Softplus Regressions and Convex Polytopes

Mingyuan Zhou

IROM Department, McCombs School of Business
The University of Texas at Austin

University of Electronic Science and Technology of China
Chengdu, China, December 13, 2016

Binary classification

- Linear classifier:
  - Logistic regression
  - Probit regression
  - Use a single hyperplane to partition the covariate space into two halves
- Nonlinear classifier:
  - Use the kernel trick:
    - Choose a subset of covariate vectors as support vectors
    - Compute a sample's kernel distances to these support vectors
    - Regress the label on the kernel distances
  - Use a deep neural network
    - Transform the covariates with a deep neural network
    - Regress on the transformed covariates

Nonlinear binary classification

- Both kernel learning and deep learning map the original covariates into a more linearly separable space, transforming a nonlinear classification problem into a linear one.
- Kernel learning based methods such as kernel support vector machines are not scalable as the number of support vectors often increases linearly in the size of the training dataset.
- A deep neural network often requires tuning the network structure
  - the number of layers
  - the number of hidden units of each layer
  - the type of nonlinear activation functions

Motivations

- Exploit two distinct types of interactions—noisy-OR and noisy-AND—between hyperplanes to define flexible nonlinear classification decision boundaries directly on the original covariate space.
- Attribute a binary outcome to multiple hidden causes, each of which is associated with an activation probability function produced by a single hyperplane or the collaboration of multiple ones.
- Investigate the potential of using multiple hyperplanes to construct nonlinear classifiers.
- Increase the classification margin by using multiple hyperplanes to enclose one class.
- The noisy-OR and/or noisy-AND interactions of hyperplanes make it simple to interpret and quantify how each hyperplane contributes to the final classification decision boundaries.

Model construction

- Exploit the convolution and stacking operations on the gamma distributions with covariate-dependent scale parameters.
  - Convolution operation convolves differently parameterized probability distributions to increase representation power and enhance smoothness
  - Stacking operation mixes a distribution in the stack with a distribution of the same family that is subsequently pushed into the stack.
- Provide interpretable geometric constraints, which are related to either a single or a union of convex polytopes, on the classification decision boundaries defined on the original covariate space.
- The proposed nonparametric Bayesian softplus regressions naturally provide probability estimates, automatically learn the complexity of the predictive distribution, and quantify model uncertainties with posterior samples.

Bernoulli-Poisson link

- Bernoulli-Poisson link: $y = \delta(m \geq 1),\ m \sim \text{Pois}(\lambda)$
- The marginalization of the latent count m leads to

$$y \sim \text{Bernoulli}(p),\ p = 1 - e^{-\lambda}$$

We now refer to $\lambda = -\ln(1-p)$ as the Bernoulli-Poisson (BerPo) rate for y and simply denote the above equation as $y \sim \text{BerPo}(\lambda)$.

- It is instructive to notice that $1/(1+e^{-x}) = 1 - \exp[-\ln(1+e^{x})]$, and hence letting

$$y \sim \text{Bernoulli}\big[\sigma(x)\big],\ \sigma(x) = 1/(1+e^{-x})$$

is equivalent to letting

$$y \sim \text{BerPo}\big[\varsigma(x)\big],\ \varsigma(x) = \ln(1+e^{x}).$$

Softplus function

- $\varsigma(x) = \ln(1 + e^x)$ is the softplus function.
- The softplus function is a smoothed version of the rectifier, or rectified linear unit

$$\mathrm{ReLU}(x) = \max(0, x).$$

- The rectifier function is now widely used in deep neural networks, replacing other canonical nonlinear activation functions such as the sigmoid and hyperbolic tangent functions.

A family of softplus functions

- Stack-softplus function:

$$\varsigma(x_1, \ldots, x_t) = \ln\left(1 + e^{x_t} \ln\left\{1 + e^{x_{t-1}} \ln\left[1 + \ldots \ln\left(1 + e^{x_1}\right)\right]\right\}\right)$$

  Recursive definition: $\varsigma(x_1, \ldots, x_t) = \ln[1 + e^{x_t}\varsigma(x_1, \ldots, x_{t-1})]$.

- Sum-softplus:

$$\sum_{k=1}^{\infty} r_k \varsigma(x_k),$$

  where r_k are the countably infinite weights of a gamma process.

- Sum-stack-softplus (SS-softplus) function:

$$\sum_{k=1}^{\infty} r_k \varsigma(x_{k1}, \ldots, x_{kt}).$$

- While the softplus function is monotonic, the stack-, sum-, and SS-softplus functions could produce a single peak, a single valley, and multiple change points, respectively, along the real line.

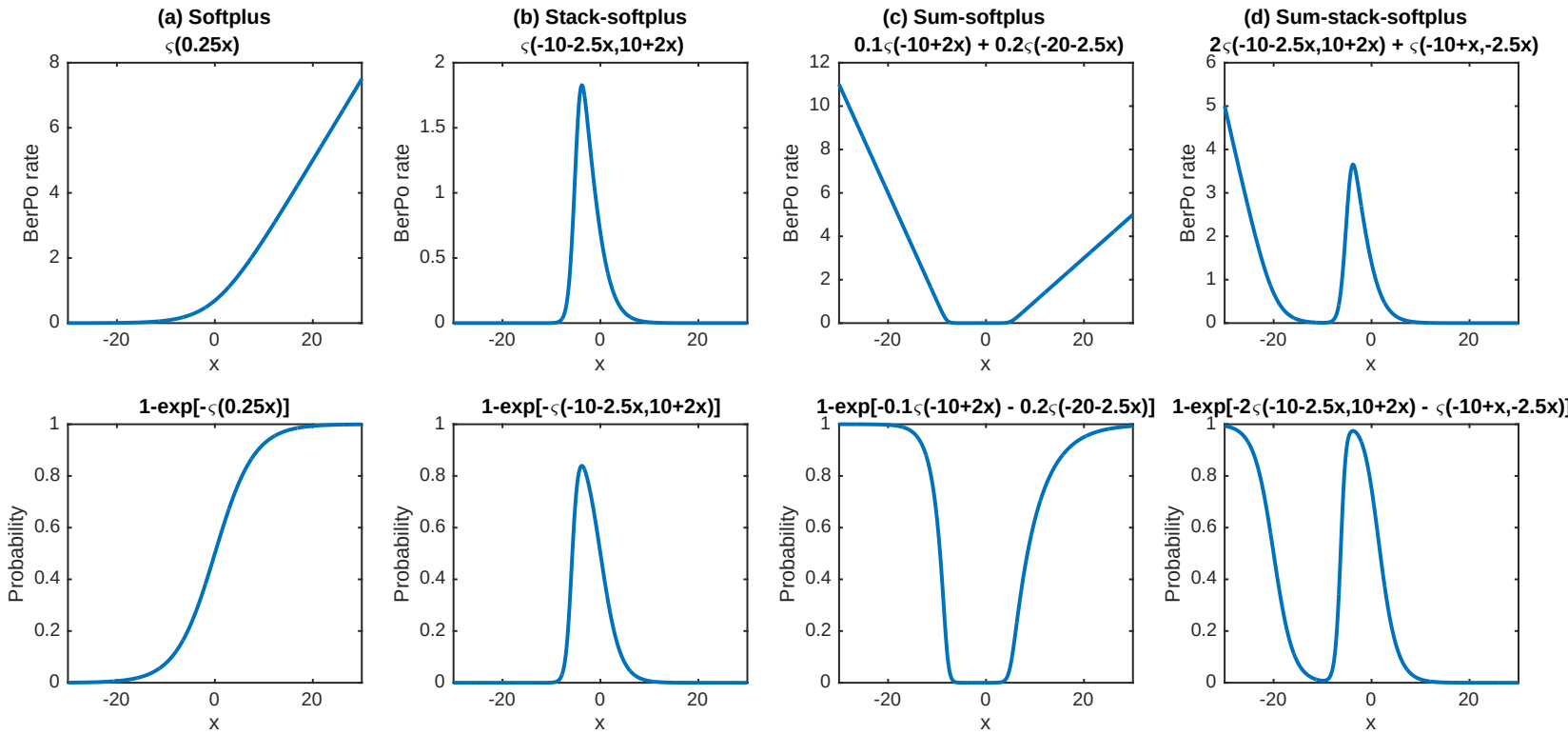

Figure: Columns from left to right illustrate softplus, stack-softplus, sum-softplus, and sum-stack-softplus functions, respectively, on the real line $x \in (-\infty, \infty)$. The first and second rows of each column illustrate a softplus function $\lambda(x)$ and its corresponding probability $1 - e^{-\lambda(x)}$, respectively.

A family of softplus regressions

- For the ith covariate vector $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{iV})' \in \mathbb{R}^{V+1}$, we model its binary class label $y_i \in \{0, 1\}$ using

$$y_i \,|\, \boldsymbol{x}_i \sim \text{BerPo}[\lambda(\boldsymbol{x}_i)].$$

- Equivalent representation:

$$y_i \,|\, \boldsymbol{x}_i \sim \text{Bernoulli}(p_i), \;\; p_i = 1 - e^{-\lambda(\boldsymbol{x}_i)}.$$

- $\lambda(\boldsymbol{x}_i)$ is a nonnegative deterministic function of $\boldsymbol{x}_i$ that may contain countably infinite parameters drawn from a completely random measure.

Definition (Softplus regression)

Given $\boldsymbol{x}_i$, weight $r \in \mathbb{R}_+$, and a regression coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^{V+1}$, softplus regression parameterizes $\lambda(\boldsymbol{x}_i)$ using a softplus function as

$$\lambda(\boldsymbol{x}_i) = r\varsigma(\boldsymbol{x}_i'\boldsymbol{\beta}) = r\ln(1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}}).$$

- Softplus regression is equivalent to the binary regression model

$$y_i \sim \text{Bernoulli}(p_i), \quad p_i = 1 - \left(1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}}\right)^{-r}.$$

- It can be constructed using the hierarchical model

$$y_i = \delta(m_i \geq 1), \; m_i \sim \text{Pois}(\theta_i), \; \theta_i \sim \text{Gamma}\left(r, e^{\boldsymbol{x}_i'\boldsymbol{\beta}}\right).$$

- Softplus regression with $r = 1$ is equivalent to logistic regression

$$y_i \sim \text{Bernoulli}[1/(1 + e^{-\boldsymbol{x}_i'\boldsymbol{\beta}})],$$

  which uses a single hyperplane dividing the covariate space into two halves to separate one class from the other.
- With $p_0 \in (0, 1)$ defined as the probability threshold to make a binary decision, softplus regression defines a hyperplane to partition the V dimensional covariate space into two halves:

$$y_i = \begin{cases} 1, & \text{if } \boldsymbol{x}_i'\boldsymbol{\beta} > \ln\left[(1 - p_0)^{-1/r} - 1\right] \\ 0, & \text{otherwise} \end{cases}$$

- The three following generalizations all partition the covariate space using a confined space that is related to a single convex polytope or the union of multiple convex polytopes.

