PAC-Bayesian Bounds for Learning Partially Observed Stochastic LTI State-Space Systems with Inputs and Sub-Gaussian Noise

Mihály Petreczky,[1] Mohamad Al Ahdab[2], John Leth[2]

[1] CNRS, Ecole Centrale Lille, Univ. Lille, UMR 9189 CRIStAL, France,
[2] Dept. Electonic Systems, Aalborg University, Denmark







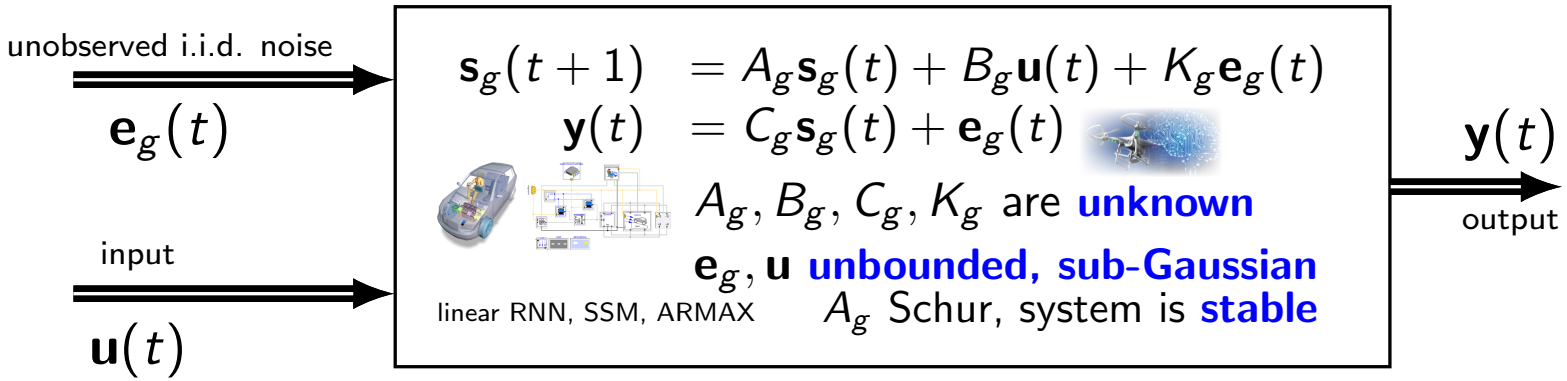




NeurIPS, Paris, France 2026.

Learning problem: ML, system identification (RL),time-series prediction

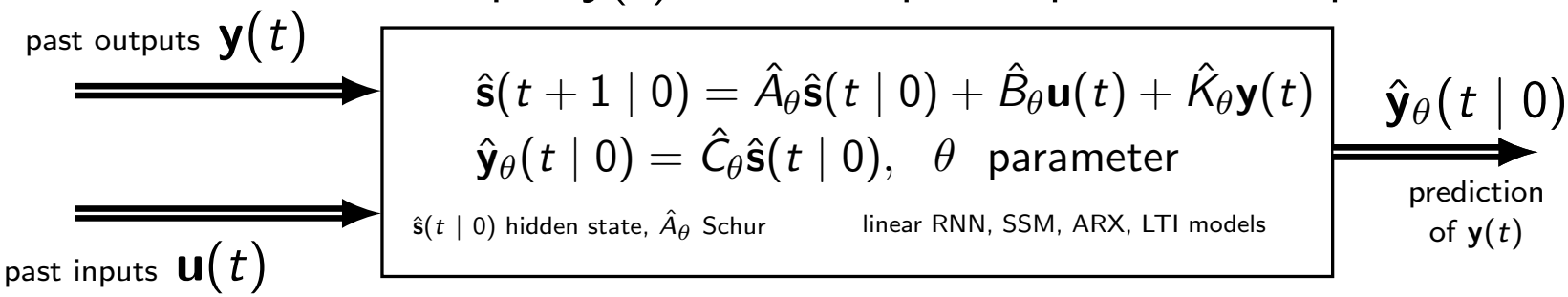
Data generator (unknown): stochastic linear dynamical system (LTI)



From single input/output time-series $\mathcal{D}_N = \{\mathbf{u}(t), \mathbf{y}(t)\}_{t=0}^{N}$ learn a

parametrized stable linear dynamical system which predicts

the current output $\mathbf{y}(t)$ based on past inputs and outputs



$\theta_{true} = (A_g, K_g, B_g, C_g) \overset{\text{Kalman filter}}{\Longrightarrow}$ best (minimum variance) predictor of $\mathbf{y}(t)$

Learning a predictor $\theta \iff$ estimating data generator.

