Logistic Regression

Lecturer:
Ondrej Drbohlav

Authors:
Ondřej Drbohlav, Jiří Matas

Centre for Machine Perception
Czech Technical University, Prague
http://cmp.felk.cvut.cz

Oct 2022, Ver. 1.1.

Logistic Regression

Outline of the talk

- Motivation
- Model, relationship between log odds and posteriors
- Cross entropy objective function
- Gradient descent for fitting the model
- Examples
- Conclusion

Logistic Regression, Motivation

Consider a classification problem with 0/1 loss matrix. Recall that given an observation $\mathbf{x}$, the optimal Bayesian strategy $q(\mathbf{x})$ decides for a class k which maximizes the posterior:

$$q(\mathbf{x}) = \operatorname*{argmax}_k p(k|\mathbf{x}) = \operatorname*{argmax}_k p(\mathbf{x}, k) = \operatorname*{argmax}_k p(\mathbf{x}|k)p(k)\,. \tag{1}$$

For a binary (2-class) classification,

$$q(\mathbf{x}) = 1 \quad \text{if} \quad p(1|\mathbf{x}) > p(2|\mathbf{x}) \quad \Leftrightarrow \quad \frac{p(1|\mathbf{x})}{p(2|\mathbf{x})} > 1 \quad \Leftrightarrow \quad \ln \frac{p(1|\mathbf{x})}{p(2|\mathbf{x})} > 0\,, \tag{2}$$

$$q(\mathbf{x}) = 2 \quad \text{if} \quad p(1|\mathbf{x}) < p(2|\mathbf{x}) \quad \Leftrightarrow \quad \frac{p(1|\mathbf{x})}{p(2|\mathbf{x})} < 1 \quad \Leftrightarrow \quad \ln \frac{p(1|\mathbf{x})}{p(2|\mathbf{x})} < 0\,. \tag{3}$$

The ratio of posteriors $\dfrac{p(1|\mathbf{x})}{p(2|\mathbf{x})}$ is called **odds ratio**, the logarithm of this ratio, $\ln \dfrac{p(1|\mathbf{x})}{p(2|\mathbf{x})}$, is called **log odds**.

Logistic Regression, Motivation

Why are we interested in log odds? Because for the following problems, the log odds are linear (therefore simple) function of the observation variable $\mathbf{x}$:

- Normal distributions with equal variances
- Independent features with binary outcomes
- Multinomial naive Bayes

Normal distributions with equal covariance matrices

$$p(\mathbf{x}|1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)}, \tag{4}$$

$$p(\mathbf{x}|2) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)}, \tag{5}$$

$$(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^D, \boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}, \boldsymbol{\Sigma} \succ 0, \boldsymbol{\Sigma} = \boldsymbol{\Sigma}^\top) \tag{6}$$

The log odds:

$$\ln \frac{p(1|\mathbf{x})}{p(2|\mathbf{x})} = \ln \frac{p(\mathbf{x}|1)p(1)}{p(\mathbf{x}|2)p(2)} = \ln \frac{p(\mathbf{x}|1)}{p(\mathbf{x}|2)} + \ln \frac{p(1)}{p(2)} = \tag{7}$$

$$= \ln \frac{\cancel{(2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}}}}{\cancel{(2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}}}} e^{-\frac{1}{2}\{(\mathbf{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - (\mathbf{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\}} + \ln \frac{p(1)}{p(2)} \tag{8}$$

Logistic Regression, Motivation

$$\ln \frac{p(1|\mathbf{x})}{p(2|\mathbf{x})} = \ln \frac{\cancel{(2\pi)^{-\frac{D}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}}{\cancel{(2\pi)^{-\frac{D}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}} e^{-\frac{1}{2}\{(\mathbf{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)-(\mathbf{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\}} + \ln \frac{p(1)}{p(2)} \tag{9}$$

$$= -\frac{1}{2}(\cancel{\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}} - \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1) \tag{10}$$

$$+ \frac{1}{2}(\cancel{\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}} - \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2) + \ln \frac{p(1)}{p(2)} \tag{11}$$

(Note: $\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{y} = \mathbf{y}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}$ because $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^\top$)