Definition (Sum-softplus regression)

Given a gamma process draw $G = \sum_{k=1}^{\infty} r_k \delta_{\boldsymbol{\beta}_k}$, sum-softplus regression parameterizes $\lambda(\boldsymbol{x}_i)$ using a sum-softplus function as

$$\lambda(\boldsymbol{x}_i) = \sum_{k=1}^{\infty} r_k \, \varsigma(\boldsymbol{x}_i' \boldsymbol{\beta}_k) = \sum_{k=1}^{\infty} r_k \ln(1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k}).$$

- Sum-softplus regression is equivalent to a noisy-OR binary regression model

$$y_i \sim \text{Bernoulli}\left[1 - \prod_{k=1}^{\infty}(1 - p_{ik})\right], \quad p_{ik} = 1 - \left(\frac{1}{1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k}}\right)^{r_k}.$$

- It can be constructed using the hierarchical model

$$y_i = \delta(m_i \geq 1), \; m_i \sim \text{Pois}(\theta_i), \; \theta_i = \sum_{k=1}^{\infty} \theta_{ik}, \; \theta_{ik} \sim \text{Gamma}\left(r_k, e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k}\right).$$

- Sum-softplus regression can also be constructed with

$$y_i = \delta(m_i \geq 1), \quad m_i = \sum_{k=1}^{\infty} m_{ik}, \; m_{ik} \sim \text{NB}\left[r_k, 1/(1+e^{-\boldsymbol{x}_i'\boldsymbol{\beta}_k})\right],$$

Proposition

The infinite product

$$e^{-\sum_{k=1}^{\infty} r_k \varsigma(\boldsymbol{x}_i'\boldsymbol{\beta}_k)} = \prod_{k=1}^{\infty} \left(1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}_k}\right)^{-r_k}$$

in sum-softplus regression is smaller than one and has a finite expectation that is greater than zero.

Geometric constraint of sum-softplus regression

Sum-softplus regression:

- uses the **noisy-OR** hyperplane interactions
  to define a **convex-polytope**-bounded confined space
  to enclose **negative** examples (i.e., data samples with $y_i = 0$)

Theorem

For sum-softplus regression, the confined space specified by the inequality $P(y_i = 1 \mid \boldsymbol{x}_i) = 1 - e^{-\lambda(\boldsymbol{x}_i)} \leq p_0$, *which can be expressed as*

$$\lambda(\boldsymbol{x}_i) = \sum_{k=1}^{\infty} r_k \ln(1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k}) \leq -\ln(1 - p_0), \tag{1}$$

is bounded by a convex polytope defined by the set of solutions to countably infinite inequalities $p_{ik} \leq p_0$ *that can be expressed as*

$$\boldsymbol{x}_i' \boldsymbol{\beta}_k \leq \ln\left[(1 - p_0)^{-1/r_k} - 1\right], \quad k \in \{1, 2, \ldots\}. \tag{2}$$

Proposition

For any data point $\boldsymbol{x}_i$ *that resides outside the convex polytope defined by* (2), *which means* $\boldsymbol{x}_i$ *violates at least one of the inequalities in* (2) *a.s., it will be labeled under sum-softplus regression with* $y_i = 1$ *with a probability greater than* p_0, *and* $y_i = 0$ *with a probability no greater than* $1 - p_0$.

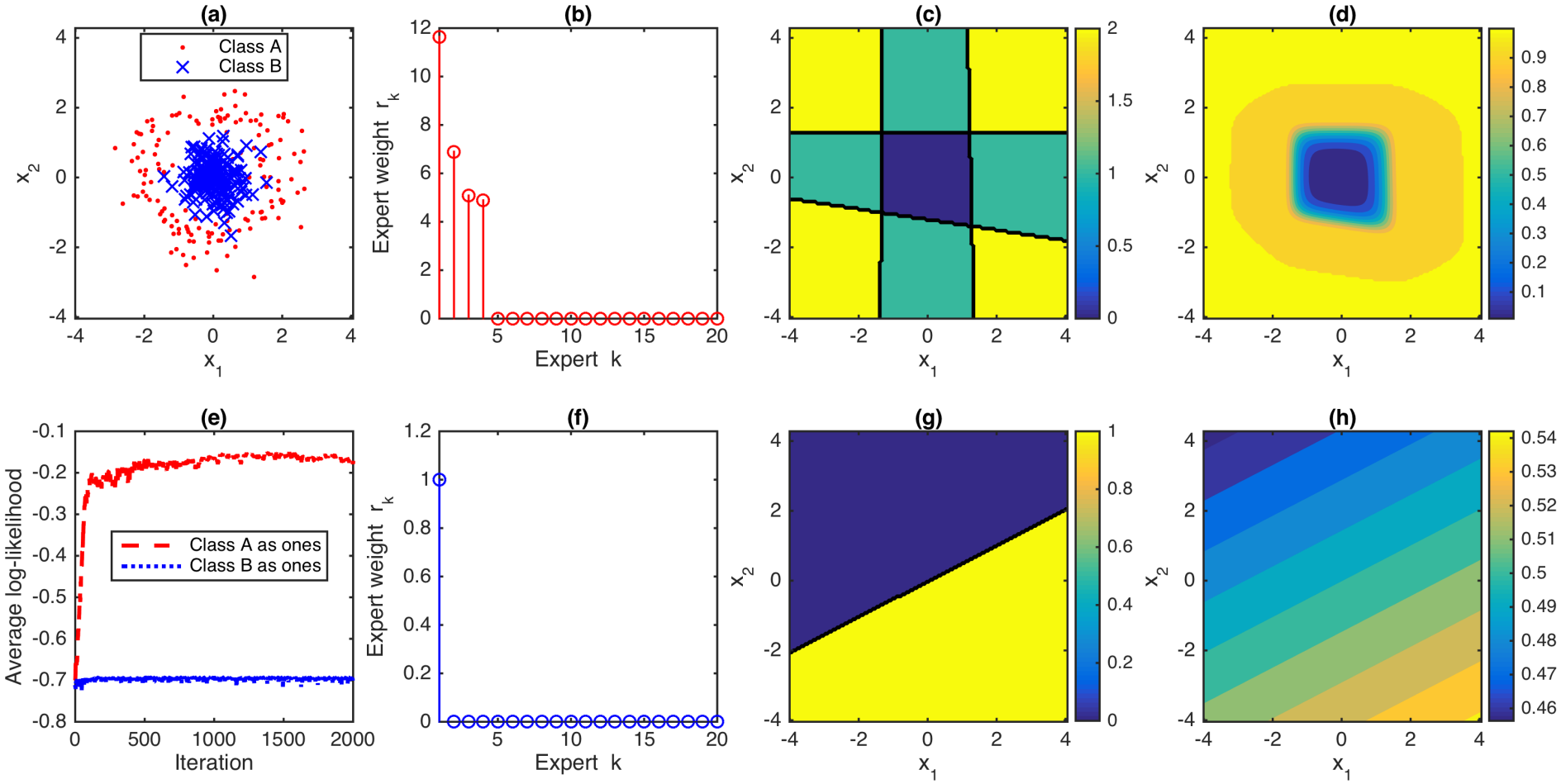

Figure: Visualization of sum-softplus regression with $K_{\max} = 20$ experts on a binary classification problem under two opposite labeling settings.
First row: Red and Blue points are labeled as "1" and "0," respectively.
Second row: Blue and Red points are relabeled as "1" and "0," respectively.

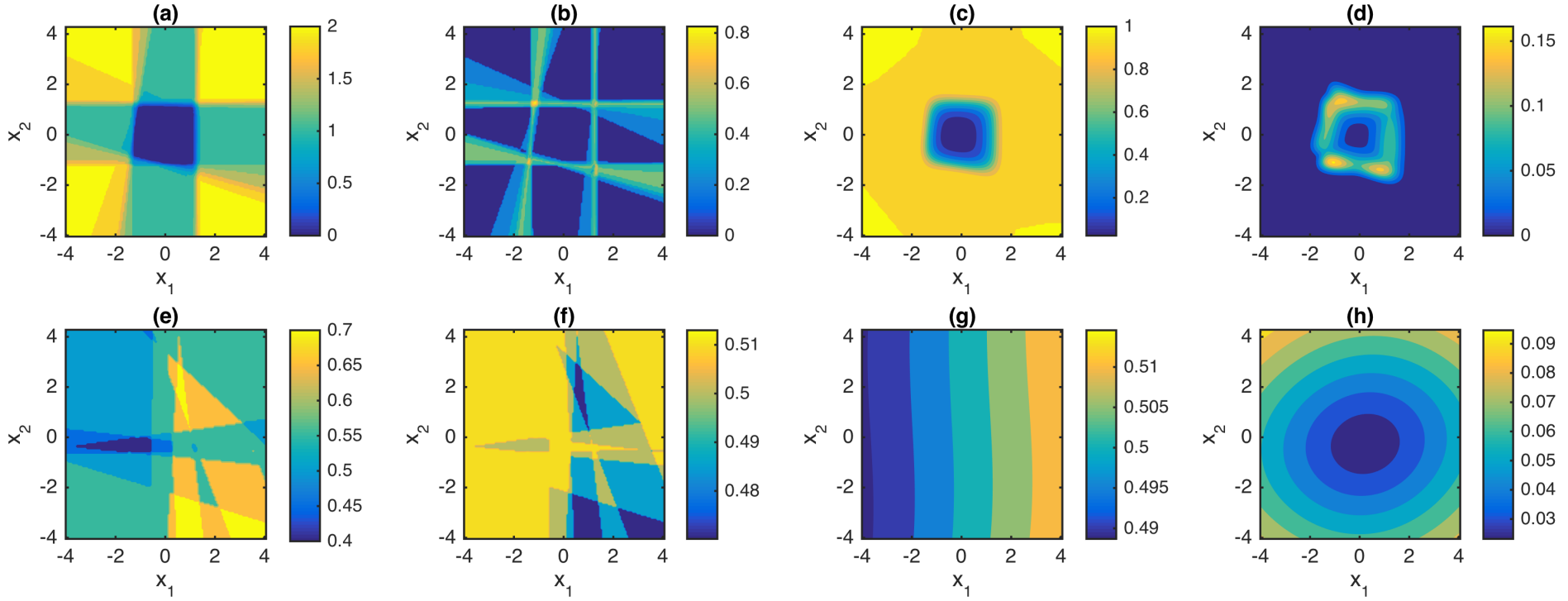

Figure: Visualization of the posteriors of sum-softplus regression based on 20 MCMC samples, collected once per every 50 iterations during the last 1000 MCMC iterations. (a) and (b) show the contour maps of the posterior means and standard deviations, respectively, of the number of inequalities specified in (2) that are violated, and (c) and (d) show the contour maps of the posterior means and standard deviations, respectively, of predicted class probabilities. (e)-(h) are analogous plots to (a)-(d), with the data points in Classes A and B relabeled as "0" and "1," respectively.

Definition (Stack-softplus regression)

With weight $r \in \mathbb{R}_+$ and T regression coefficient vectors $\boldsymbol{\beta}^{(2:T+1)} := (\boldsymbol{\beta}^{(2)}, \ldots, \boldsymbol{\beta}^{(T+1)}) \in \mathbb{R}^{(V+1)\times T}$, stack-softplus regression with T layers parameterizes $\lambda(\boldsymbol{x}_i)$ using a stack-softplus function as

$$\lambda(\boldsymbol{x}_i) = r\,\varsigma\left(\boldsymbol{x}_i'\boldsymbol{\beta}^{(2)}, \ldots, \boldsymbol{x}_i'\boldsymbol{\beta}^{(T+1)}\right)$$

$$= r \ln\left(1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}^{(T+1)}} \ln\left\{1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}^{(T)}} \ln\left[1 + \ldots \ln\left(1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}^{(2)}}\right)\right]\right\}\right).$$

- Stack-softplus regression is equivalent to a noisy-AND regression model

$$y_i \sim \text{Bernoulli}(p_i),$$

$$p_i = 1 - \left(1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}^{(T+1)}} \ln\left\{1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}^{(T)}} \ln\left[1 + \ldots \ln\left(1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}^{(2)}}\right)\right]\right\}\right)^{-r}.$$

- Stack-softplus regression can be constructed using the hierarchical model that stacks T gamma distributions, whose scales are differently parameterized by the covariates, as

$$\theta_i^{(T)} \sim \text{Gamma}\left(r, e^{\boldsymbol{x}_i'\boldsymbol{\beta}^{(T+1)}}\right),$$

$$\cdots$$

$$\theta_i^{(t)} \sim \text{Gamma}\left(\theta_i^{(t+1)}, e^{\boldsymbol{x}_i'\boldsymbol{\beta}^{(t+1)}}\right),$$

$$\cdots$$

$$y_i = \delta(m_i \geq 1),\ m_i \sim \text{Pois}(\theta_i^{(1)}),\ \theta_i^{(1)} \sim \text{Gamma}\left(\theta_i^{(2)}, e^{\boldsymbol{x}_i'\boldsymbol{\beta}^{(2)}}\right).$$

Geometric constraint of stack-softplus regression

Stack-softplus regression:

- uses the **noisy-AND** hyperplane interactions
  to define a **convex-polytope**-like confined space
  to enclose **positive** examples (i.e., data samples with $y_i = 1$)

- Define g_t with $g_1 = 1$ and the recursion $g_t = \ln(1 + g_{t-1})$ for $t = 2, \ldots, T$
- Define h_t with $h_{T+1} = (1 - p_0)^{-\frac{1}{r}} - 1$ and the recursion $h_t = e^{h_{t+1}} - 1$ for $t = T, T-1, \ldots, 2$.
- The confined space of stack-softplus regression can be roughly related to a convex polytope specified by the solutions to a set of T inequalities as

$$\boldsymbol{x}_i' \boldsymbol{\beta}^{(t)} > \ln(h_t) - \ln(g_{t-1}), \ t \in \{2, \ldots, T+1\}. \tag{3}$$

- The convex polytope is enclosed by the intersection of T V-dimensional hyperplanes, and since none of the T criteria would be strongly violated inside the convex polytope, the label $y_i = 1$ ($y_i = 0$) would be assigned to an $\boldsymbol{x}_i$ inside (outside) the convex polytope with a relatively high (low) probability.

