Uncertainty quantification for iterative algorithms in linear models with application to early stopping

Bellec C. Pierre and **Kai Tan**

Department of Statistics, Rutgers University

RUTGERS THE STATE UNIVERSITY OF NEW JERSEY

Gap between theory and practice

- For linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b}^* + \boldsymbol{\varepsilon}$, statistical theories (e.g., error bounds, asymptotic distributions, risk characterization) focus on the optimizer $\hat{\boldsymbol{b}}$:

$$\hat{\boldsymbol{b}} \in \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + g(\boldsymbol{b})$$

☹ In practice, $\hat{\boldsymbol{b}}$ cannot be solved exactly; iterative algorithms are used to produce iterates $\hat{\boldsymbol{b}}^1, \hat{\boldsymbol{b}}^2, ..., \hat{\boldsymbol{b}}^T$ (stop after T iterations).

☹ $\hat{\boldsymbol{b}}^T$ can be far from $\hat{\boldsymbol{b}}$, and the theories about $\hat{\boldsymbol{b}}$ do not apply to $\hat{\boldsymbol{b}}^T$.

☹ There is no guarantee that $\hat{\boldsymbol{b}}^t$ will get closer to $\boldsymbol{b}^*$ as t increases.

Q1: How can we quantify the predictive performance of $\hat{\boldsymbol{b}}^t$ at each iteration?
Q2: How does the performance of $\hat{\boldsymbol{b}}^t$ depend on the previous iterates?
Q3: How can we perform statistical inference on $\boldsymbol{b}^*$ using the iterate $\hat{\boldsymbol{b}}^t$?

Estimation Target

Consider an algorithm of the following form:

$$\hat{\boldsymbol{b}}^t = \boldsymbol{g}_t(\hat{\boldsymbol{b}}^{t-1}, \hat{\boldsymbol{b}}^{t-2}, \dots, \hat{\boldsymbol{b}}^2, \hat{\boldsymbol{b}}^1,\ \boldsymbol{v}^{t-1}, \dots, \boldsymbol{v}^2, \boldsymbol{v}^1), \qquad (1)$$

where $\boldsymbol{v}^t = \frac{1}{n}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}}^t)$.

Table 1. Examples of several algorithms

GD	$\hat{\boldsymbol{b}}^t = \hat{\boldsymbol{b}}^{t-1} + \eta\boldsymbol{v}^t$
AGD	$\hat{\boldsymbol{b}}^t = (1 - w_{t-1})(\hat{\boldsymbol{b}}^{t-1} + \eta\boldsymbol{v}^{t-1}) + w_{t-1}(\hat{\boldsymbol{b}}^{t-2} + \eta\boldsymbol{v}^{t-2})$
ISTA	$\hat{\boldsymbol{b}}^t = \mathrm{soft}_{\eta\lambda}(\hat{\boldsymbol{b}}^{t-1} + \eta\boldsymbol{v}^t)$
FISTA	$\hat{\boldsymbol{b}}^t = \mathrm{soft}_{\eta\lambda}((1 - w_{t-1})(\hat{\boldsymbol{b}}^{t-1} + \eta\boldsymbol{v}^{t-1}) + w_{t-1}(\hat{\boldsymbol{b}}^{t-2} + \eta\boldsymbol{v}^{t-2}))$

Estimation target: The generalization error r_t for each $\hat{\boldsymbol{b}}^t$:

$$r_t \stackrel{\text{def}}{=} \mathbb{E}\Big[(y_{new} - \boldsymbol{x}_{new}^\top\hat{\boldsymbol{b}}^t)^2 \mid (\boldsymbol{X}, \boldsymbol{y})\Big] = \|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{b}}^t - \boldsymbol{b}^*)\|^2 + \sigma^2.$$

Contributions

- Propose a novel estimator of r_t:

$$\underbrace{\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{b}}^t - \boldsymbol{b}^*)\|^2 + \sigma^2}_{r_t} \approx \underbrace{\frac{1}{n}\Big\|\sum_{s=1}^t \hat{w}_{t,s}\big(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}}^s\big)\Big\|^2}_{\hat{r}_t}$$