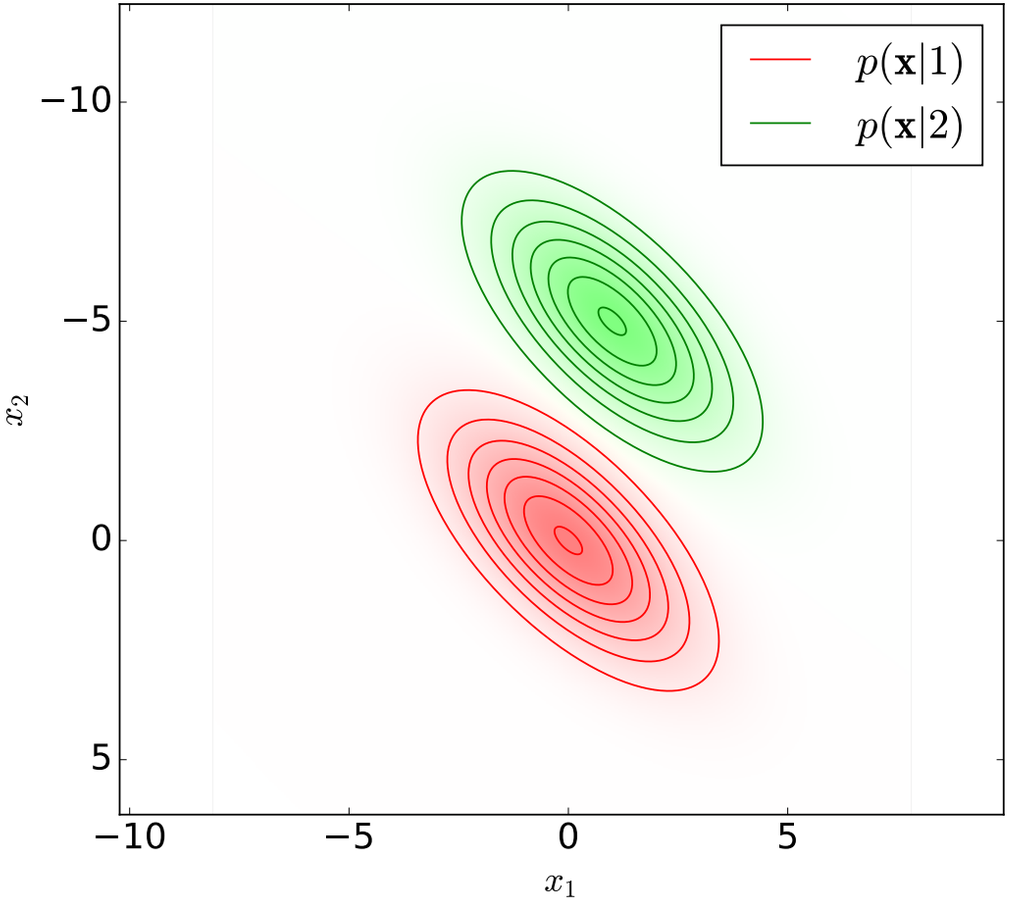
$$= -\frac{1}{2}(-2\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - (-2\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2) + \ln \frac{p(1)}{p(2)} \tag{12}$$

$$= \underbrace{[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}]}_{\mathbf{w}}\mathbf{x} + \underbrace{\frac{1}{2}(\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2) + \ln \frac{p(1)}{p(2)}}_{w_0} \tag{13}$$

$$= \mathbf{w} \cdot \mathbf{x} + w_0\,, \qquad (\mathbf{w} \in \mathbb{R}^D, w_0 \in R) \tag{14}$$

Conclusion: When two classes are normally distributed with equal covariance matrices, the log odds are a **linear** function of observation vector $\mathbf{x}$.

Example



Multinomial normal distribution, two classes

Centers:

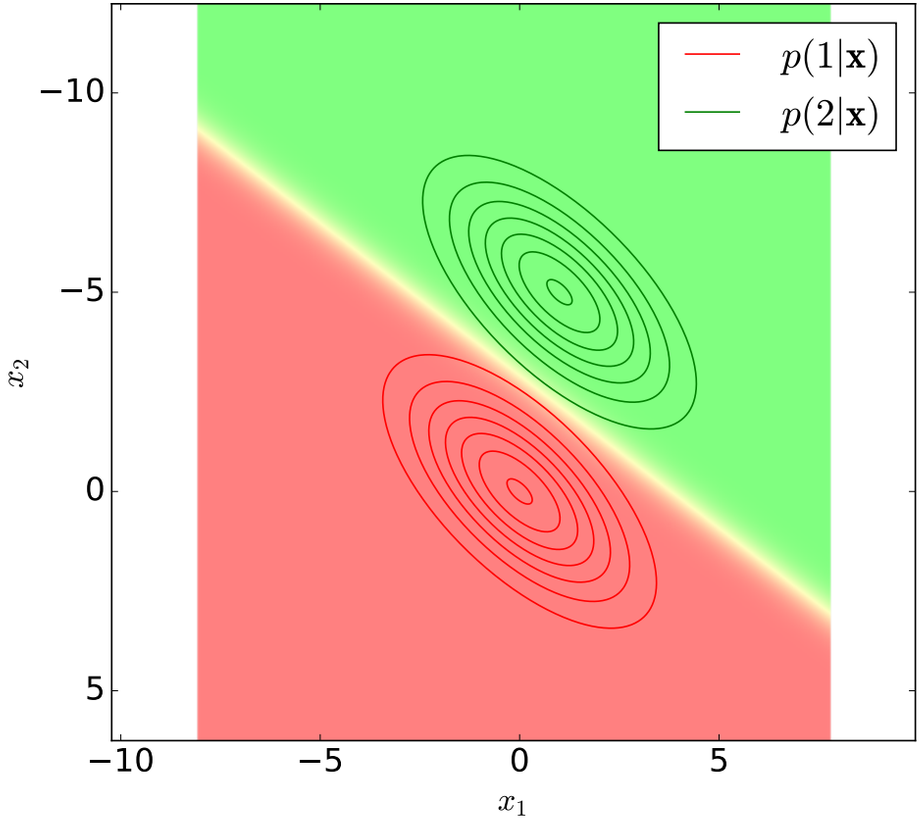
$(0, 0)$, $(1, -5)$.

