequentist Bayesian Investigation (FBI)

Statistical and Computational Guarantees of EM with Random Initialization

Harrison H. Zhou
Department of Statistics and Data Science
Yale University

Jointly with Yihong Wu



Expectation-Maximization (EM) Algorithm

Model: Parametric family $p_{\theta}(y, z)$

Y : observed

Z : latent

θ : unknown parameter

Maximal likelihood estimator (MLE):

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} \log p_{\theta}(Y),$$

where $p_{\theta}(y) = \int p_{\theta}(y, z) dz$.

Expectation-Maximization (EM) Algorithm

EM algorithm: Dempster, Laird, Rubin (1977)

Given current estimate θ_t ,

- E step

$$Q(\theta|\theta_t) \triangleq \int p_{\theta_t}(z|y) \log p_{\theta}(y, z) dz.$$

- M step

$$\theta_{t+1} = \arg \max_{\theta} Q(\theta|\theta_t).$$

Interpretation:

$$\log p_{\theta}(y) - \log p_{\theta_t}(y) \geq Q(\theta|\theta_t) - Q(\theta_t|\theta_t).$$

Mixture of Gaussians: $k = 2$

Model: Observe $Y_1, \dots, Y_n \in \mathbb{R}^d$ from

$$Y_i = Z_i \cdot \theta^* + W_i,$$

where $Z_i \in \{-1, 1\}$ are i.i.d. $\text{Rad}(1/2)$, and $\{W_i\}$ are independent with $W_i \sim \mathcal{N}(0, I_d)$.

Goal: Estimate θ^* up to a global sign flip, i.e.,

$$\ell(\hat{\theta}, \theta^*) = \min\{\|\hat{\theta} + \theta^*\|, \|\hat{\theta} - \theta^*\|\}.$$

Minimax rate: Assume $\|\theta^*\| = O(1)$. For $n \geq d$,

$$\inf_{\hat{\theta}} \sup_{\theta^*} \mathbb{E} \ell(\hat{\theta}, \theta^*) \asymp (d/n)^{1/4}.$$

EM Iteration

Sample version:

$$\hat{\theta}_{t+1} = f_n(\hat{\theta}_t) = \mathbb{E}_n[Y \tanh(\hat{\theta}_t^T Y)] = \frac{1}{n} \sum_{i=1}^n Y_i \tanh(\hat{\theta}_t^T Y_i).$$

Population version:

$$\theta_{t+1} = f(\theta_t), \text{ where } f(\theta) = \mathbb{E}_{\theta^*}[Y \tanh(\theta^T Y)].$$

Non-convex optimization by first-order methods:

$$\hat{\theta}_{t+1} = \mathbb{E}_n[Y \tanh(\hat{\theta}_t^T Y)] = \hat{\theta}_t + \nabla \ell_n(\hat{\theta}_t),$$

where ℓ_n is the log-likelihood.

Remark: $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$

Some Recent Progress

- Balakrishnan, Wainwright, Yu (14): For symmetric 2-GM, requires $\|\theta^*\| \geq \text{large constant}$ and initialization satisfies $\|\theta_0 - \theta^*\| \leq \frac{1}{4}\|\theta^*\|$.
- Xu, Hsu, Maleki (16): Analysis of population EM in symmetric 2-GM.
- Daskalakis, Tzamos, Zampetakis (17): Use new samples for each iteration and need to run EM for a number of iterations, normalize and restart.

Question:

Faithful finite sample analysis without strong separation condition and warm initialization conditions?

An Encouraging First Result: $d = 1$

Theorem. Assume $|\theta^*| = O(1)$. With probability $1 - \delta$, the following holds:
For any $\hat{\theta}_0 > 0$ and t sufficiently large,

$$\ell(\hat{\theta}_t, \theta^*) \lesssim \underbrace{\min\left\{\frac{1}{|\theta^*|\sqrt{n}}, \frac{1}{n^{1/4}}\right\}}_{\text{optimal rate}} \sqrt{\log \frac{1}{\delta}}.$$