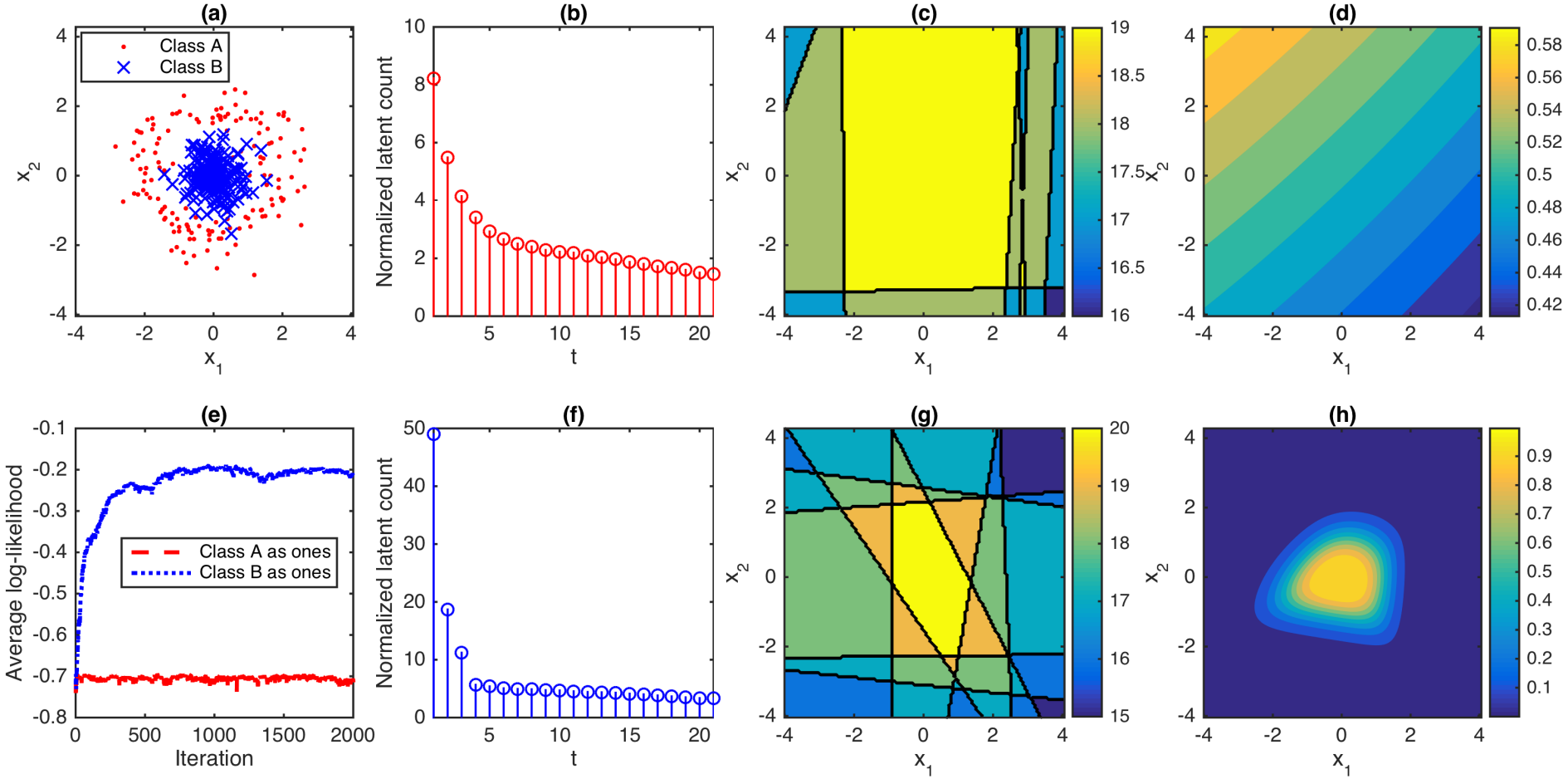

Figure: Visualization of stack-softplus regression with $T = 20$ expert criteria on a binary classification problem under two opposite labeling settings.
First row: Red and Blue points are labeled as "1" and "0," respectively.
Second row: Blue and Red points are relabeled as "1" and "0," respectively.

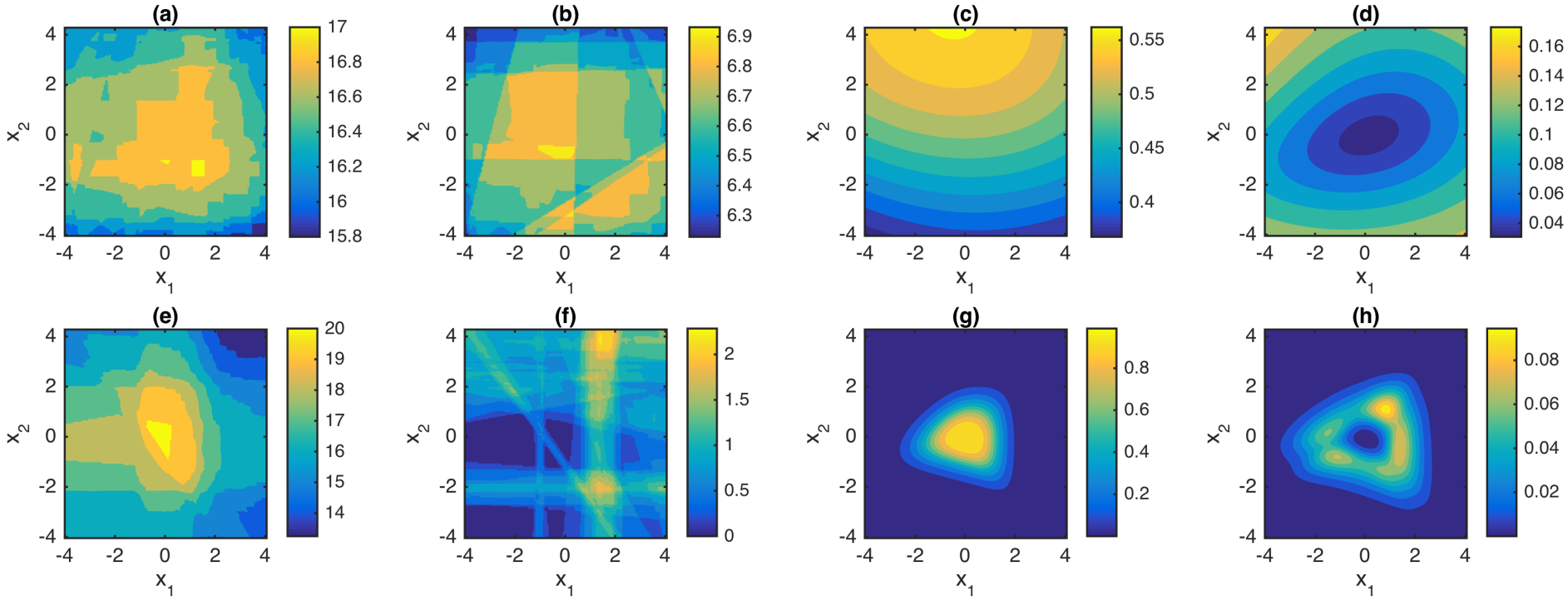

Figure: Analogous figure for stack-softplus regression to that for sum-softplus regression, with the following differences: (a) and (b) show the contour maps of the posterior means and standard deviations, respectively, of the number of inequalities specified in (3) that are satisfied. (e)-(f) are analogous plots to (a)-(b) under the opposite labeling setting.

Definition (Sum-stack-softplus regression)

Given a gamma process draw $G = \sum_{k=1}^{\infty} r_k \delta_{\boldsymbol{\beta}_k^{(2:T+1)}}$, with each $\boldsymbol{\beta}_k^{(t)} \in \mathbb{R}^{V+1}$, sum-stack-softplus (SS-softplus) regression with $T \in \{1, 2, \ldots\}$ layers parameterizes $\lambda(\boldsymbol{x}_i)$ using a SS-softplus function as

$$\lambda(\boldsymbol{x}_i) = \sum_{k=1}^{\infty} r_k \, \varsigma\left(\boldsymbol{x}_i' \boldsymbol{\beta}_k^{(2)}, \ldots, \boldsymbol{x}_i' \boldsymbol{\beta}_k^{(T+1)}\right)$$

$$= \sum_{k=1}^{\infty} r_k \ln\left(1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k^{(T+1)}} \ln\left\{1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k^{(T)}} \ln\left[1 + \ldots \ln\left(1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k^{(2)}}\right)\right]\right\}\right).$$

- SS-softplus regression is equivalent to the following noisy-OR-AND regression model

$$\begin{aligned} y_i &\sim \text{Bernoulli}\left[1 - \prod_{k=1}^{\infty}(1 - p_{ik})\right], \\ p_{ik} &= 1 - \left(1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k^{(T+1)}} \ln\left\{1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k^{(T)}} \ln\left[1 + \ldots \ln\left(1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k^{(2)}}\right)\right]\right\}\right)^{-r_k} \end{aligned}$$

- Sum-stack-softplus regression can be constructed by convolving countably infinite stacked gamma distributions that have covariate-dependent scale parameters as

$$\begin{gathered}
\theta_{ik}^{(T)} \sim \text{Gamma}\left(r_k, e^{\boldsymbol{x}_i'\boldsymbol{\beta}_k^{(T+1)}}\right), \\
\cdots \\
\theta_{ik}^{(t)} \sim \text{Gamma}\left(\theta_{ik}^{(t+1)}, e^{\boldsymbol{x}_i'\boldsymbol{\beta}_k^{(t+1)}}\right), \\
\cdots \\
\theta_{ik}^{(1)} \sim \text{Gamma}\left(\theta_{ik}^{(2)}, e^{\boldsymbol{x}_j'\boldsymbol{\beta}_k^{(2)}}\right), \\
y_i = \delta(m_i \geq 1),\ m_i = \textstyle\sum_{k=1}^{\infty} m_{ik}^{(1)},\ m_{ik}^{(1)} \sim \text{Pois}(\theta_{ik}^{(1)}).
\end{gathered}$$

Proposition

The infinite product in sum-stack-softplus regression as

$$e^{-\sum_{k=1}^{\infty} r_k \varsigma\left(\boldsymbol{x}_i'\boldsymbol{\beta}_k^{(2:T+1)}\right)} = \prod_{k=1}^{\infty}\left(1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}_k^{(T+1)}} \ln\left\{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}_k^{(T)}} \ln\left[1+\ldots \ln\left(1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}_k^{(2)}}\right)\right]\right\}\right)^{-r_k}$$

is smaller than one and has a finite expectation that is greater than zero.