Cov. matrices (are equal):

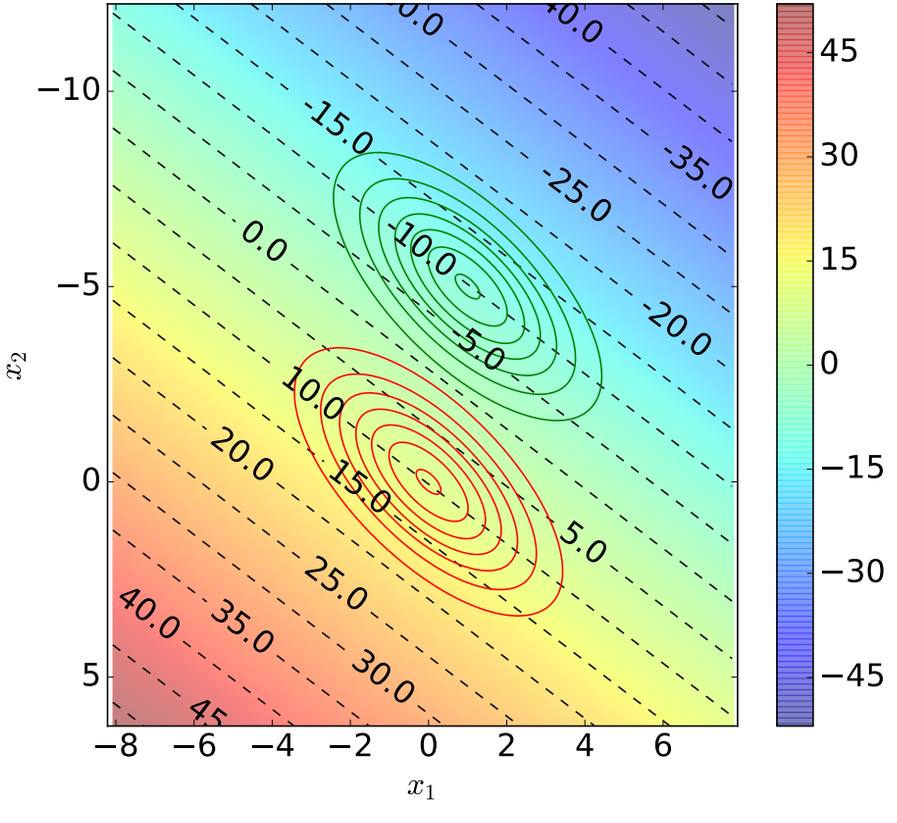
$$\Sigma = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}$$

Example (contd.)

Priors: $p(\text{red}) = 0.5$, $p(\text{green}) = 0.5$



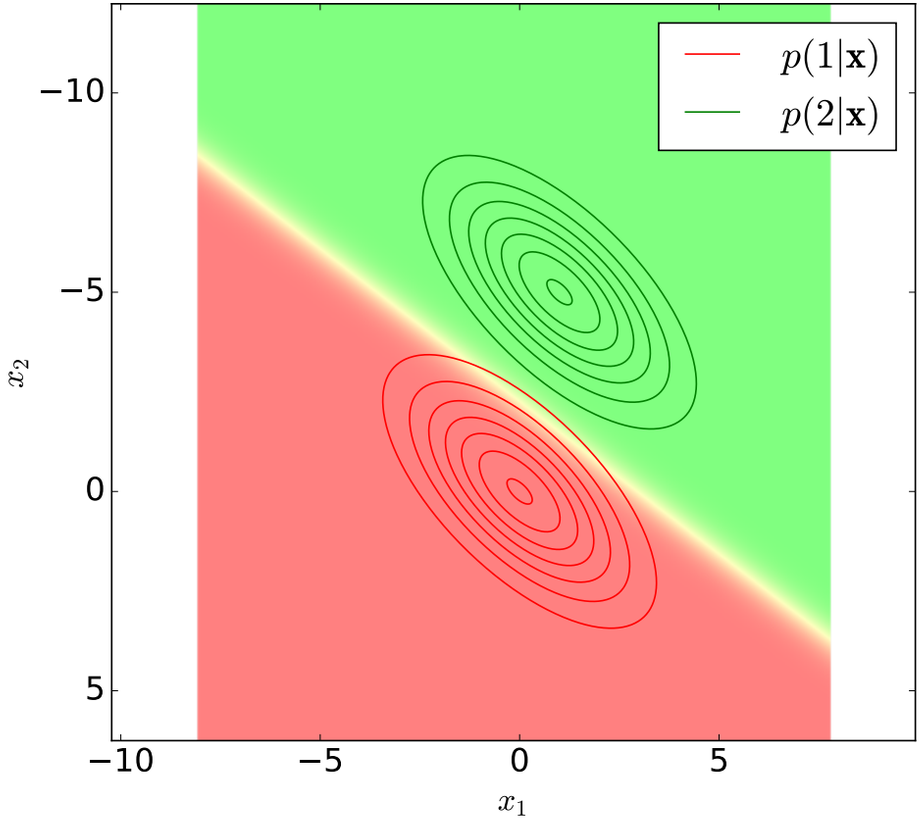
Probabilities $p(1|\mathbf{x})$ and $p(2|\mathbf{x})$.



