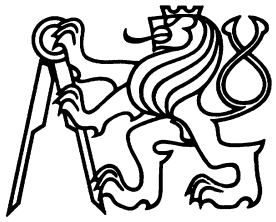
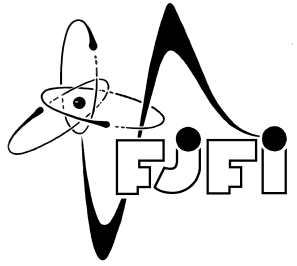


Czech Technical University in Prague
Faculty of Nuclear Sciences and Physical Engineering



The Role of a priori Distributions for Sparse Parametrizations of Models

Bc. Lukáš Kulička

Student's Scientific Conference 2021
Supported from CTU Grant SVK 29/21/F4

June 24, 2021

Contents

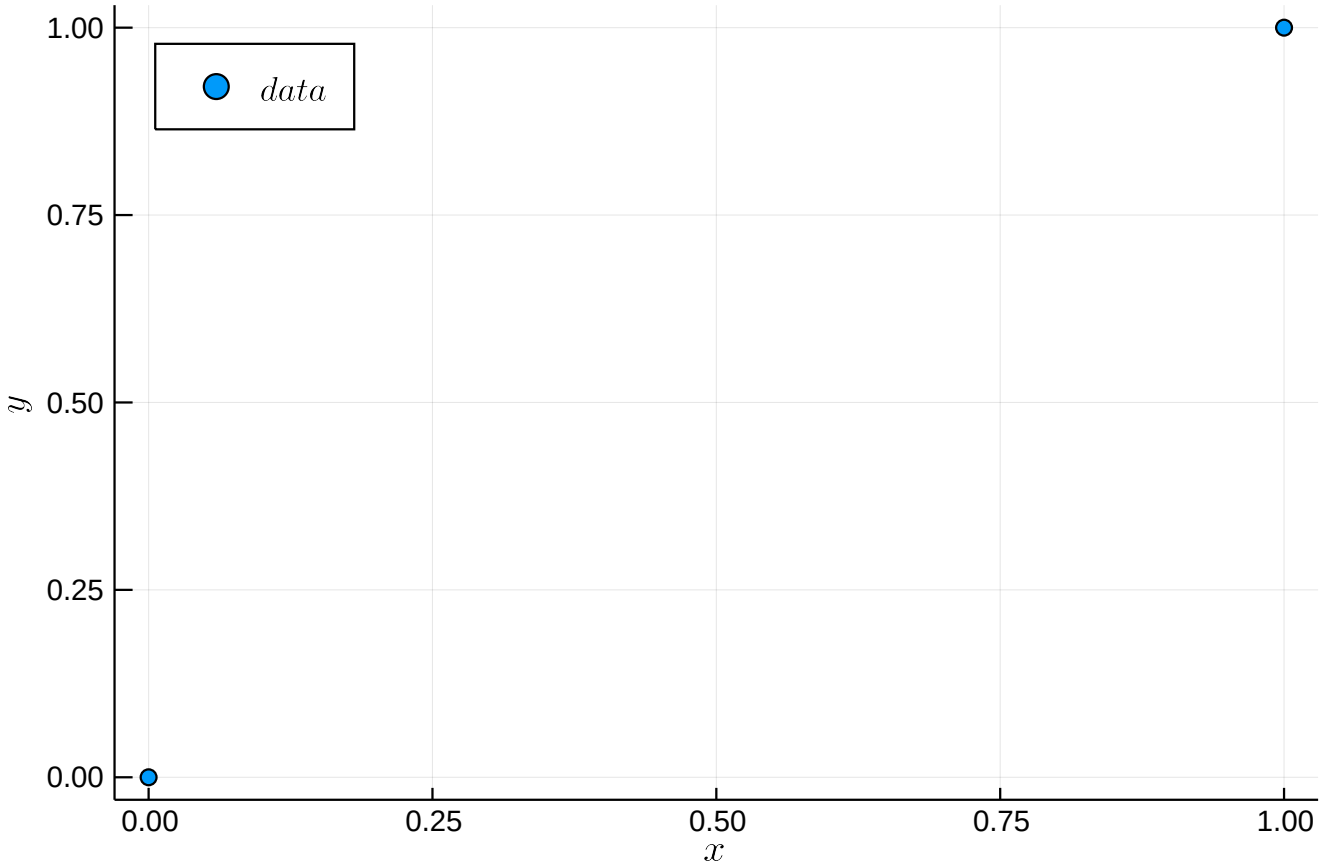


Figure 1: Example of sparse data space.