Geometric constraint of sum-stack-softplus regression

Sum-stack-softplus regression:

- uses the **noisy-OR of noisy-AND** hyperplane interactions
  to define a **union of convex-polytope**-like confined space
  to enclose **positive** examples (i.e., data samples with $y_i = 1$)

Theorem

For sum-stack-softplus regression, the confined space specified by the inequality $P(y_i = 1 \mid \boldsymbol{x}_i) = 1 - e^{-\lambda(\boldsymbol{x}_i)} > p_0$, *which can be expressed as*

$$\lambda(\boldsymbol{x}_i) = \sum_{k=1}^{\infty} r_k \, \varsigma\left(\boldsymbol{x}' \boldsymbol{\beta}_k^{(2)}, \ldots, \boldsymbol{x}' \boldsymbol{\beta}_k^{(T+1)}\right) > -\ln(1 - p_0),$$

encompasses the union of convex-polytope-like confined spaces, expressed as

$$\mathcal{D}_\star = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \ldots,$$

where the k*th convex-polytope-like confined space* $\mathcal{D}_k$ *is specified by the inequality* $p_{ik} > p_0$ *that can be expressed as*

$$\boldsymbol{x}_i' \boldsymbol{\beta}_k^{(T+1)} + \ln \ln \left\{ 1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k^{(T)}} \ln \left[1 + \ldots \ln \left(1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k^{(2)}} \right) \right] \right\} > \ln \left[(1 - p_0)^{-\frac{1}{r_k}} - 1 \right]. \quad (4)$$

Corollary

For sum-stack-softplus regression, the confined space specified by the inequality $P(y_i = 1 \mid \boldsymbol{x}_i) = 1 - e^{-\lambda(\boldsymbol{x}_i)} \leq p_0$ *is bounded by* $\bar{\mathcal{D}}_\star = \bar{\mathcal{D}}_1 \cap \bar{\mathcal{D}}_2 \cap \ldots$

Proposition

For any data point $\boldsymbol{x}_i$ *that resides inside the union of countably infinite convex-polytope-like confined spaces* $\mathcal{D}_\star = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \ldots$, *which means* $\boldsymbol{x}_i$ *satisfies at least one of the inequalities in* (4), *it will be labeled under sum-stack-softplus regression with* $y_i = 1$ *with a probability greater than* p_0, *and* $y_i = 0$ *with a probability no greater than* $1 - p_0$.

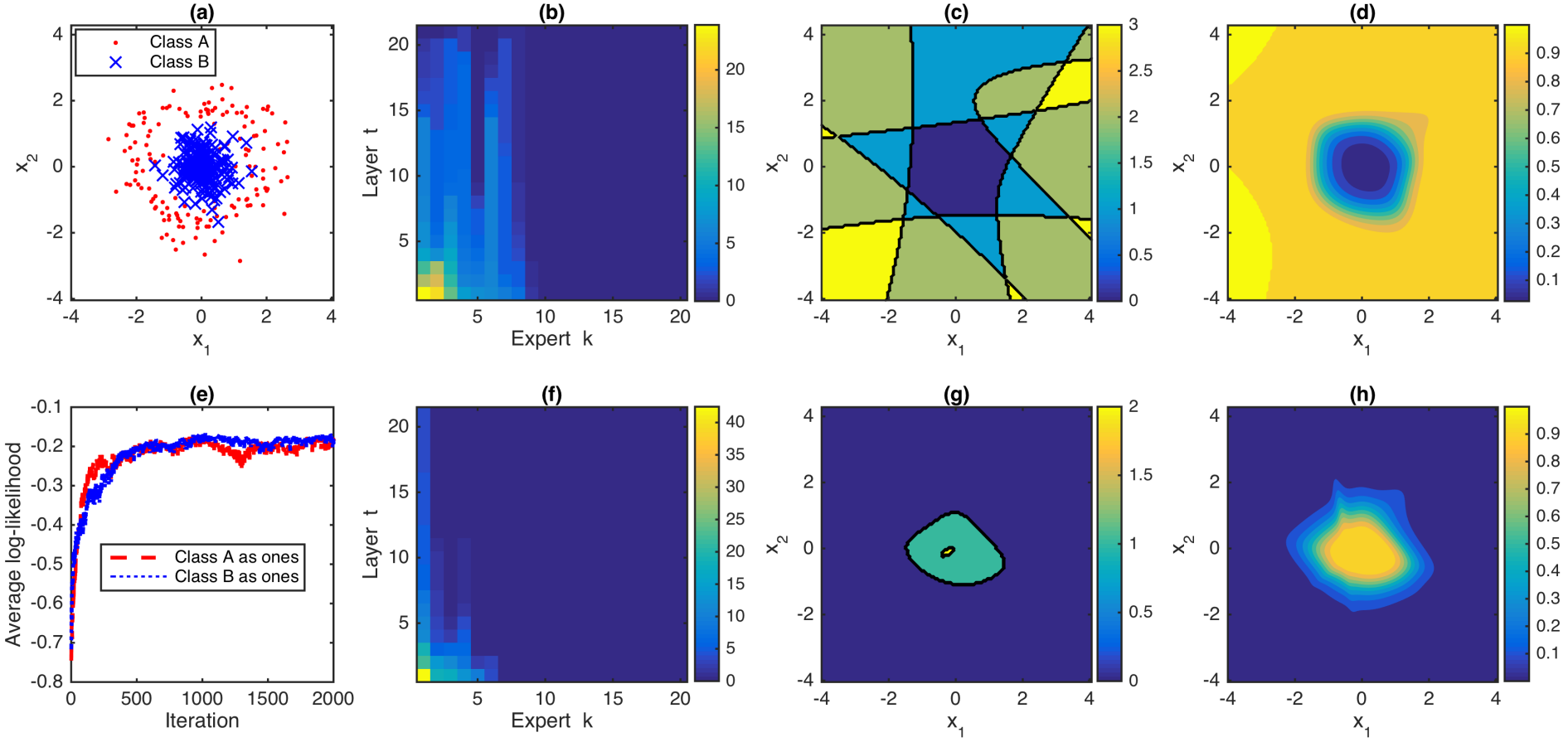

Figure: Visualization of sum-stack-softplus regression, with $K_{\max} = 20$ experts and $T = 20$ criteria for each expert, under two opposite labeling settings. (b) shows the average latent count per positive sample, $\sum_i m_{ik}^{(t)} / \sum_i \delta(y_i = 1)$, as a function of both the expert index k and layer index t, where the experts are ordered based on the values of $\sum_i m_{ik}^{(1)}$. (c) shows a contour map, whose region with nonzero values corresponds to the union of convex-polytope-like confined spaces. (d) shows the the contour map of the predicted class probabilities.

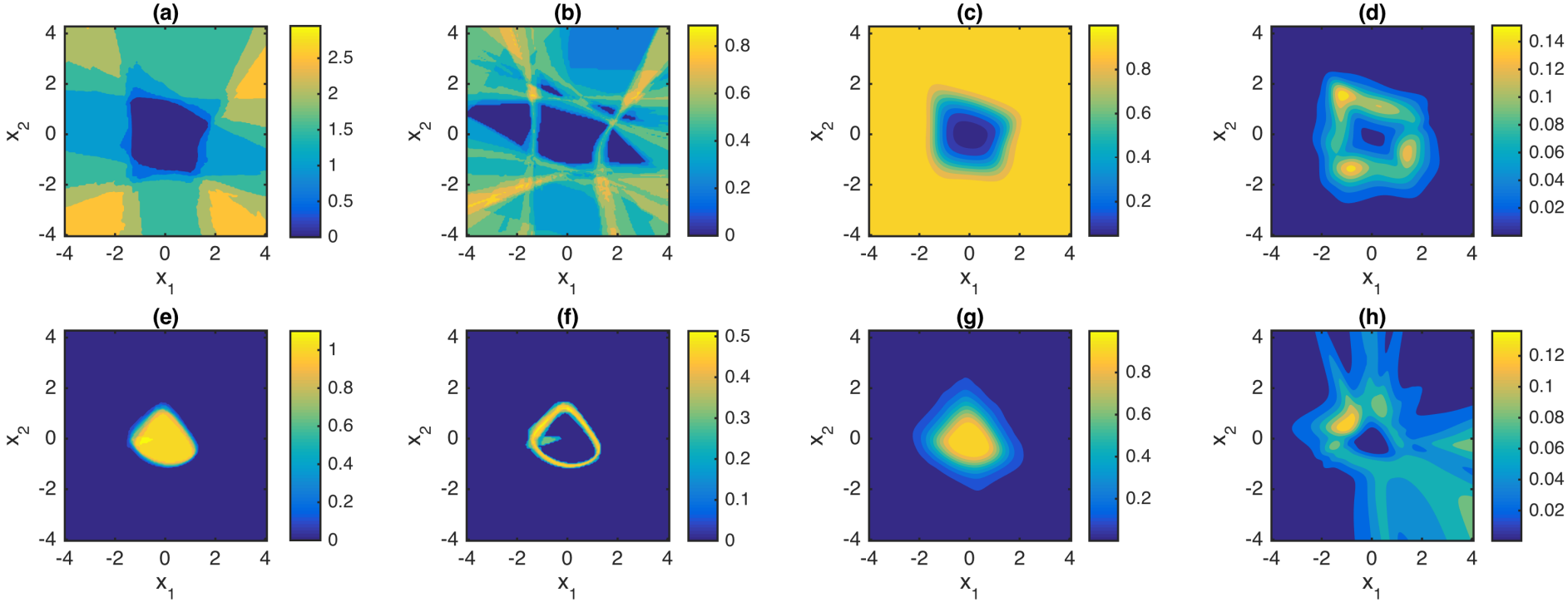

Figure: Analogous figure for sum-stack-softplus regression to that for sum-softplus regression, with the following differences: (a) and (b) show the contour maps of the posterior means and standard deviations, respectively, of the number of inequalities specified in (4) that are satisfied, and (c) and (d) show the contour maps of the posterior means and standard deviations, respectively, of predicted class probabilities. (e)-(h) are analogous plots to (a)-(d), with the data points in Classes A and B relabeled as "0" and "1," respectively.