Learning algorithm: empirical risk minimization

Loss function: $\ell(\mathbf{y}(t), \hat{\mathbf{y}}_\theta(t \mid 0))$ measures quality of prediction

- L-Lipschitz, e.g., $\ell(y, y') = \|y - y'\|_2$, or
- Quadratic: $\ell(y, y') = \|y - y'\|_2^2$

True error for a predictor θ: long-term expected prediction error

$$\mathcal{L}(\theta) \triangleq \lim_{t \to \infty} \mathbf{E}[\ell(\mathbf{y}(t), \hat{\mathbf{y}}_\theta(t \mid 0))]$$

Good predictors: $\mathcal{L}(\theta)$ small.

Learning by empirical risk minimization

1. Approximate the true error by the empirical error

$$\hat{\mathcal{L}}_N(\theta) \triangleq \frac{1}{N} \sum_{t=0}^{N-1} \ell(\mathbf{y}(t), \hat{\mathbf{y}}_\theta(t \mid 0)).$$

2. Choose $\theta_\star$ for which the empirical error is small.

Problem: What can we say about the true error for $\theta_\star$

Contribution: Probably Approximately Correct (PAC) (Bayesian) bounds

For a large class of learning algorithms, including:

- regularized empirical error minimization, and
- Bayesian-like methods:

  (a prior density over models is updated based on data, and a model is chosen from the updated density (posterior))

the following inequalities hold with high probability (at least $1 - 2\delta$)

Empirical bound: $\mathcal{L}(\theta_\star) - \hat{\mathcal{L}}_N(\theta_\star) \leq \frac{(\ln(N))^{r+1}}{\sqrt{N}} C_{em}(\theta_\star, \delta)$

over data (and over models if Bayesian).

Contribution: Probably Approximately Correct (PAC) (Bayesian) bounds

For a large class of learning algorithms, including:

- regularized empirical error minimization, and
- Bayesian-like methods:

  (a prior density over models is updated based on data, and a model is chosen from the updated density (posterior))

the following inequalities hold with high probability (at least $1 - 2\delta$)

Empirical bound: $\mathcal{L}(\theta_\star) - \hat{\mathcal{L}}_N(\theta_\star) \leq \frac{(\ln(N))^{r+1}}{\sqrt{N}} C_{em}(\theta_\star, \delta)$

Oracle bound: $\mathcal{L}(\theta_\star) - \min_{\theta \in \Theta} \mathcal{L}(\theta) \leq \frac{(\ln(N))^{r+1}}{\sqrt{N}} C_{or}(\delta)$

over data (and over models if Bayesian).

Contribution: Probably Approximately Correct (PAC) (Bayesian) bounds

For a large class of learning algorithms, including:

- regularized empirical error minimization, and
- Bayesian-like methods:

  (a prior density over models is updated based on data, and a model is chosen from the updated density (posterior))

the following inequalities hold with high probability (at least $1-2\delta$)

Empirical bound: $\mathcal{L}(\theta_\star) - \hat{\mathcal{L}}_N(\theta_\star) \leq \frac{(\ln(N))^{r+1}}{\sqrt{N}} C_{em}(\theta_\star, \delta)$

Oracle bound: $\mathcal{L}(\theta_\star) - \min_{\theta\in\Theta} \mathcal{L}(\theta) \leq \frac{(\ln(N))^{r+1}}{\sqrt{N}} \mathcal{C}_{or}(\delta)$

Parameter estimation error: $\|\theta_\star - \theta_{true}\|_2^2 \leq \frac{(\ln(N))^{r+1}}{\sqrt{N}} \frac{\mathcal{C}_{or}(\delta)}{m_\Theta}$

over data (and over models if Bayesian).

Contribution: Probably Approximately Correct (PAC) (Bayesian) bounds

For a large class of learning algorithms, including:

- regularized empirical error minimization, and
- Bayesian-like methods:

  (a prior density over models is updated based on data, and a model is chosen from the updated density (posterior))

the following inequalities hold with high probability (at least $1 - 2\delta$)

Empirical bound: $\mathcal{L}(\theta_\star) - \hat{\mathcal{L}}_N(\theta_\star) \leq \frac{(\ln(N))^{r+1}}{\sqrt{N}} C_{em}(\theta_\star, \delta)$

Oracle bound: $\mathcal{L}(\theta_\star) - \min_{\theta \in \Theta} \mathcal{L}(\theta) \leq \frac{(\ln(N))^{r+1}}{\sqrt{N}} \mathcal{C}_{or}(\delta)$

Parameter estimation error: $\|\theta_\star - \theta_{true}\|_2^2 \leq \frac{(\ln(N))^{r+1}}{\sqrt{N}} \frac{\mathcal{C}_{or}(\delta)}{m_\Theta}$

H_2 **distance bound:** $\|H_{\theta_{true}} - H_{\theta_\star}\|_{H_2}^2 \leq \frac{(\ln(N))^{r+1}}{\sqrt{N}} \frac{\mathcal{C}_{or}(\delta)}{m_{\mathbf{w}}}$

over data (and over models if Bayesian).