- Least squares problem:

$$\begin{aligned} y_1 &= w_1 x_{11} + w_2 x_{12}^2 \\ &\Rightarrow \quad \mathbb{X} = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} \\ y_2 &= w_1 x_{21} + w_2 x_{22}^2 \end{aligned} \tag{1}$$

- $h(\mathbb{X}) = 1$.
- How to find $\hat{\boldsymbol{\theta}}$?

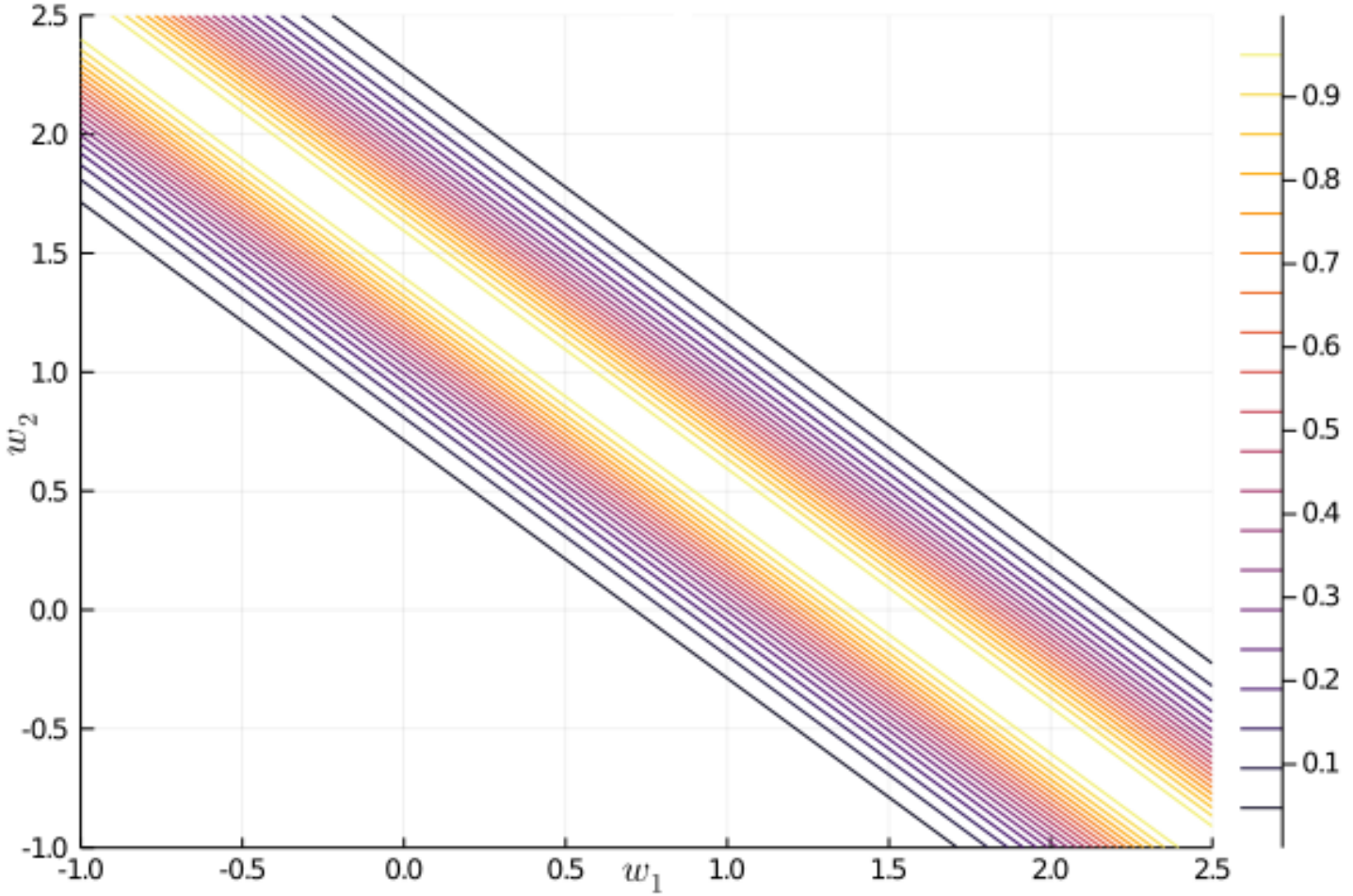


Figure 2: Contour plot of parameters likelihood w_1, w_2 from Eq. (1).