  - The $\hat{r}_t$ depends on the **weighted residual vector** of all previous iterates.
  - The weights $\hat{w}_{t,s}$ can be easily computed using observational quantities.
- Introduce the debiased estimate for the component of each iterate $\hat{\boldsymbol{b}}^t$:

$$\hat{\boldsymbol{b}}_j^{t,\text{debias}} \stackrel{\text{def}}{=} \underbrace{\hat{\boldsymbol{b}}_j^t}_{\text{iterate}} + \underbrace{\sum_{s=1}^t \hat{w}_{t,s}\big(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}}^s\big)^\top \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j/n}_{\text{bias correction}}$$

- Establish asymptotic normality results for $\boldsymbol{b}_j^*$ using the debiased estimate:

$$\frac{\sqrt{n}(\hat{\boldsymbol{b}}_j^{t,\text{debias}} - \boldsymbol{b}_j^*)}{\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_j\|\sqrt{\hat{r}_t}} \xrightarrow{\text{d}} N(0, 1).$$

Assumptions

A1: The design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ has i.i.d. rows from $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ with an invertible $\boldsymbol{\Sigma}$.

A2: The noise $\boldsymbol{\varepsilon}$ is independent of $\boldsymbol{X}$ and has i.i.d. entries from $N(0, \sigma^2)$.

A3: The asymptotic regime is $n \to \infty$ and $p \to \infty$ with $\frac{p}{n} \le \gamma \in (0, \infty)$.

A4: The algorithm starts with $\hat{\boldsymbol{b}}^1 = \boldsymbol{0}_p$ and $\boldsymbol{g}_t$ in Equation (1) is ζ-Lipschitz with $\boldsymbol{g}_t(0) = 0$.

Theorem 1: Estimation of r_t

Assume conditions A1 - A4 hold, then for any $t \in [T]$:

$$\mathbb{E}[|\hat{r}_t - r_t|] \le \frac{1}{\sqrt{n}} C(\zeta, T, \gamma, \kappa)\mathrm{var}(y_1). \qquad (2)$$

Let $\hat{t} \stackrel{\text{def}}{=} \arg\min_{t\in[T]} \hat{r}_t$. For any $c \in (0, 1/2)$, we have:

$$\mathbb{P}\Big(\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{b}}^{\hat{t}} - \boldsymbol{b}^*)\|^2 \le \min_{s\in[T]} \|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{b}}^s - \boldsymbol{b}^*)\|^2 + \frac{\mathrm{var}(y_1)}{n^{1/2-c}}\Big) \ge 1 - \frac{C(\zeta, \gamma, T, \kappa)}{n^c} \to 1.$$

☺ The proposed estimator $\hat{r}_t$ is $\sqrt{n}$-consistent for r_t.

☺ Minimizing $\hat{r}_t$ can lead to an optimal stopping time with negligible error.

Theorem 2: Inference for b_j^*

Under Assumptions A1 - A4. There exists a set $J_{n,p} \subset [p]$ with $|J_{n,p}| \ge p - \log p$ such that

$$\frac{\sqrt{n}(\hat{\boldsymbol{b}}_j^{t,\text{debias}} - \boldsymbol{b}_j^*)}{\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_j\|\sqrt{\hat{r}_t}} \xrightarrow{\text{d}} N(0, 1) \quad \text{for any } j \in J_{n,p}. \qquad (3)$$

- The asymptotic variance is proportional to $\hat{r}_t$.
- Suggest picking the t with smallest $\hat{r}_t$ for inference tasks.