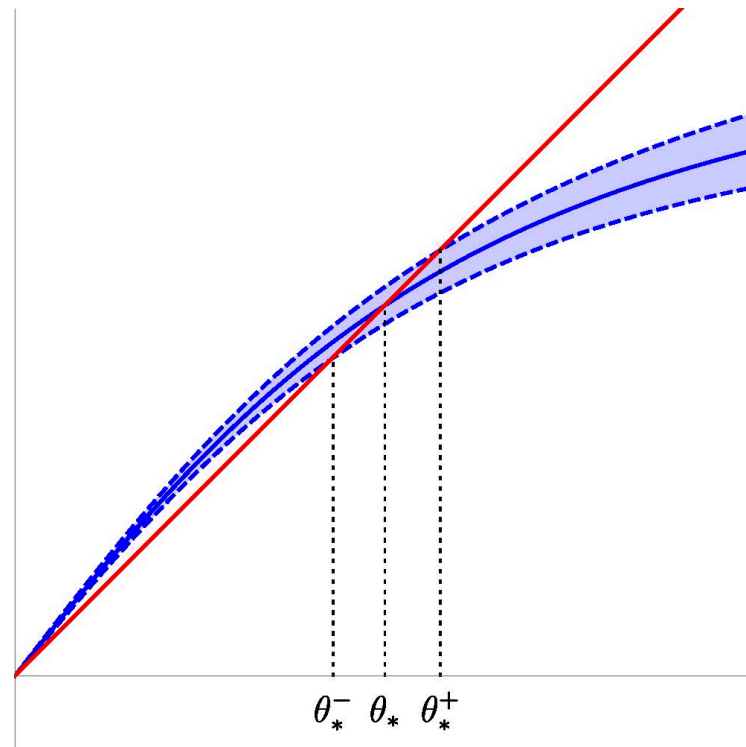
Remarks:

- Convergence for any $\hat{\theta}_0$.
- Starting from a constant $\hat{\theta}_0$, $\hat{\theta}_t$ reaches minimax optimal rate in at most $O(\sqrt{n})$ iterations, regardless of θ^* .

Reason: Slightly Perturbed EM Iteration

The sample EM path $\hat{\theta}_{t+1} = f_n(\hat{\theta}_t)$ is sandwiched between two auxiliary sequences:

$$\begin{cases} \hat{\theta}_{t+1}^+ = f(\hat{\theta}_t^+) + \frac{C_0}{\sqrt{n}} \hat{\theta}_t^+ \\ \hat{\theta}_{t+1}^- = f(\hat{\theta}_t^-) - \frac{C_0}{\sqrt{n}} \hat{\theta}_t^- \end{cases}, \quad \hat{\theta}_0^- = \hat{\theta}_0^+ = \hat{\theta}_0$$



General d

EM iteration:

$$\hat{\theta}_{t+1} = f_n(\hat{\theta}_t)$$

where $f_n : R^d \rightarrow R^d$ is given by

$$f_n(\theta) \triangleq \mathbb{E}_n[Y \tanh(\theta^T Y)] = \frac{1}{n} \sum_{i=1}^n Y_i \tanh(\theta^T Y_i).$$

Definition:

Let $\eta_* = \frac{1}{\|\theta_*\|} \theta_*$. Define

$$\alpha_t = \langle \hat{\theta}_t, \eta_* \rangle, \quad \beta_t = \|(I - \eta_* \eta_*^\top) \hat{\theta}_t\|.$$

Rotation Invariance:

We assume $\theta_* = \|\theta_*\| \cdot e_1$ so that

$$\alpha_t = \hat{\theta}_{t,1}, \quad \beta_t = \|\hat{\theta}_{t,\perp}\| = \|(\hat{\theta}_{t,2}, \dots, \hat{\theta}_{t,d})\|.$$

Main Result

Theorem. Let θ_0 be drawn uniformly at random from the unit sphere S^{d-1} . Assume that

$$n \geq Cd \log^4 d.$$

Then with high probability we have

$$\ell(\hat{\theta}_t, \theta^*) \lesssim \left(\frac{d}{n} \log^3 n \right)^{1/4},$$

for $t \gtrsim \frac{\log n}{\max\{\|\theta_*\|^2, \sqrt{d/n}\}}$. In addition, if $\|\theta_*\| \gtrsim \left(\frac{d \log n}{n} \right)^{1/4}$, we have

$$\mathbb{E} \ell^2(\hat{\theta}_t, \theta^*) = (1 + o(1)) \frac{\text{tr}(I^{-1}(\theta^*))}{n}.$$

Two important remarks:

- A random initialization suffices.
- No need to split samples.

Proof

- Stage 1 – Better than Random Guess.
- Stage 2 – Perturbed EM Iteration.
- Stage 3 – Asymptotic Efficiency.