Sum-stack-softplus regression on "banana"

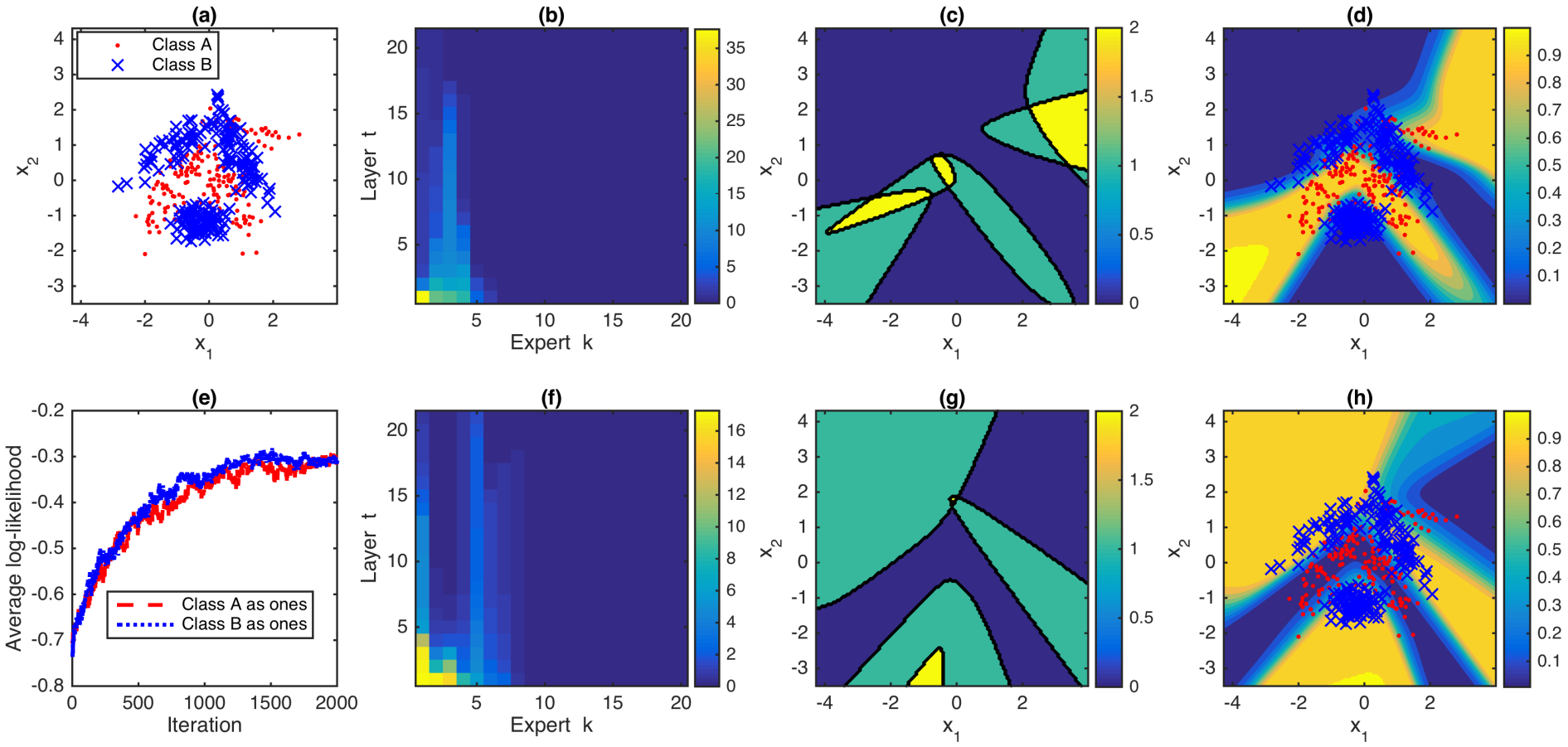

Figure: Visualization of sum-stack-softplus regression, with $K_{\max} = 20$ experts and $T = 20$ criteria for each expert, on classifying the banana dataset under two opposite labeling settings.

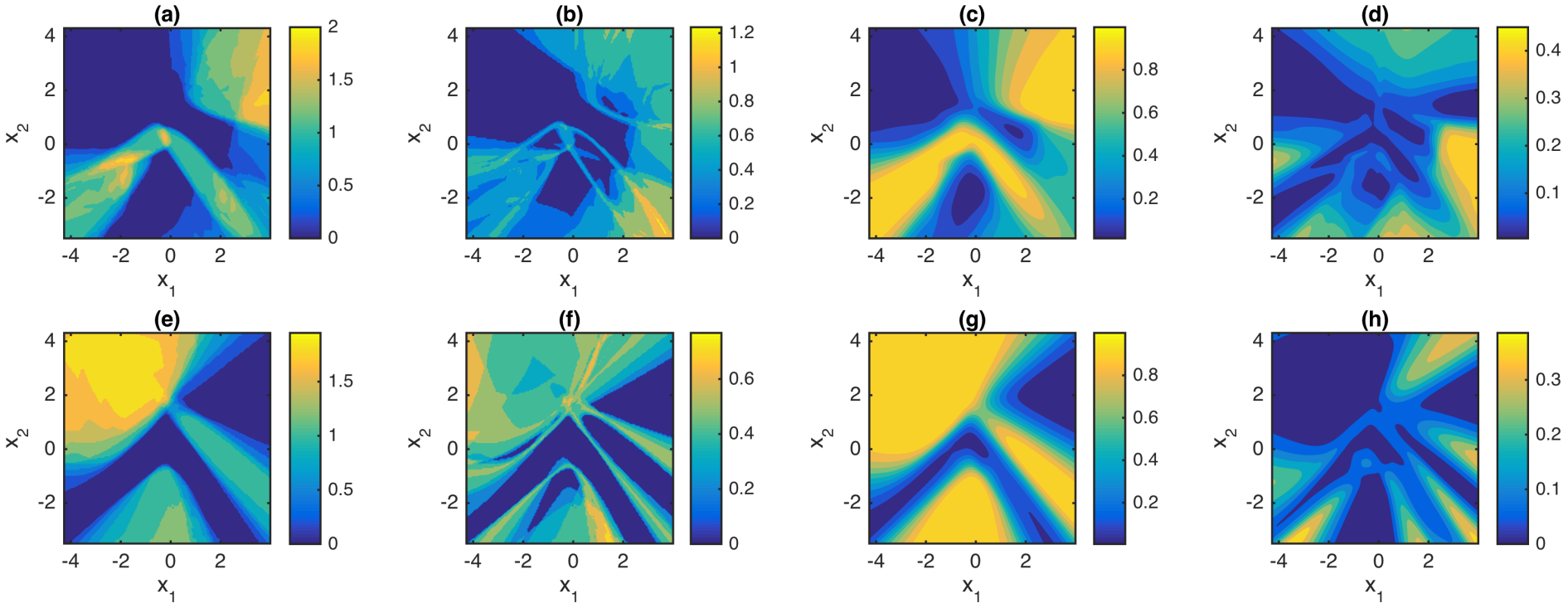

Figure: Visualization of the posteriors of sum-statck-softplus regression on the banana dataset.

Sum-stack-softplus regression on "XOR"

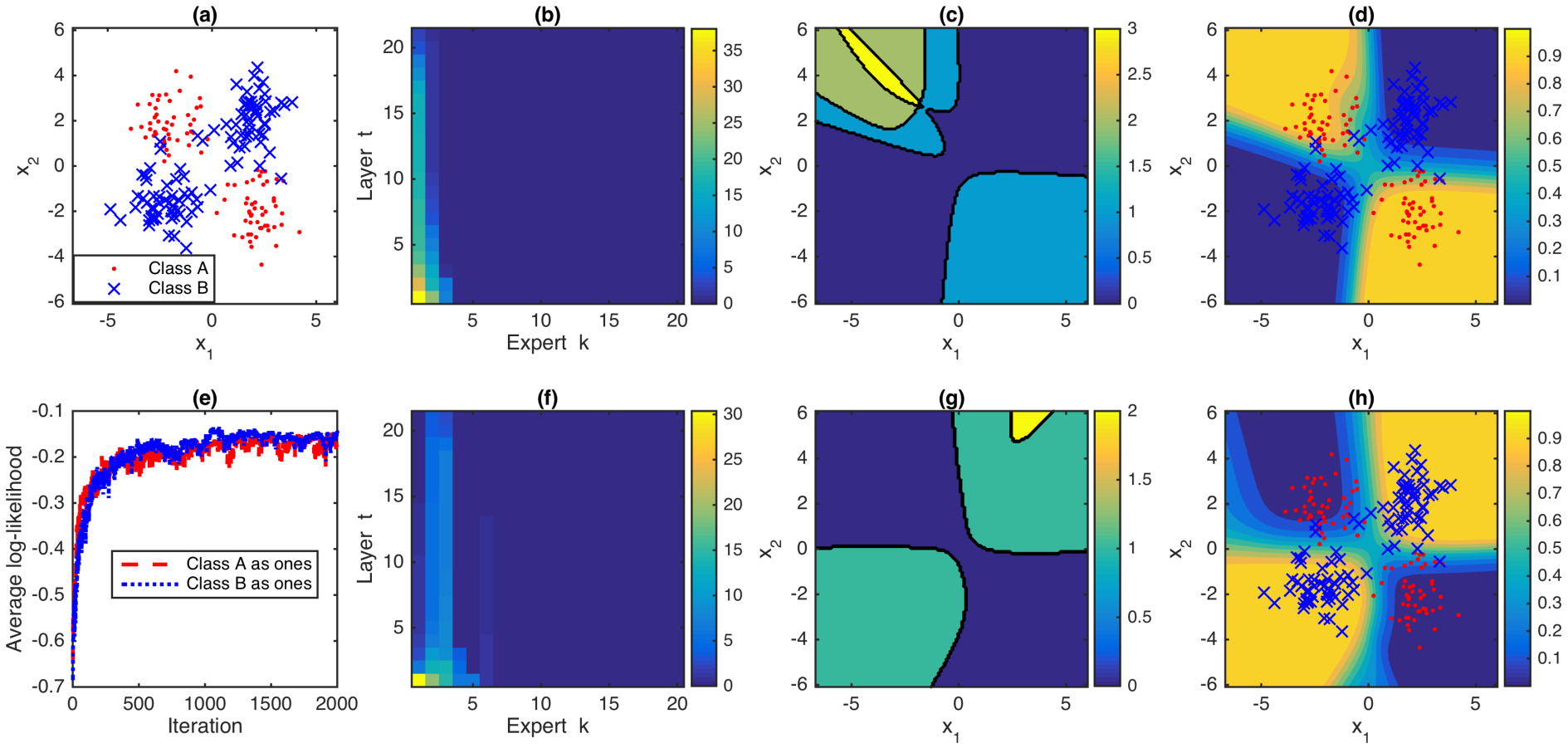

Figure: Visualization of sum-stack-softplus regression, with $K_{\max} = 20$ experts and $T = 20$ criteria for each expert, on classifying the XOR dataset under two opposite labeling settings.

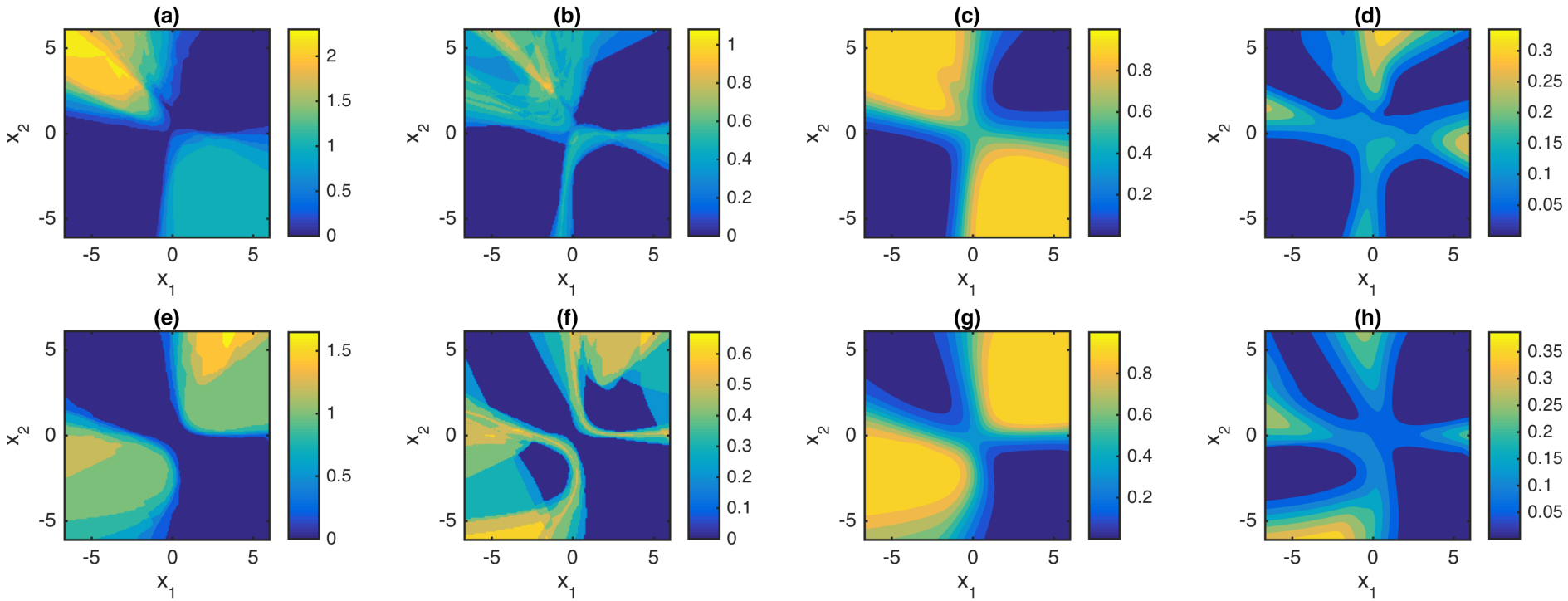

Figure: Visualization of the posteriors of sum-statck-softplus regression on the XOR dataset.