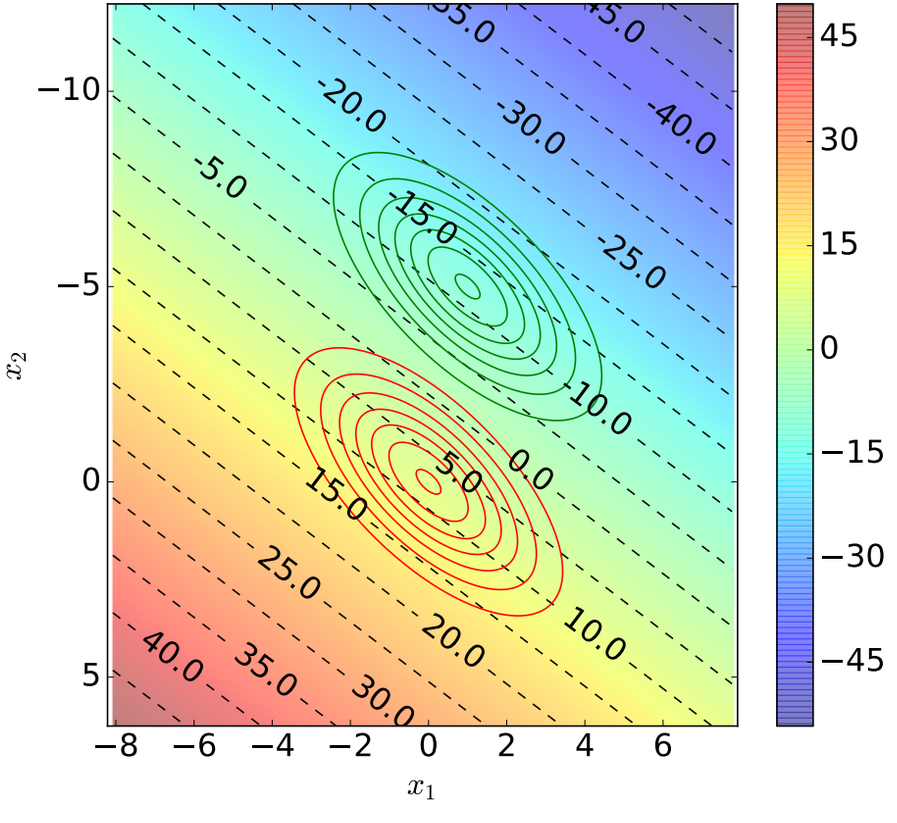
Log odds, shown as a map with contour plot. Note that the log odds are a linear function.

Example (contd.)

Priors: $p(\text{red}) = 0.1$, $p(\text{green}) = 0.9$



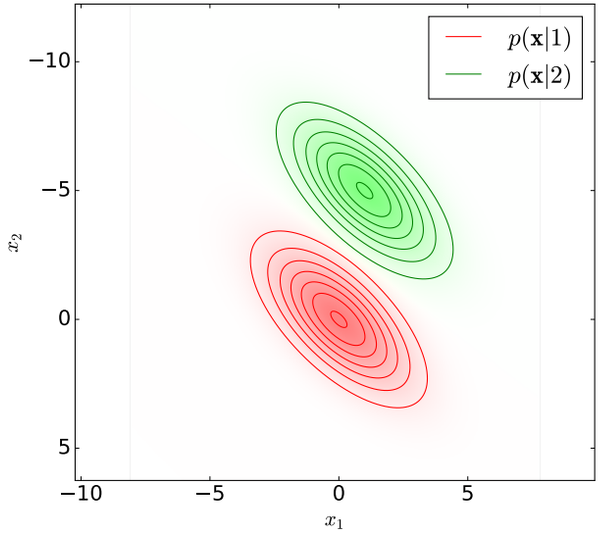
Probabilities $p(1|\mathbf{x})$ and $p(2|\mathbf{x})$.



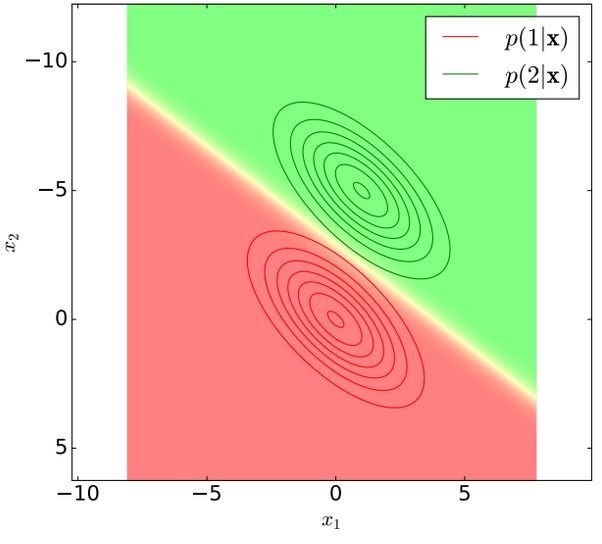
Log odds, shown as a map with contour plot. Note that the log odds are a linear function. Also note the shift of the zero level w.r.t. previous slide.

Example (contd.)

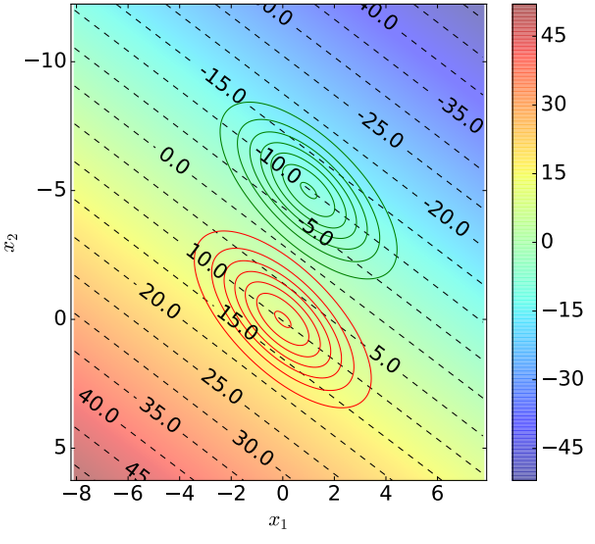
conditionals



posteriors



log odds



It will soon be shown for linear log odds, $a(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + w_0$, the posteriors can be written as

$$p(1|\mathbf{x}) = \frac{1}{1 + e^{-a(\mathbf{x})}} \tag{15}$$

$$p(2|\mathbf{x}) = \frac{1}{1 + e^{a(\mathbf{x})}} \tag{16}$$

The idea of logistic regression is that posteriors are modeled this way directly, without first estimating the priors and conditionals.

Other models with linear log odds (1)

Independent features with binary outcomes.