Summary

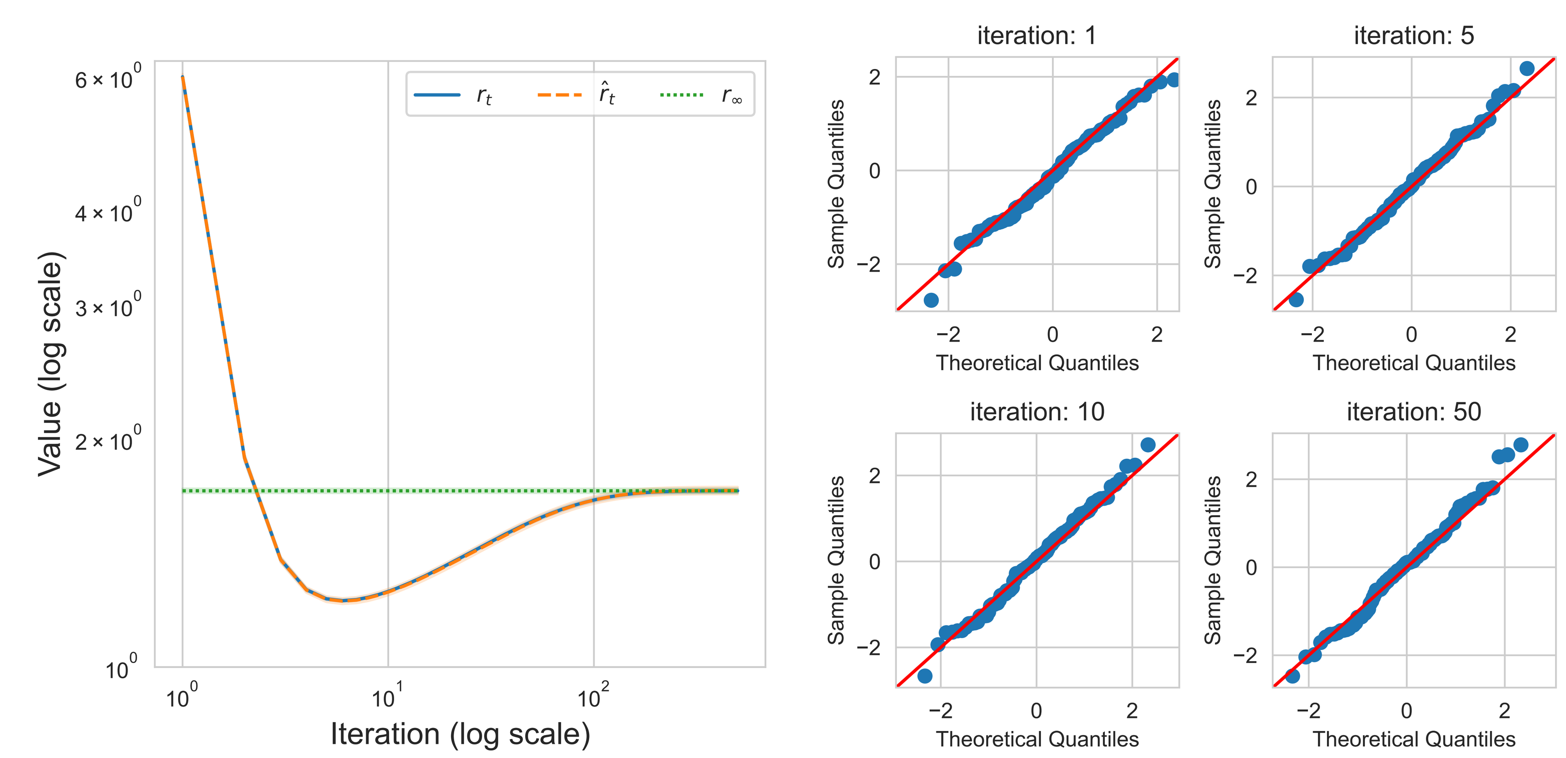
- Proposed a novel $\sqrt{n}$-consistent estimator for the generalization error of iterates along the trajectories of widely used algorithms.
- The form of the estimator depends on a weighted residual vector of all previous iterates. The weights are algorithm-specific and can be efficiently calculated.
- The proposed risk estimators can serve as a proxy for the generalization error, aiding in early stopping decisions.
- Established a valid asymptotic normality result by debiasing each $\hat{\boldsymbol{b}}^t$, which can be used for statistical inferences.

Reference

Bellec, Pierre C., and Kai Tan. Uncertainty quantification for iterative algorithms in linear models with application to early stopping. arXiv preprint arXiv:2404.17856 (2024).