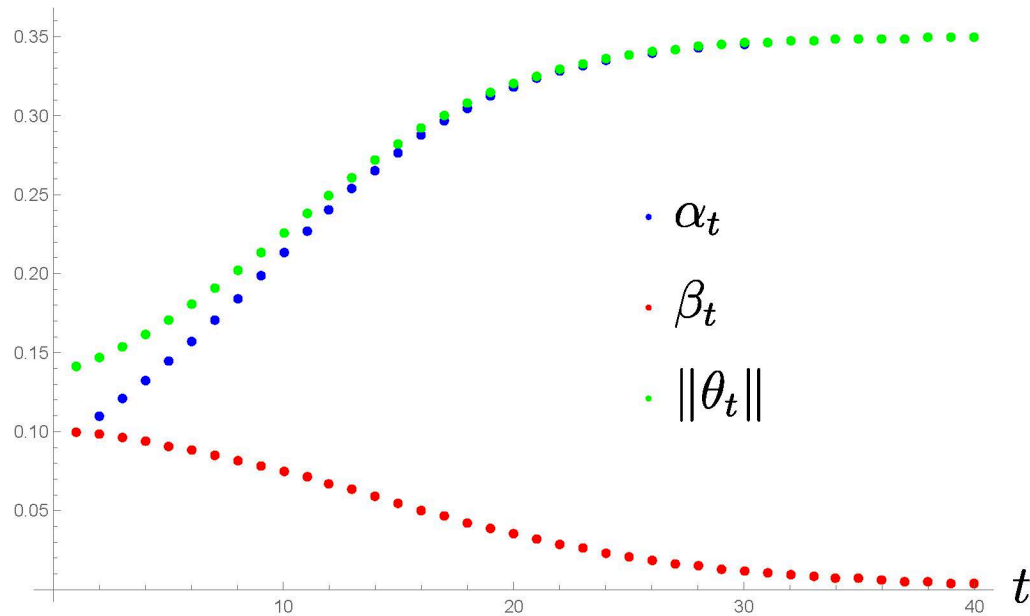
Proof – Stage 1 – Better than Random Guess

Key argument: To show that the signal component α_t grows exponentially according to

$$\alpha_{t+1} \geq \alpha_t(1 + \|\theta_*\|^2 - o(\|\theta_*\|^2)),$$

as long as $\alpha_t = o(\|\theta^*\|)$.

Monotonicity of α_t :



Proof – Stage 1 – Better than Random Guess

Sign perturbation:

Define $\tilde{Y}_i$ by multiplying the first coordinate of Y_i by an independent $\text{Rad}(1/2)$, $i = 1, 2, \dots, n$.