Feature vector $\mathbf{x} = (x_1, x_2, ..., x_D) \in \mathbb{R}^D$ $\quad$ $x_i \in \{0, 1\}$ (binary outcomes)

Each feature:

$$p(x_i = 1) = \pi_i, \tag{17}$$

$$p(x_i = 0) = 1 - \pi_i, \tag{18}$$

$\Rightarrow$ can be written as: $\quad p(x_i) = \pi_i^{x_i}(1 - \pi_i)^{1-x_i}$ (19)

Conditional probabilities for two classes 1, 2:

$$p(\mathbf{x}|1) = \prod_{i=1}^{D} \pi_i^{x_i}(1 - \pi_i)^{1-x_i} \tag{20}$$

$$p(\mathbf{x}|2) = \prod_{i=1}^{D} \kappa_i^{x_i}(1 - \kappa_i)^{1-x_i} \qquad (\pi_i, \kappa_i \in (0, 1), x_i \in \{0, 1\}) \tag{21}$$

Other models with linear log odds (1)

$$p(\mathbf{x}|1) = \prod_{i=1}^{D} \pi_i^{x_i}(1-\pi_i)^{1-x_i} \tag{22}$$

$$p(\mathbf{x}|2) = \prod_{i=1}^{D} \kappa_i^{x_i}(1-\kappa_i)^{1-x_i} \qquad (\pi_i, \kappa_i \in (0,1), x_i \in \{0,1\}) \tag{23}$$

The log odds are:

$$a(\mathbf{x}) = \ln \frac{p(\mathbf{x}|1)p(1)}{p(\mathbf{x}|2)p(2)} = \sum_{i=1}^{D} \{x_i \ln \pi_i + (1-x_i)\ln(1-\pi_i) - x_i \ln \kappa_i - (1-x_i)\ln(1-\kappa_i)\} \tag{24}$$

$$+ \ln \frac{p(1)}{p(2)} = \mathbf{w} \cdot \mathbf{x} + w_0 \qquad (\mathbf{w} \in \mathbb{R}^D, w_0 \in \mathbb{R}) \tag{25}$$

Note that the assumption that the features are independent may be quite strong. If this assumption is true (or anyway adopted), we talk about **naive Bayes** approach.

Other models with linear log odds (1)

Example

Problem: male/female classification

x_1 : hair length $>$ 5cm

x_2 : shoe size $>$ 41

Other models with linear log odds (2)

Multinomial naive Bayes

The analysis is similar to the case of binary outcomes. Here, the feature components x_i are not binary but they represent counts, summing to certain constant n ($\sum_{i=1}^{D} x_i = n$). The probabilities of observing the counts $\{x_i, i = 1, 2, .., D\}$ are

$$p(\mathbf{x}|1) = \frac{n!}{\prod_{i=1}^{D} x_i!} \prod_{i=1}^{D} \pi_i^{x_i} \tag{26}$$

$$p(\mathbf{x}|2) = \frac{n!}{\prod_{i=1}^{D} x_i!} \prod_{i=1}^{D} \kappa_i^{x_i} \qquad (x_i \in \mathbb{N}_0, \pi_i > 0, \kappa_i > 0, \sum_i \pi_i = 1, \sum_i \kappa_i = 1, \sum_{i=1}^{D} x_i = n) \tag{27}$$

It is easy to see that the log odds $a(\mathbf{x})$ are again linear in $\mathbf{x}$.

Other models with linear log odds (2)

Example

Two factories produces smarties. One of them ($k = 1$) is in Spain, the other ($k = 2$) is in Finland.

					
κ_i/Spain factory	0.5	0.3	0.05	0.05	0.1
π_i/Finland factory	0.05	0.05	0.1	0.5	0.3

(28)

$$p(\mathbf{x}|1) = \frac{n!}{\prod_{i=1}^{D} x_i!} \prod_{i=1}^{D} \pi_i^{x_i} \tag{29}$$

$$p(\mathbf{x}|2) = \frac{n!}{\prod_{i=1}^{D} x_i!} \prod_{i=1}^{D} \kappa_i^{x_i} \qquad (x_i \in \mathbb{N}_0, \pi_i > 0, \kappa_i > 0,$$

$$\sum_i \pi_i = 1, \sum_i \kappa_i = 1, \textstyle\sum_{i=1}^{D} x_i = n) \tag{30}$$

Logistic Regression, Model

Idea: Let us look for the log of the ratio of posteriors (log odds) $a(\mathbf{x})$ directly as a linear function of the input vector $\mathbf{x} = (x_1, x_2, ..., x_D) \in \mathbb{R}^D$ (D is the dimensionality of the feature space):

$$a(\mathbf{x}) = \ln \frac{p(1|\mathbf{x})}{p(2|\mathbf{x})} = \mathbf{w} \cdot \mathbf{x} + w_0 \,, \qquad \mathbf{w} = (w_1, w_2, ..., w_D) \in \mathbb{R}^D \,,$$
$$w_0 \in \mathbb{R} \,, \tag{31}$$

where w_0 is the bias term.

Let us rewrite this as

$$a(\mathbf{x}) = \underset{\substack{\uparrow \\ x_0}}{w_0 \cdot 1} + \mathbf{w} \cdot \mathbf{x} = [w_0, w_1, w_2, ..., w_D] \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_D \end{bmatrix} = \mathbf{w}' \cdot \mathbf{x}'. \qquad (x_0 = 1) \tag{32}$$

Note: From now on, we will drop the dash sign and write again only '$\mathbf{x}$' or '$\mathbf{w}$', with the understanding that these **include the zero-index components** $x_0 = 1$ and $w_0 \in \mathbb{R}$ implementing the **bias**.