Sum-stack-softplus regression on "double moons"

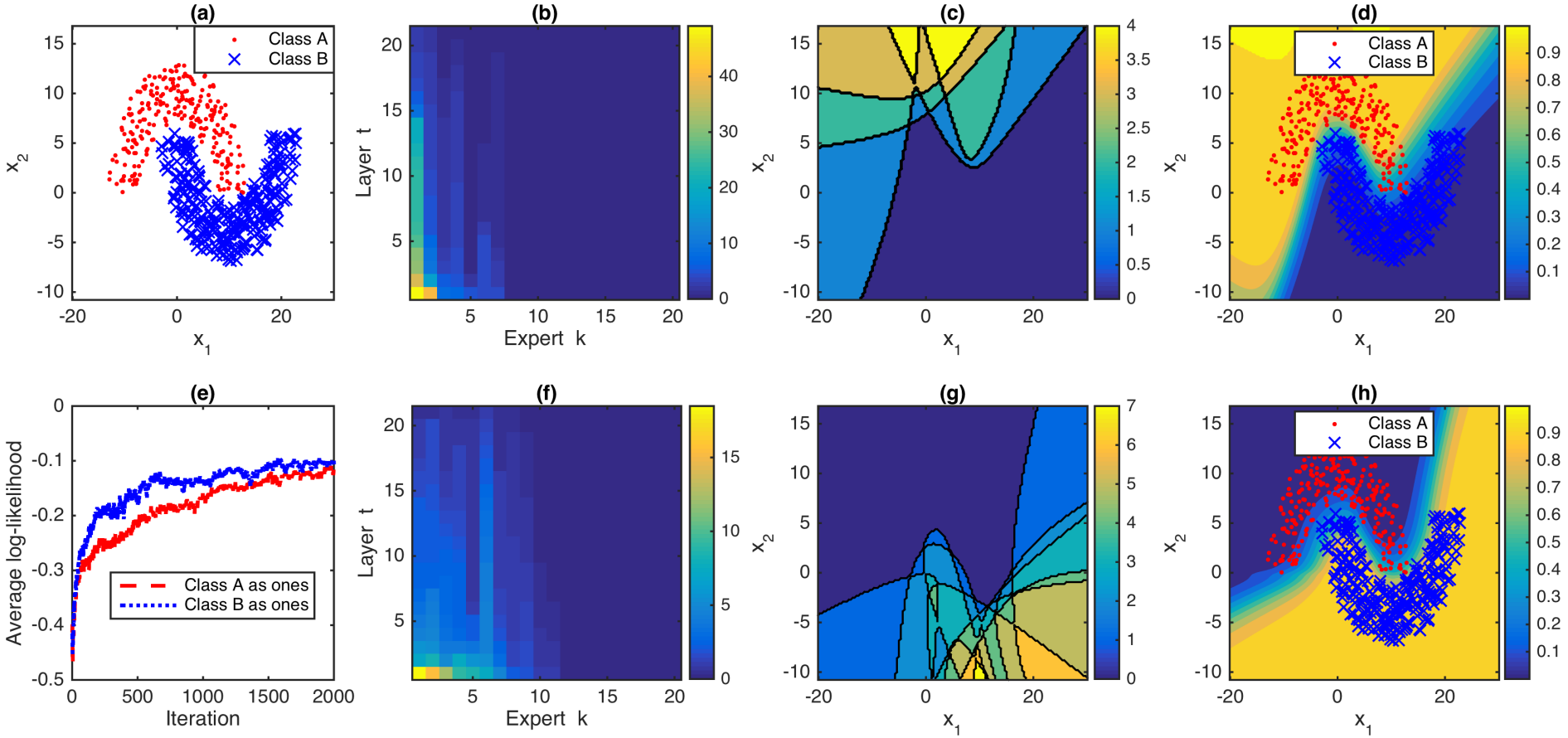

Figure: Visualization of sum-stack-softplus regression, with $K_{\max} = 20$ experts and $T = 20$ criteria for each expert, on classifying the double moons dataset under two opposite labeling settings.

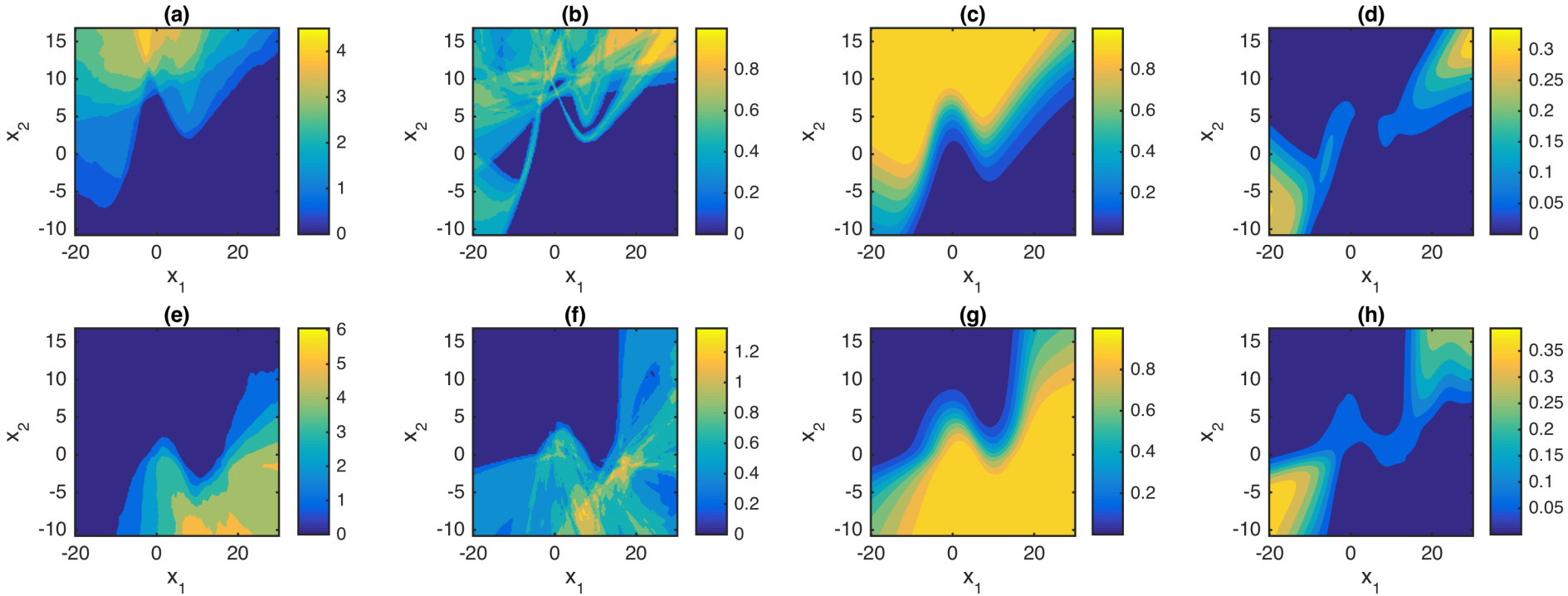

Figure: Visualization of the posteriors of sum-statck-softplus regression on the double moons dataset.

Hierarchical model

- Truncated sum-stack-softplus (SS-softplus) regression:

$$\theta_{ik}^{(T)} \sim \text{Gamma}\left(r_k, e^{\boldsymbol{x}_i'\boldsymbol{\beta}_k^{(T+1)}}\right),$$
$$\dots$$
$$\theta_{ik}^{(t)} \sim \text{Gamma}\left(\theta_{ik}^{(t+1)}, e^{\boldsymbol{x}_i'\boldsymbol{\beta}_k^{(t+1)}}\right),$$
$$\dots$$
$$\theta_{ik}^{(1)} \sim \text{Gamma}\left(\theta_{ik}^{(2)}, e^{\boldsymbol{x}_j'\boldsymbol{\beta}_k^{(2)}}\right),$$
$$y_i = \delta(m_i \geq 1),\ m_i = \textstyle\sum_{k=1}^{\infty} m_{ik}^{(1)},\ m_{ik}^{(1)} \sim \text{Pois}(\theta_{ik}^{(1)}).$$

- We complete the model by letting

$$r_k \sim \text{Gamma}(\gamma_0/K, 1/c_0),\ \gamma_0 \sim \text{Gamma}(a_0, 1/b_0),\ c_0 \sim \text{Gamma}(e_0, 1/f_0),$$
$$\boldsymbol{\beta}_k^{(t)} \sim \prod_{v=0}^{V} \mathcal{N}(0, \alpha_{vtk}^{-1}),\ \alpha_{vtk} \sim \text{Gamma}(a_t, 1/b_t),$$

where $t \in \{2, \dots, T+1\}$

Upward-downward Gibbs sampling

- Downward sampling for $t = T, T-1, \ldots, 1$
  - Sample $\theta_{ik}^{(t)}$ from Gamma
- Sample m_i from Truncated Poisson
- Sample $\{m_{ik}^{(1)}\}_k$ from Multinomial
- Upward sampling for $t = 2, 3, \ldots, T+1$
  - Sample $m_{ik}^{(t)}$ for $t > 1$ from Chinese restaurant table (CRT)
  - Sample $\omega_{ik}^{(t)}$ from Polya-Gamma
  - Sample $\beta_k^{(t)}$ from Gaussian
- Sample γ_0 and c_0 from Gamma
- Sample r_k from Gamma

Sample m_i

Denote $\theta_{i\cdot} = \sum_{k=1}^{K} \theta_{ik}^{(1)}$. Since $m_i = 0$ a.s. given $y_i = 0$ and $m_i \geq 1$ given $y_i = 1$, and in the prior we have $m_i \sim \text{Pois}(\theta_{i\cdot})$, following the inference for the Bernoulli-Poisson link in [Zhou, 2015], we can sample m_i from the truncated Poisson distribution as

$$(m_i \,|\, -) \sim y_i \text{Pois}_+ (\theta_{i\cdot})$$

Sample $m_{ik}^{(1)}$

Since letting

$$m_i = \sum_{k=1}^{K} m_{ik}^{(1)}, \;\; m_{ik}^{(1)} \sim \text{Pois}(\theta_{ik}^{(1)})$$

is equivalent in distribution to letting

$$(m_{i1}^{(1)}, \ldots, m_{iK}^{(1)}) \,|\, m_i \sim \text{Mult}\left(m_i, \theta_{i1}^{(1)}/\theta_{i\cdot}, \ldots, \theta_{iK}^{(1)}/\theta_{i\cdot}\right), \;\; m_i \sim \text{Pois}\left(\theta_{i\cdot}\right),$$

similar to [Zhou et al., 2012], we sample $m_{ik}^{(1)}$ as

$$(m_{i1}^{(1)}, \ldots, m_{iK}^{(1)} \,|\, -) \sim \text{Mult}\left(m_i, \theta_{i1}^{(1)}/\theta_{i\cdot}, \ldots, \theta_{iK}^{(1)}/\theta_{i\cdot}\right)$$