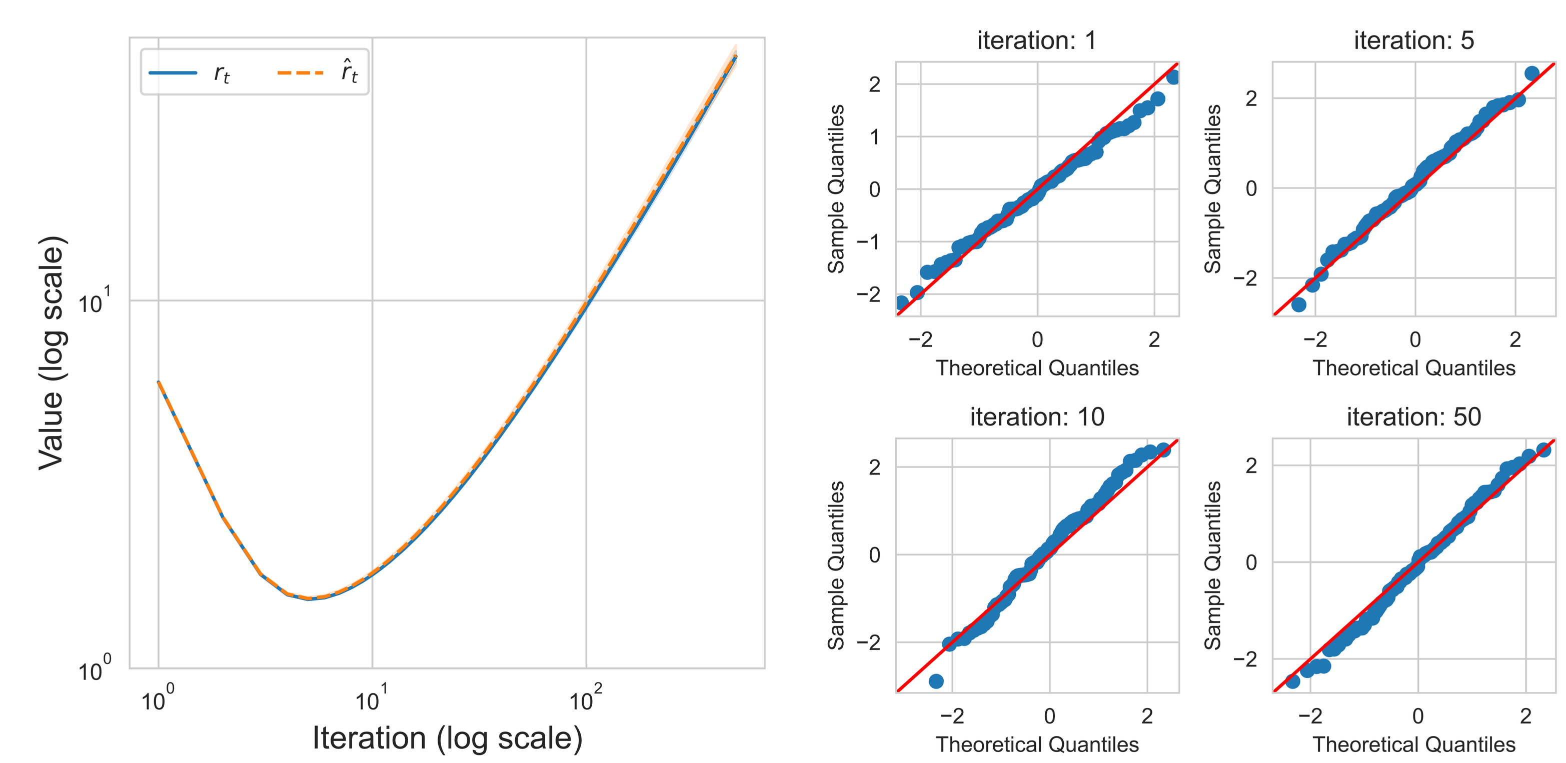
Numerical experiments

The goals: to confirm (1) $\hat{r}_t \approx r_t$, (2) $\frac{\sqrt{n}(\hat{\boldsymbol{b}}_j^{t,\text{debias}} - \boldsymbol{b}_j^*)}{\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_j\|\sqrt{\hat{r}_t}} \xrightarrow{\text{d}} N(0, 1)$ for all t.