Logistic Regression, Log Odds and Posteriors (1)

Here is the relationship between the the log odds $a(\mathbf{x})$ and the posterior probabilities $p(1|\mathbf{x})$ and $p(2|\mathbf{x})$.

The log odds $a(\mathbf{x})$ is (remember the bias term is consumed in the $\mathbf{x}$ and $\mathbf{w}$)

$$a(\mathbf{x}) = \ln \frac{p(1|\mathbf{x})}{p(2|\mathbf{x})} = \mathbf{w} \cdot \mathbf{x}\,. \tag{33}$$

From this, it follows that

$$\frac{p(1|\mathbf{x})}{p(2|\mathbf{x})} = \exp(\mathbf{w} \cdot \mathbf{x}) \qquad \frac{p(2|\mathbf{x})}{p(1|\mathbf{x})} = \exp(-\mathbf{w} \cdot \mathbf{x}) \tag{34}$$

$$p(1|\mathbf{x}) = p(2|\mathbf{x}) \exp(\mathbf{w} \cdot \mathbf{x}) \qquad p(2|\mathbf{x}) = p(1|\mathbf{x}) \exp(-\mathbf{w} \cdot \mathbf{x}) \tag{35}$$

$$1 = p(1|\mathbf{x}) + p(2|\mathbf{x}) = p(1|\mathbf{x}) \left(1 + e^{-\mathbf{w}\cdot\mathbf{x}}\right) = p(2|\mathbf{x}) \left(1 + e^{\mathbf{w}\cdot\mathbf{x}}\right) \tag{36}$$

$$p(1|\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}\cdot\mathbf{x}}}\,, \qquad p(2|\mathbf{x}) = \frac{1}{1 + e^{\mathbf{w}\cdot\mathbf{x}}}\,. \tag{37}$$

Logistic Regression, Log Odds and Posteriors (2)

Again,

$$a(\mathbf{x}) = \ln \frac{p(1|\mathbf{x})}{p(2|\mathbf{x})} = \mathbf{w} \cdot \mathbf{x} \tag{38}$$

$$p(1|\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}\cdot\mathbf{x}}} = \sigma(\mathbf{w} \cdot \mathbf{x}) \tag{39}$$

$$p(2|\mathbf{x}) = \frac{1}{1 + e^{\mathbf{w}\cdot\mathbf{x}}} = \sigma(-\mathbf{w} \cdot \mathbf{x}) \tag{40}$$

where $\sigma(u) = 1/(1 + \exp(-u))$ is the **logistic sigmoid** function.

It will be advantageous to rename the clases from $(1, 2)$ to $(1, -1)$. Then we can rewrite the equations (39, 40) as

$$p(k|\mathbf{x}) = \frac{1}{1 + e^{-k\mathbf{w}\cdot\mathbf{x}}}, \qquad k \in \{-1, 1\} \tag{41}$$

Finding $\mathbf{w}$: Objective $E(\mathbf{w})$

$$p(k|\mathbf{x}) = \frac{1}{1+e^{-k\mathbf{w}\cdot\mathbf{x}}}, \qquad k \in \{-1, 1\} \tag{42}$$

How do we find $\mathbf{w}$?

We apply the Maximum Likelihood approach for finding $\mathbf{w}$. Let us have the training set $\mathcal{T} = \{(\mathbf{x}_1, k_1), (\mathbf{x}_2, k_2), ..., (\mathbf{x}_N, k_N)\}$. The optimal $\mathbf{w}^*$ would be the one maximizing the log-likelihood $l(\mathbf{w})$,

$$l(\mathbf{w}) = \sum_{(\mathbf{x},k)\in\mathcal{T}} \ln p(\mathbf{x},k)_{[\mathbf{w}]} = \sum_{(\mathbf{x},k)\in\mathcal{T}} \ln p(k|\mathbf{x})_{[\mathbf{w}]} + \sum_{(\mathbf{x},k)\in\mathcal{T}} \ln p(\mathbf{x})_{[\mathbf{w}]}\,, \tag{43}$$

where for the sake of clarity, dependence on $\mathbf{w}$ is denoted by a subscript $[\mathbf{w}]$.

As there are no assumptions about the form of $p(\mathbf{x})$ as a function of $\mathbf{w}$, logistic regression employs instead the maximization of *conditional log-likelihood* $l'(\mathbf{w})$

$$l'(\mathbf{w}) = \sum_{(\mathbf{x},k)\in\mathcal{T}} \ln p(k|\mathbf{x})_{[\mathbf{w}]} = -\sum_{(\mathbf{x},k)\in\mathcal{T}} \ln(1+e^{-k\mathbf{w}\cdot\mathbf{x}}) \qquad \text{(conditional log likelihood)} \tag{44}$$

$$\mathbf{w}^* = \operatorname*{argmax}_{\mathbf{w}} l'(\mathbf{w}) \qquad \text{(optimal } \mathbf{w}^*\text{)} \tag{45}$$

Finding $\mathbf{w}$: Objective $E(\mathbf{w})$

(copied from previous slide)

$$l'(\mathbf{w}) = \sum_{(\mathbf{x},k)\in\mathcal{T}} \ln p(k|\mathbf{x})_{[\mathbf{w}]} = -\sum_{(\mathbf{x},k)\in\mathcal{T}} \ln(1+e^{-k\mathbf{w}\cdot\mathbf{x}}) \qquad \text{(conditional log likelihood)} \tag{46}$$

$$\mathbf{w}^* = \operatorname*{argmax}_{\mathbf{w}} l'(\mathbf{w}) \qquad \text{(optimal } \mathbf{w}^*\text{)} \tag{47}$$

In order for the optimization to fit into the **minimization** framework, we define the objective function $E(\mathbf{w})$ as the negative conditional log likelihood, $E(\mathbf{w}) = -l'(\mathbf{w})$. This objective function corresponds to **cross entropy**. Let us now analyze the properties of $E(\mathbf{w})$.