Bayesian Approach

- Bayes' rule:

$$\boxed{\text{Aposteriori distribution} \propto \text{Likelihood} \times \text{Apriori distribution.}}$$

- L_1 norm:

$$w_j \sim \text{Laplace}\,(0, \lambda) \propto \exp\left(-\lambda \|\boldsymbol{\theta}\|_1\right), \tag{2}$$

  we choose λ arbitrarily firmly.

- ARD:

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\alpha}^{-1}\mathbb{I}) \ \wedge \ p(\alpha_j) = \text{St}(0, \sigma^2, \nu), \ \forall j \in \hat{d} \tag{3}$$

- Spike & Slab:

$$w_j \sim \lambda_j \mathcal{N}(0, c^2) + (1 - \lambda_j)\mathcal{N}(0, \epsilon^2), \tag{4}$$

  where $\epsilon \ll c$ and $\lambda_j \in \{0, 1\} \sim$ Bernoulli

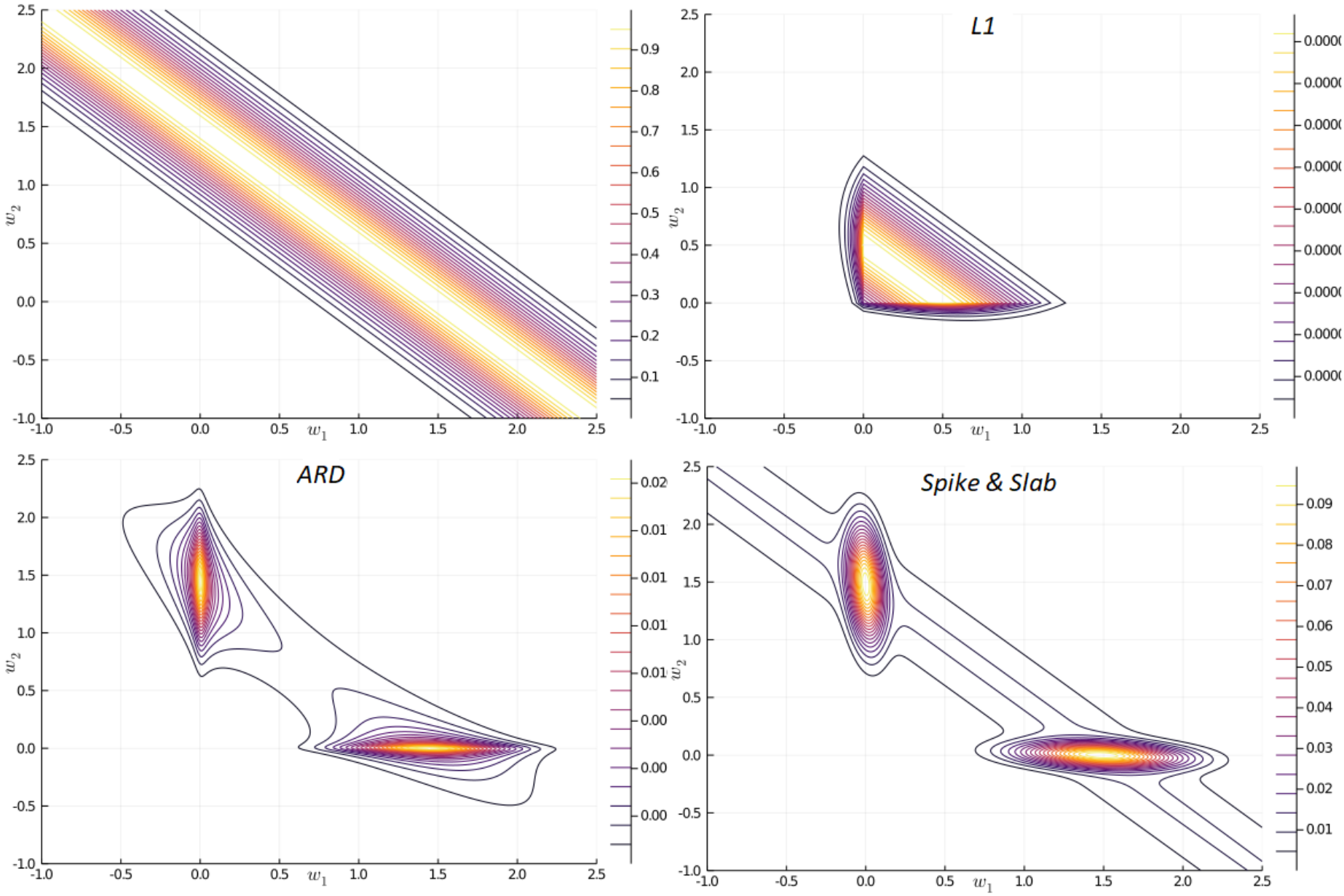


Figure 3: Contour plots of likelihood with priors (2), (3), (4).

Mathematical Tools – ELBO

- Evidence Lower Bound:

$$\begin{aligned}\ln p(\mathbf{x}) = &\int q(\mathbf{z}_{\boldsymbol{\theta}}) \left(\ln \left\{ \frac{p(\mathbf{x}, \mathbf{z}_{\boldsymbol{\theta}})}{q(\mathbf{z}_{\boldsymbol{\theta}})} \right\} \right) d\mathbf{z}_{\boldsymbol{\theta}} - \int q(\mathbf{z}_{\boldsymbol{\theta}}) \left(\ln \left\{ \frac{p(\mathbf{z}_{\boldsymbol{\theta}}|\mathbf{x})}{q(\mathbf{z}_{\boldsymbol{\theta}})} \right\} \right) d\mathbf{z}_{\boldsymbol{\theta}} \\ =&\mathcal{L}(q(\mathbf{z}_{\boldsymbol{\theta}})) + KL(q(\mathbf{z}_{\boldsymbol{\theta}}) \| p(\mathbf{z}_{\boldsymbol{\theta}}|\mathbf{x}))\end{aligned} \tag{5}$$