Theorem

Let us denote $p_{ik}^{(t)} = 1 - e^{-q_{ik}^{(t)}}$, *i.e.*, $q_{ik}^{(t)} = -\ln(1 - p_{ik}^{(t)})$, *and* $\theta_{ik}^{(T+1)} = r_k$. *With* $q_{ik}^{(1)} := 1$ *and*

$$q_{ik}^{(t+1)} := \ln\left(1 + q_{ik}^{(t)} e^{\boldsymbol{x}_i \beta_k^{(t+1)}}\right)$$

for $t = 1, \ldots, T$, *which means*

$$\begin{aligned} q_{ik}^{(t+1)} &= \varsigma\left(\boldsymbol{x}_i'\beta_k^{(2)}, \ldots, \boldsymbol{x}_i'\beta_k^{(t+1)}\right) \\ &= \ln\left(1+e^{\boldsymbol{x}_i'\beta_k^{(t+1)}} \ln\left\{1+e^{\boldsymbol{x}_i'\beta_k^{(t)}} \ln\left[1+\ldots \ln\left(1+e^{\boldsymbol{x}_i'\beta_k^{(2)}}\right)\right]\right\}\right), \end{aligned}$$

one may find latent counts $m_{ik}^{(t)}$ *that are connected to the regression coefficient vectors under negative binomial regression as*

$$m_{ik}^{(t)} \sim NB(\theta_{ik}^{(t+1)}, \ 1 - e^{-q_{ik}^{(t+1)}}) = NB\left(\theta_{ik}^{(t+1)}, \frac{1}{1 + e^{-\boldsymbol{x}_i'\beta_k^{(t+1)} - \ln(q_{ik}^{(t)})}}\right). \tag{5}$$

Proof.

By construction (5) is true for $t = 1$. Suppose (5) is also true for $t \geq 2$, then we can augment each $m_{ik}^{(t)}$ under its compound Poisson representation as

$$m_{ik}^{(t)} \mid m_{ik}^{(t+1)} \sim \text{SumLog}(m_{ik}^{(t+1)},\ p_{ik}^{(t+1)}), \quad m_{ik}^{(t+1)} \sim \text{Pois}\left(\theta_{ik}^{(t+1)} q_{ik}^{(t+1)}\right), \tag{6}$$

where the joint distribution of $m_{ik}^{(t)}$ and $m_{ik}^{(t+1)}$, according to Theorem 1 of Zhou and Carin (2015), is the same as that in

$$m_{ik}^{(t+1)} \mid m_{ik}^{(t)} \sim \text{CRT}(m_{ik}^{(t)},\ \theta_{ik}^{(t+1)}), \quad m_{ik}^{(t)} \sim \text{NB}(\theta_{ik}^{(t+1)},\ p_{ik}^{(t+1)}),$$

where CRT refers to the Chinese restaurant table distribution described in Zhou and Carin (2015). Marginalizing $\theta_{ik}^{(t+1)}$ from the Poisson distribution in (6) leads to $m_{ik}^{(t+1)} \sim \text{NB}(\theta_{ik}^{(t+2)}, p_{ik}^{(t+2)})$. Thus if (5) is true for layer t, then it is also true for layer $t + 1$. □

Experiments on benchmark datasets

Table: Binary classification datasets used in experiments, where V is the feature dimension.

Dataset	banana	breast cancer	titanic	waveform	german	image	ijcnn1	a9a
Train size	400	200	150	400	700	1300	49,990	32,561
Test size	4900	77	2051	4600	300	1010	91,701	16,281
V	2	9	3	21	20	18	22	123

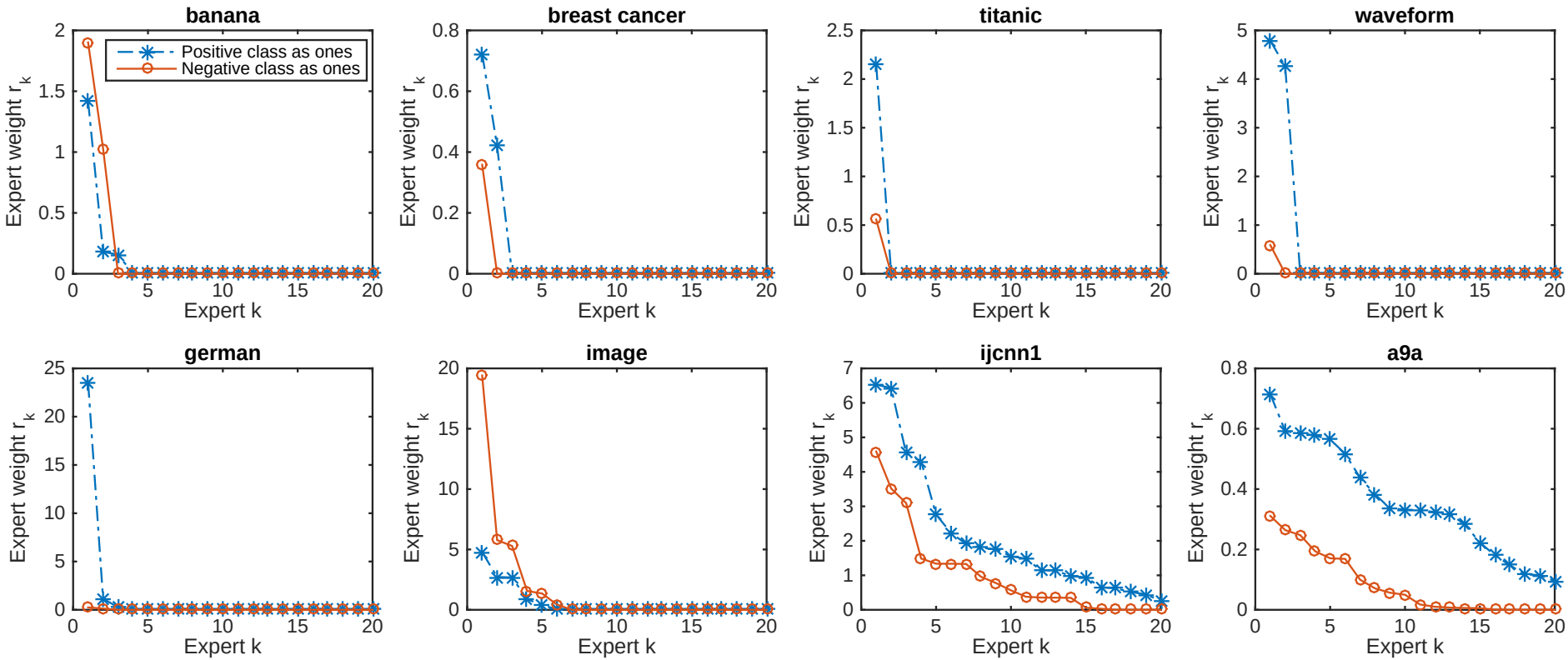

Figure: The inferred weights of the $K_{\max} = 20$ experts of sum-softplus regression, ordered from left to right according to their weights, on eight different datasets, based on the maximum likelihood sample of a single random trial.

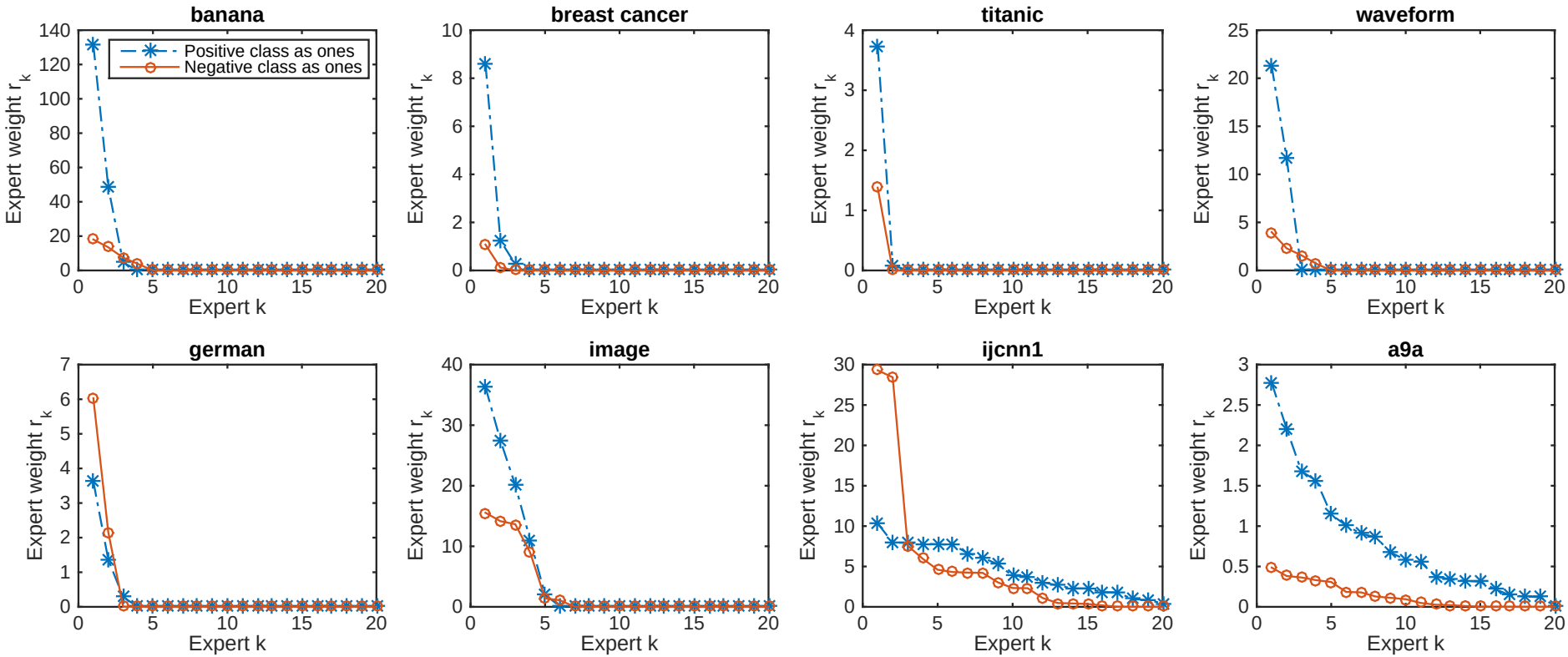

Figure: Analogous plots for SS-softplus regression with $K_{\max} = 20$ and $T = 5$.

Table: Comparison of classification errors of logistic regression (LR), RBF kernel support vector machine (SVM), relevance vector machine (RVM), adaptive multi-hyperplane machine (AMM), convex polytope machine (CPM), softplus regression, sum-softplus (sum-ς) regression with $K_{\max} = 20$, stack-softplus (stack-ς) regression with $T = 5$, SS-softplus (SS-ς) regression with $K_{\max} = 20$ and $T = 3$, and SS-ς regression with $K_{\max} = 20$ and $T = 5$. Displayed in each column of the last row is the average of the classification errors of an algorithm normalized by those of kernel SVM.