Finding $\mathbf{w}$: Gradient of $E(\mathbf{w})$

$$E(\mathbf{w}) = -\sum_{(\mathbf{x},k)\in\mathcal{T}} \ln p(k|\mathbf{x}) = \sum_{(\mathbf{x},k)\in\mathcal{T}} \ln(1+e^{-k\mathbf{w}\cdot\mathbf{x}}) \tag{48}$$

(49)

The gradient vector $g(\mathbf{w})$ of E is:

$$g(\mathbf{w}) = \frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = \sum_{(\mathbf{x},k)\in\mathcal{T}} \frac{e^{-k\mathbf{w}\cdot\mathbf{x}}}{1+e^{-k\mathbf{w}\cdot\mathbf{x}}}(-k\mathbf{x}) = -\sum_{(\mathbf{x},k)\in\mathcal{T}} \underbrace{\frac{1}{1+e^{k\mathbf{w}\cdot\mathbf{x}}}}_{p(-k|\mathbf{x})} k\mathbf{x} \tag{50}$$

$$= -\sum_{(\mathbf{x},k)\in\mathcal{T}} (1-p(k|\mathbf{x}))\, k\mathbf{x}\,. \tag{51}$$

We require $g(\mathbf{w}) = 0$ (the necessary condition for optimality). However, it seems that these equations **cannot be solved analytically**. We will need to resort to the numerical optimization methods. Let us continue and check the second order derivatives.

Finding $\mathbf{w}$: $E(\mathbf{w})$ is convex

The Hessian matrix $H(\mathbf{w})$ of the objective function E is

$$H(\mathbf{w}) = \frac{\partial^2 E(\mathbf{x})}{\partial \mathbf{w}^2} = \frac{\partial g(\mathbf{w})}{\partial \mathbf{w}} = -\frac{\partial}{\partial \mathbf{w}} \sum_{(\mathbf{x},k)\in\mathcal{T}} \frac{1}{1+e^{k\mathbf{w}\cdot\mathbf{x}}} k\mathbf{x} \tag{52}$$

$$= \sum_{(\mathbf{x},k)\in\mathcal{T}} \underbrace{\frac{e^{k\mathbf{w}\cdot\mathbf{x}}}{(1+e^{k\mathbf{w}\cdot\mathbf{x}})^2} k^2}_{>0} \mathbf{x}\mathbf{x}^\top = \sum_{(\mathbf{x},k)\in\mathcal{T}} p(-1|\mathbf{x})p(1|\mathbf{x})\ \mathbf{x}\mathbf{x}^\top \tag{53}$$

This is a very important result. It shows that the Hessian matrix $H(\mathbf{w})$ is **positive definite** in every point $\mathbf{w}$ and, therefore, the function $E(\mathbf{w})$ is **convex**. As a consequence, $E(\mathbf{w})$ has a **unique minimum**.

Note. Can you show that $H(\mathbf{w})$ is positive definite?

Finding $\mathbf{w}$: Gradient Descent

Any method of convex optimization can be used to find the optimal $\mathbf{w}^*$. For the examples in this lecture, the following gradient descent method with adaptive step size has been used:

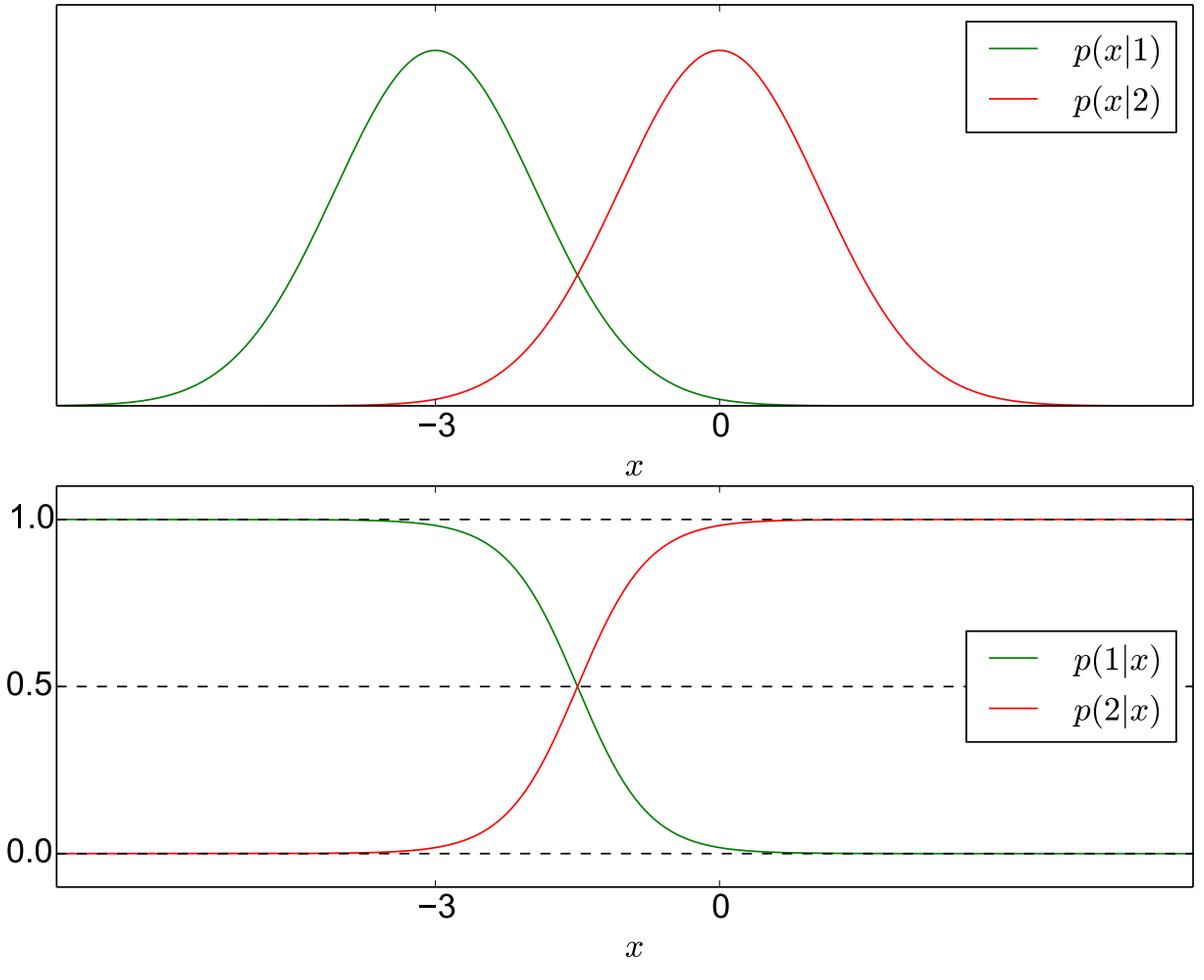
```
# input: x (observations), k (class labels), w_init (initial w)

# init:
w = w_init
step_size = 1.0
E, g = compute_E_and_gradient(x, k, w)

# iterate:
while not TERMINATION_CONDITION:
    E_new, g_new = compute_E_and_gradient(x, k, w - step_size * g)

    if E_new < E:
        # success.
        w -= step_size * g
        g = g_new
        E = E_new
        step_size *= 2
    else:
        step_size /= 2

return w
```