Conclusions

- $O\left(\frac{(\ln(N))^{r+1}}{\sqrt{N}}\right)$ bound on generalisation error & parameter estimation error

Conclusions

- $O\left(\frac{(\ln(N))^{r+1}}{\sqrt{N}}\right)$ bound on generalisation error & parameter estimation error
- r depends on whether data is bounded or ubounded, and whether the loss is Lipschitz or quadratic.

Conclusions

- $O\left(\frac{(\ln(N))^{r+1}}{\sqrt{N}}\right)$ bound on generalisation error & parameter estimation error
- r depends on whether data is bounded or ubounded, and whether the loss is Lipschitz or quadratic.
- Constants

$$\mathcal{L}(\theta_\star) - \hat{\mathcal{L}}_N(\theta_\star) \leq \frac{(\ln(N))^{r+1}}{\sqrt{N}} C_{em}(\theta_\star, \delta)$$

$$\mathcal{L}(\theta_\star) - \min_{\theta \in \Theta} \mathcal{L}(\theta) \leq \frac{(\ln(N))^{r+1}}{\sqrt{N}} C_{or}(\delta)$$

$$\|\theta_\star - \theta_{true}\|_2^2 \leq \frac{(\ln(N))^{r+1}}{\sqrt{N}} \frac{C_{or}(\delta)}{m_\Theta}$$

$$\|H_{\theta_{true}} - H_{\theta_\star}\|_{H_2}^2 \leq \frac{(\ln(N))^{r+1}}{\sqrt{N}} \frac{C_{or}(\delta)}{m_{\mathbf{w}}}$$

depend on stability of the predictors:

more stability $\implies$ smaller constants thus smaller bounds.

Conclusions

- $O\left(\frac{(\ln(N))^{r+1}}{\sqrt{N}}\right)$ bound on generalisation error & parameter estimation error
- r depends on whether data is bounded or ubounded, and whether the loss is Lipschitz or quadratic.
- Constants depend on stability of the predictors:

  more stability $\implies$ smaller constants thus smaller bounds.
- Constants m_Θ, $m_{\mathbf{w}}$ in parameter estimation: identifiability & richness of inputs (persistence of excitation).

$$\|\theta_\star - \theta_{true}\|_2^2 \leq \frac{(\ln(N))^{r+1}}{\sqrt{N}} \frac{\mathcal{C}_{or}(\delta)}{m_\Theta}$$

$$\|H_{\theta_{true}} - H_{\theta_\star}\|_{H_2}^2 \leq \frac{(\ln(N))^{r+1}}{\sqrt{N}} \frac{\mathcal{C}_{or}(\delta)}{m_{\mathbf{w}}}$$

Conclusions

- $O\left(\frac{(\ln(N))^{r+1}}{\sqrt{N}}\right)$ bound on generalisation error & parameter estimation error
- r depends on whether data is bounded or ubounded, and whether the loss is Lipschitz or quadratic.
- Constants depend on stability of the predictors:

  more stability $\Longrightarrow$ smaller constants thus smaller bounds.
- Constants m_Θ, $m_{\mathbf{w}}$ in parameter estimation: identifiability & richness of inputs (persistence of excitation).
- Guarantees for a large number of learning algorithms

Conclusions

- $O\left(\frac{(\ln(N))^{r+1}}{\sqrt{N}}\right)$ bound on generalisation error & parameter estimation error
- r depends on whether data is bounded or ubounded, and whether the loss is Lipschitz or quadratic.
- Constants depend on stability of the predictors:

  more stability $\implies$ smaller constants thus smaller bounds.
- Constants m_Θ, $m_{\mathbf{w}}$ in parameter estimation: identifiability & richness of inputs (persistence of excitation).
- Guarantees for a large number of learning algorithms
- Challenges: non i.i.d. & unbounded data, hidden states.

Conclusions

- $O\left(\frac{(\ln(N))^{r+1}}{\sqrt{N}}\right)$ bound on generalisation error & parameter estimation error
- r depends on whether data is bounded or ubounded, and whether the loss is Lipschitz or quadratic.
- Constants depend on stability of the predictors:

  more stability $\implies$ smaller constants thus smaller bounds.
- Constants m_Θ, $m_{\mathbf{w}}$ in parameter estimation: identifiability & richness of inputs (persistence of excitation).
- Guarantees for a large number of learning algorithms
- **Future work:** tighter bounds, experimental evaluation, extension to RNNs.