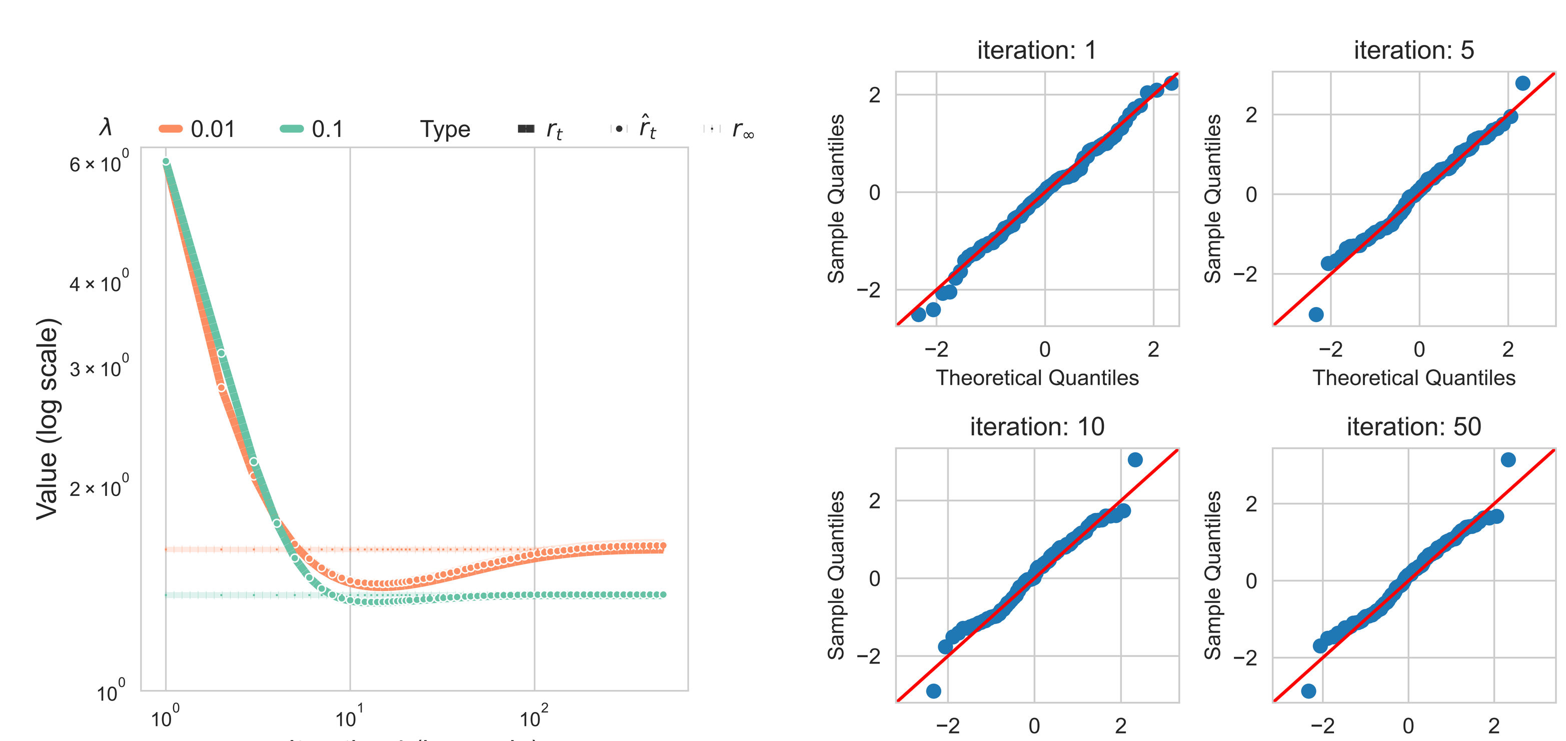
(a) risk curves (b) Q-Q plots at different iteration t

Figure 1. Risk curves and Q-Q plots for **GD** with $(n, p) = (1200, 500)$



(a) risk curves (b) Q-Q plots at different iteration t

Figure 2. Risk curves and Q-Q plots for **AGD** with $(n, p) = (1200, 1200)$



(a) risk curves (b) Q-Q plots at different iteration t

Figure 3. Risk curves and Q-Q plots for **ISTA** with $(n, p) = (1200, 1500)$