Dataset	LR	SVM	RVM	AMM	CPM	softplus	sum-ς	stack-ς (T=5)	SS-ς (T=3)	SS-ς (T=5)
banana	47.76 ±4.38	**10.85** ±0.57	11.08 ±0.69	18.76 ±4.09	21.39 ±1.72	47.87 ±4.36	30.78 ±8.68	33.21 ±5.76	12.54 ±1.18	11.89 ±0.61
breast cancer	28.05 ±3.68	28.44 ±4.52	31.56 ±4.66	31.82 ±4.47	32.08 ±4.29	28.70 ±4.76	30.13 ±4.23	**27.92** ±3.31	30.39 ±4.94	28.83 ±3.40
titanic	22.67 ±0.98	22.33 ±0.63	23.20 ±1.08	28.85 ±8.56	22.37 ±0.45	22.53 ±0.43	22.48 ±0.25	22.71 ±0.70	22.42 ±0.45	**22.29** ±0.80
waveform	13.33 ±0.59	**10.73** ±0.86	11.16 ±0.72	11.81 ±1.13	12.76 ±1.17	13.62 ±0.71	11.51 ±0.65	12.25 ±0.69	11.34 ±0.70	11.69 ±0.69
german	23.63 ±1.70	23.30 ±2.51	23.67 ±2.28	25.13 ±3.73	25.03 ±2.49	24.07 ±2.11	23.60 ±2.39	**22.97** ±2.22	23.30 ±2.20	24.23 ±2.46
image	17.53 ±1.05	2.84 ±0.52	3.82 ±0.59	3.82 ±0.87	3.25 ±0.41	17.55 ±0.75	3.50 ±0.73	7.97 ±0.52	**2.59** ±0.47	2.73 ±0.53
Mean of SVM normalized errors	2.472	1	1.095	1.277	1.251	2.485	1.370	1.665	1.033	1.033

Table: Comparison of the number of experts (times the number of hyperplanes per expert), where an expert contains T hyperplanes for both stack- and SS-softplus regressions and contains a single hyperplane/support vector for all the others. The computational complexity for out-of-sample prediction is about linear in the number of hyperplanes/support vectors. Displayed in each column of the last row is the average of the number of experts (times the number of hyperplanes per expert) of an algorithm normalized by those of RBF kernel SVM.

Dataset	LR	SVM	RVM	AMM	CPM	softplus	sum-ς	stack-ς (T=5)	SS-ς (T=3)	SS-ς (T=5)
banana	1	129.20 ±32.76	22.30 ±26.02	9.50 ±2.80	14.60 ±7.49	2	3.70 ±0.95	2 (×5)	6.80 (×3) ±0.79 (×3)	7.60 (×5) ±1.17 (×5)
breast cancer	1	115.10 ±11.16	24.80 ±28.32	13.40 ±0.84	12.00 ±8.43	2	3.10 ±0.74	2 (×5)	5.70 (×3) ±1.70 (×3)	6.40 (×5) ±1.43 (×5)
titanic	1	83.40 ±13.28	5.10 ±3.03	14.90 ±3.14	5.20 ±2.53	2	2.30 ±0.48	2 (×5)	3.80 (×3) ±0.92 (×3)	4.00 (×5) ±0.94 (×5)
waveform	1	147.00 ±38.49	21.10 ±10.98	9.50 ±1.18	6.40 ±2.27	2	4.40 ±0.84	2 (×5)	7.00 (×3) ±2.21 (×3)	8.90 (×5) ±2.33 (×5)
german	1	423.60 ±55.02	11.00 ±3.20	18.80 ±1.81	8.80 ±7.79	2	6.70 ±0.95	2 (×5)	11.10 (×3) ±2.64 (×3)	14.70 (×5) ±1.77 (×5)
image	1	211.60 ±47.51	35.80 ±9.19	10.50 ±1.08	23.00 ±6.75	2	11.20 ±1.32	2 (×5)	14.60 (×3) ±2.07 (×3)	17.60 (×5) ±1.90 (×5)
Mean of SVM normalized K	0.007	1	0.131	0.088	0.075	0.014	0.030	0.014 (×5)	0.048 (×3)	0.057 (×5)

Table: Performance of stack-softplus regression with the depth set as $T \in \{1, 2, 3, 5, 10\}$, where stack-softplus regression with $T = 1$ reduces to softplus regression.

Dataset	softplus	stack-ς (T=2)	stack-ς (T=3)	stack-ς (T=5)	stack-ς (T=10)
banana	47.87 ± 4.36	34.66 ± 5.58	32.19 ± 4.76	33.21 ± 5.76	$\mathbf{30.67} \pm 4.23$
breast cancer	28.70 ± 4.76	29.35 ± 2.31	29.48 ± 4.94	$\mathbf{27.92} \pm 3.31$	28.31 ± 4.36
titanic	22.53 ± 0.43	22.80 ± 0.59	22.48 ± 0.55	22.71 ± 0.70	22.84 ± 0.54
waveform	13.62 ± 0.71	12.52 ± 1.14	$\mathbf{12.23} \pm 0.79$	12.25 ± 0.69	12.33 ± 0.65
german	24.07 ± 2.11	23.73 ± 1.99	23.67 ± 1.89	$\mathbf{22.97} \pm 2.22$	23.80 ± 1.64
image	17.55 ± 0.75	9.11 ± 0.99	8.39 ± 1.05	7.97 ± 0.52	$\mathbf{7.50} \pm 1.17$
Mean of SVM normalized errors	2.485	1.773	1.686	1.665	1.609

Table: Performance of SS-softplus regression with $K_{\max} = 20$ and the depth set as $T \in \{1, 2, 3, 5, 10\}$, where SS-softplus regression with $T = 1$ reduces to sum-softplus regression.

Dataset	sum-ς	SS-ς (T=2)	SS-ς (T=3)	SS-ς (T=5)	SS-ς (T=10)
banana	30.78 ± 8.68	15.00 ± 5.31	12.54 ± 1.18	$\mathbf{11.89} \pm 0.61$	11.93 ± 0.59
breast cancer	30.13 ± 4.23	29.74 ± 3.89	30.39 ± 4.94	28.83 ± 3.40	$\mathbf{28.44} \pm 4.60$
titanic	22.48 ± 0.25	22.56 ± 0.65	22.42 ± 0.45	22.29 ± 0.80	$\mathbf{22.20} \pm 0.48$
waveform	11.51 ± 0.65	11.41 ± 0.96	$\mathbf{11.34} \pm 0.70$	11.69 ± 0.69	12.92 ± 1.00
german	23.60 ± 2.39	$\mathbf{23.30} \pm 2.54$	$\mathbf{23.30} \pm 2.20$	24.23 ± 2.46	23.90 ± 1.50
image	3.50 ± 0.73	2.76 ± 0.47	$\mathbf{2.59} \pm 0.47$	2.73 ± 0.53	2.93 ± 0.46
Mean of SVM normalized errors	1.370	1.079	1.033	1.033	1.059

Table: Comparison of the number of experts (times the number of hyperplanes per expert) for SS-softplus regression.

Dataset	sum-ς	SS-ς (T=2)	SS-ς (T=3)	SS-ς (T=5)	SS-ς (T=10)
banana	3.70 ± 0.95	5.70 ± 0.67	6.80 ± 0.79	7.60 ± 1.17	9.80 ± 2.39
breast cancer	3.10 ± 0.74	4.10 ± 0.88	5.70 ± 1.70	6.40 ± 1.43	9.50 ± 1.51
titanic	2.30 ± 0.48	3.30 ± 0.82	3.80 ± 0.92	4.00 ± 0.94	6.20 ± 1.23
waveform	4.40 ± 0.84	6.20 ± 1.62	7.00 ± 2.21	8.90 ± 2.33	11.50 ± 2.72
german	6.70 ± 0.95	9.80 ± 1.48	11.10 ± 2.64	14.70 ± 1.77	20.00 ± 2.40
image	11.20 ± 1.32	13.20 ± 2.30	14.60 ± 2.07	17.60 ± 1.90	21.40 ± 2.22
Mean of SVM normalized K	0.030	0.041 ($\times$2)	0.048 ($\times$3)	0.057 ($\times$5)	0.077 ($\times$10)

Table: Comparison of classification errors of logistic regression (LR), support vector machine (SVM), adaptive multi-hyperplane machine (AMM), convex polytope machine (CPM), softplus regression, sum-softplus (sum-ς) regression with $K_{\max} = 20$, stack-softplus (stack-ς) regression with $T = 5$, and SS-softplus regression with $K_{\max} = 20$ and $T = 5$.

Dataset	LR	SVM	RVM	AMM	CPM	softplus	sum-ς	stack-ς (T=5)	SS-ς (T=5)
ijcnn1	8.00	1.30	**1.29**	2.06 ±0.27	2.57 ±0.17	8.41 ±0.03	3.39 ±0.17	6.43 ±0.15	2.24 ±0.12
a9a	15.00	**14.88**	14.95	15.03 ±0.17	15.08 ±0.07	15.02 ±0.06	**14.88** ±0.05	15.00 ±0.06	15.02 ±0.11

Table: Comparison of the number of inferred experts (hyperplanes).

	LR	SVM	RVM	AMM	CPM	softplus	sum-ς	stack-ς (T=5)	SS-ς (T=5)
ijcnn1	1	2477	296	8.20 ±0.84	58.00 ±13.04	2	37.60 ±1.52	2 (×5)	38.80 (×5) ±0.84 (×5)
a9a	1	11506	109	28.00 ±4.12	7.60 ±2.19	2	37.60 ±0.55	2 (×5)	40.00 (×5) ±0.00 (×5)

Table: Performance of stack-softplus regression with the depth set as $T \in \{1, 2, 3, 5, 10\}$, where stack-softplus regression with $T = 1$ reduces to softplus regression.

Dataset	softplus	stack-ς (T=2)	stack-ς (T=3)	stack-ς (T=5)	stack-ς (T=10)
ijcnn1	8.41 ± 0.03	6.73 ± 0.13	6.44 ± 0.21	6.43 ± 0.15	$\mathbf{6.39} \pm 0.08$
a9a	15.02 ± 0.06	14.96 ± 0.04	$\mathbf{14.93} \pm 0.06$	15.00 ± 0.06	14.97 ± 0.08

Table: Performance of SS-softplus regression with $K_{\max} = 20$ and the depth set as $T \in \{1, 2, 3, 5, 10\}$, where SS-softplus regression with $T = 1$ reduces to sum-softplus regression.

Dataset	sum-ς	SS-ς (T=2)	SS-ς (T=3)	SS-ς (T=5)	SS-ς (T=10)
ijcnn1	3.39 ± 0.17	2.32 ± 0.18	2.31 ± 0.17	2.24 ± 0.12	$\mathbf{2.19} \pm 0.11$
a9a	$\mathbf{14.88} \pm 0.05$	14.98 ± 0.03	15.07 ± 0.20	15.02 ± 0.11	15.09 ± 0.06

Conclusions

- We propose sum-, stack-, and sum-stack-softplus regressions that combine multiple hyperplanes, respectively,
  - via the noisy-OR interaction to construct a convex-polytope-bounded confined space to enclose the negative class,
  - via the noisy-AND interaction to construct a convex-polytope-bounded confined space to enclose the negative class,
  - and via the noisy-OR-AND interaction to construct a union of convex-polytope-like confined spaces to enclose the positive class.
- Sum-stack-softplus regression, including logistic regression and all the other softplus regressions as special examples, constructs a highly flexible nonparametric Bayesian predictive distribution by mixing the convolved and stacked covariate-dependent gamma distributions with the Bernoulli-Poisson distribution.

- The predictive distribution is deconvolved and demixed by inferring the parameters of the underlying nonparametric Bayesian hierarchical model using a series of data augmentation and marginalization techniques.
- In the proposed Gibbs sampler that has closed-form update equations, the parameters of different stacked gamma distributions can be updated in parallel within each iteration.
- Example results demonstrate that the proposed softplus regressions
  - can achieve classification accuracies comparable to those of kernel support vector machine,
  - but consume significant less computation for out-of-sample predictions,
  - provide probability estimates, quantify uncertainties,
  - and place interpretable geometric constraints on its classification decision boundaries directly in the original covariate space.