- $p(\mathbf{z}_{\boldsymbol{\theta}}|\mathbf{x})$ – unknown.
- $q(\mathbf{z}_{\boldsymbol{\theta}}|\mathbf{x})$ – known, chosen.

$$\boxed{\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathcal{L}} \tag{6}$$

Mathematical Tools – Reparameterization Trick

- Optimalization $\mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{z})}[f(\mathbf{z})]$.

$$\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{z}) = \varepsilon \sim q(\varepsilon) \qquad \mathbf{z} = \mathcal{S}_{\boldsymbol{\theta}}^{-1}(\varepsilon) \tag{7}$$

$$\mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{z})}[f(\mathbf{z})] = \mathbb{E}_{q(\varepsilon)}\left[f(\mathcal{S}_{\boldsymbol{\theta}}^{-1}(\varepsilon))\right] \tag{8}$$

- $\mathcal{N}(\mu, \sigma^2) \Rightarrow \mathcal{S}_{\mu,\sigma^2}(\mathbf{z}) = \frac{\mathbf{z}-\mu}{\sigma^2} \sim \mathcal{N}(0,1)$.

$$\mathbb{E}_{\mathcal{N}(\mu,\sigma^2)}[f(\mathbf{z})] = \mathbb{E}_{\mathcal{N}(0,1)}\left[f(\mu + \varepsilon\sigma^2)\right] \tag{9}$$

- Non-linear tasks.

Classic Regression

- Probability model

$$p(\mathbf{y}, \theta \mid \mathbb{X}, \alpha, \omega) = \mathcal{N}(\mathbb{X}\theta, \omega\mathbb{I})\mathcal{N}(0, \mathrm{diag}(\boldsymbol{\alpha})) \prod_i \Gamma(\gamma_0, \delta_0) \tag{10}$$

- ETEX data
- ARD with $\alpha = 0.1$, $\omega = 5$ and $\Gamma(0.1, 0.1)$ prior.