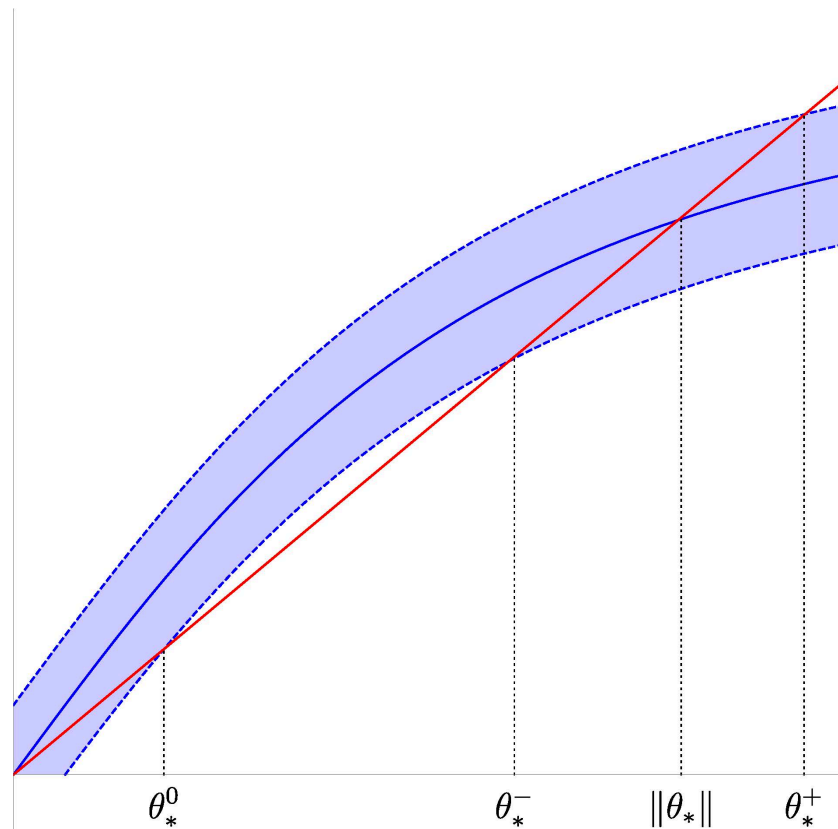
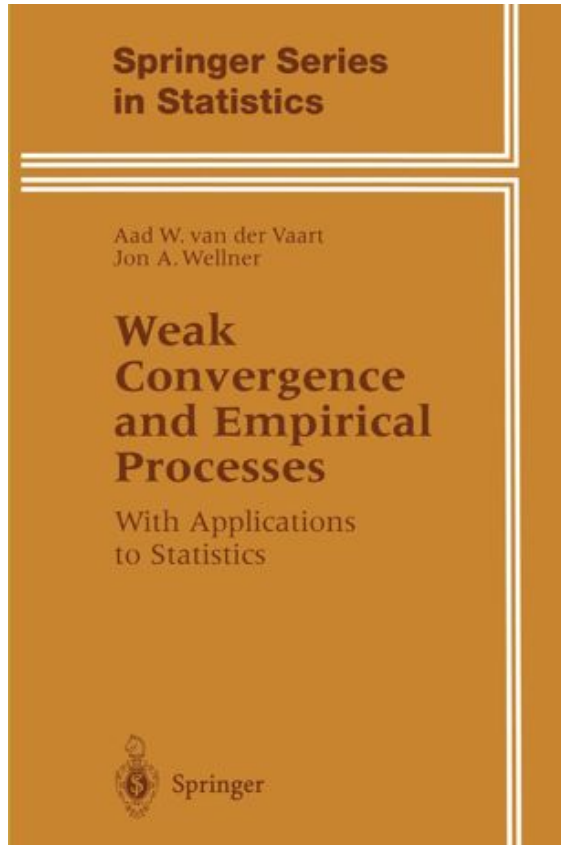
For $i = 1$, $\tilde{Y}_1 = (\xi_1 Y_{11}, Y_{21}, \dots, Y_{d1})^T$, for $Y_1 = (Y_{11}, Y_{21}, \dots, Y_{d1})^T$, where ξ_1 is $\text{Rad}(1/2)$.

Independence:

Correspondingly, we define $\tilde{\theta}_t$, which is independent of the sign of Y_{i1} , the first coordinate of Y_i , and has the same distribution as $\hat{\theta}_t$.

Reference: Chen, Chi, Fan, Ma (2018).

Proof – Stage 2 – Perturbed EM Iteration



Eventually, w.h.p. we have

$$\ell(\hat{\theta}_t, \theta^*) \lesssim \left(\frac{d}{n} \log^3 n \right)^{1/4}.$$

Proof – Stage 3 – Asymptotic Efficiency

Gradient descent

$$\hat{\theta}_{t+1} = \mathbb{E}_n[Y \tanh(\hat{\theta}_t^T Y)] = \hat{\theta}_t + \nabla \ell_n(\hat{\theta}_t)$$

where ℓ_n is the log-likelihood, which leads

$$\mathbb{E} \ell^2(\hat{\theta}_t, \theta^*) = (1 + o(1)) \frac{\text{tr}(I^{-1}(\theta^*))}{n},$$

for t sufficiently large.

Remark: Bickel (1975) needs only one Newton - Raphson step to attain efficiency, while for EM we need many steps.

Proof – Stage 3 – Asymptotic Efficiency

Contraction:

Applying the Taylor expansion of $\nabla \ell_n$ at $\hat{\theta}_{MLE}$, we have

$$\begin{aligned}\hat{\theta}_{t+1} - \hat{\theta}_{MLE} &= \hat{\theta}_t - \hat{\theta}_{MLE} + \nabla \ell_n(\hat{\theta}_t) - \nabla \ell_n(\hat{\theta}_{MLE}) \\ &= (I + \nabla^2 \ell_n(\xi_t))(\hat{\theta}_t - \hat{\theta}_{MLE}),\end{aligned}$$

where $\xi_t = \gamma \hat{\theta}_t + (1 - \gamma) \hat{\theta}_{MLE}$ for some $\gamma \in [0, 1]$.





Happy Birthday Aad

Question: Can we go beyond the two-component symmetric Gaussian mixtures case?