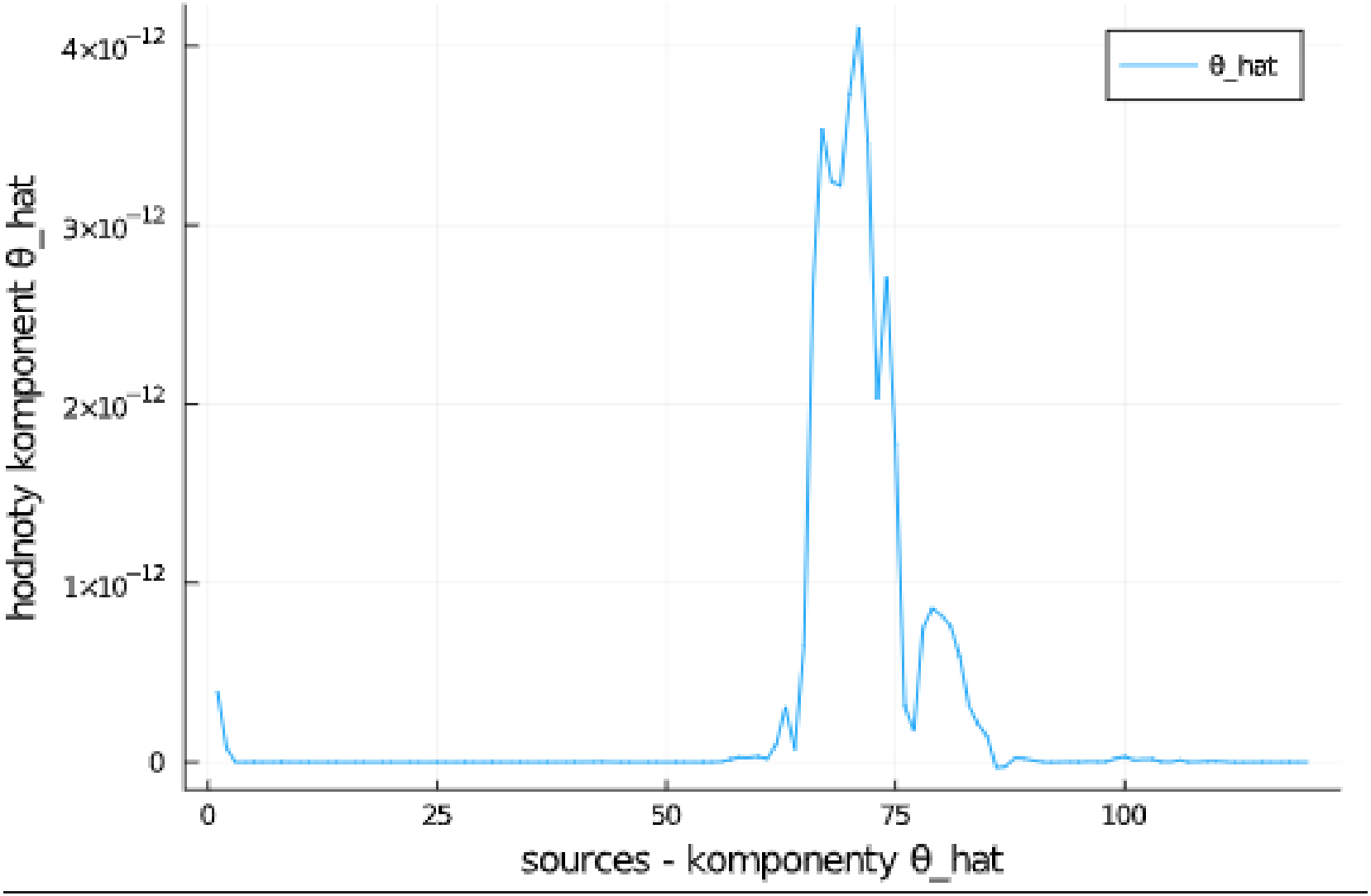


Figure 4: Finding sparse solution for ETEX data with ARD.

Logistic Regression

- 1000 generated observations from $\mathcal{N}_2(\mathbf{0}, \mathbb{I})$. Let $\boldsymbol{\theta}^{(\text{true})} = (0, 10)$.
- Generating of $\mathbf{y}^{(\text{true})}$:

$$\mathbf{y}^{(\text{true})} = \boldsymbol{\sigma}\left(\mathbb{X} \cdot \boldsymbol{\theta}^{(\text{true})\,T}\right) \overset{\text{Bernoulli}}{\Rightarrow} \text{binary-class vector.} \tag{11}$$

- Model:

$$\text{model}(\boldsymbol{\theta}) = \boldsymbol{\sigma}\left(\mathbb{X} \cdot \boldsymbol{\theta}^{T}\right). \tag{12}$$

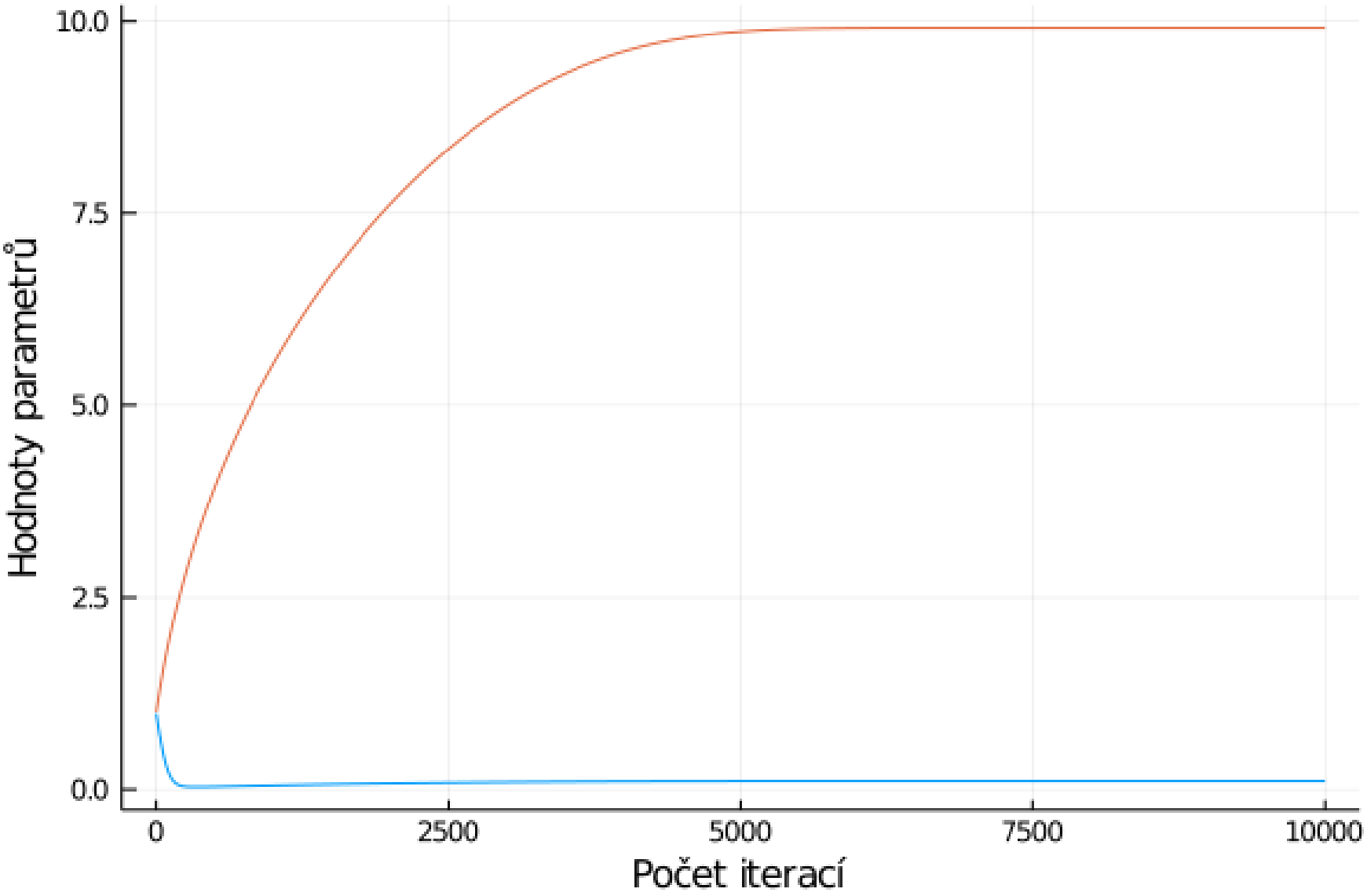
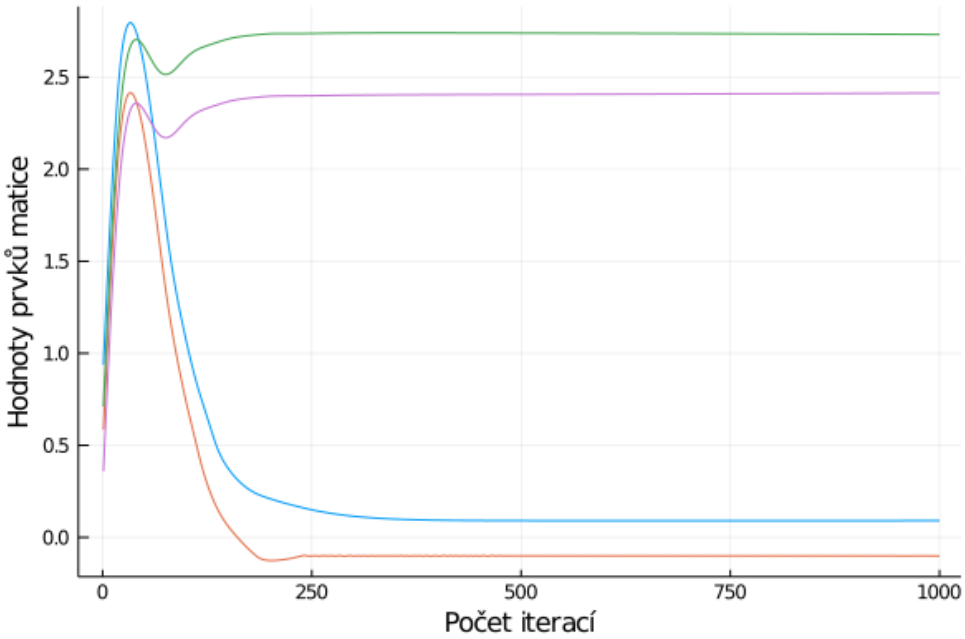


Figure 5: Learning process of θ with L_1 norm.

Tree Structures

- Model:

$$\mathbf{y} = \boldsymbol{\sigma}\left(w_1\mathbf{x}_1 + \max(\mathbb{W} \cdot \mathbb{Z})\right) \tag{13}$$



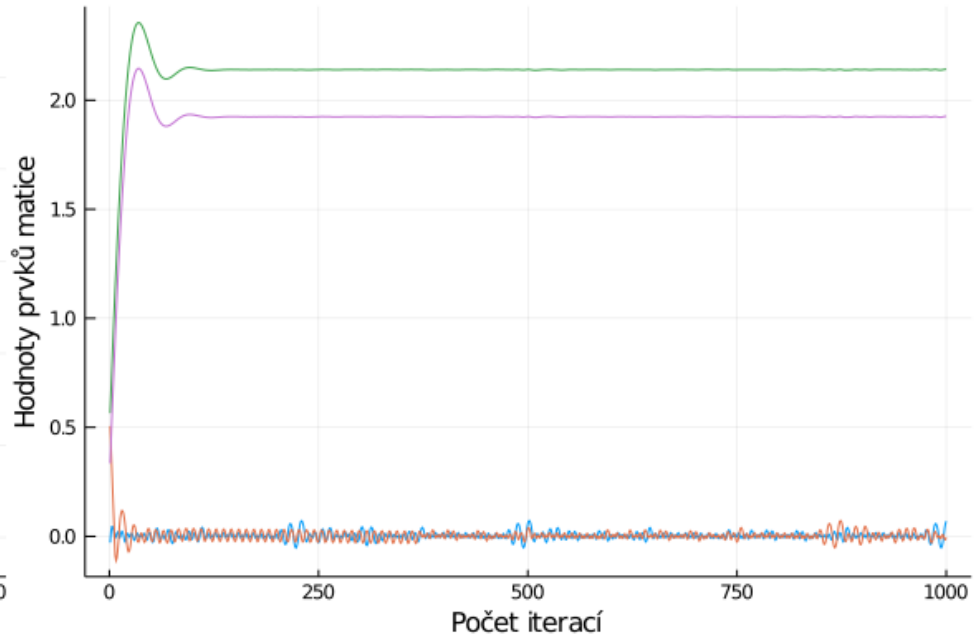


Figure 6: $\mathbb{W}^{(\text{true})} = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}$, right with penalty L_1 norm ($\lambda = 0.01$).

Thank you for your attention.

References

- KULIČKA, Lukáš. *Klasifikace dat popsaných stromovou strukturou*. Praha, 2020. Bakalářská práce. České vysoké učení technické v Praze, Fakulta jaderná a fyzikálně inženýrská, Katedra matematiky. Vedoucí práce doc. Ing. Václav Šmídl, Ph.D.

Author: Bc. Lukáš Kulička

Title: The Role of a priori Distributions for Sparse Parametrizations of Models

Programme: Applied Mathematical Stochastic Methods

Student's Scientific Conference 2021

Supported from CTU Grant SVK 29/21/F4

Lukas.Kulicka@cvut.cz