Notes:

i) Iteration is accepted if $E(\mathbf{w})$ decreases. If it hasn't decreased, either the step size is too high (thus it is halved), or optimum has been already found.

ii) We normalize the gradient by the number of training data N because otherwise its magnitude scales linearly with N, causing the necessity for smaller step sizes with higher N.

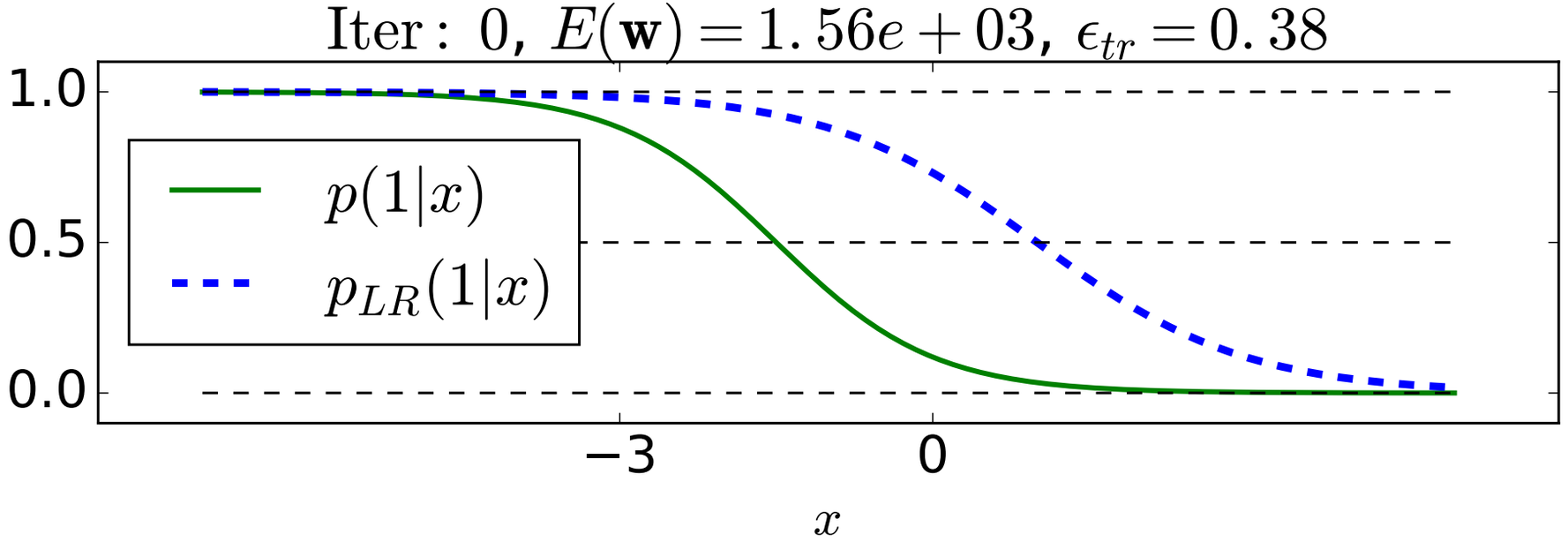
Example 1, Two Normal Distribs with Equal Variance (1)



$p(x|1) = \mathcal{N}(x|\mu_1 = -3, \sigma_1 = 1.5)$
$p(x|2) = \mathcal{N}(x|\mu_2 = 0, \sigma_2 = 1.5)$
$p(1) = p(2) = 0.5$. **Bayesian error is** $\epsilon_B = 0.16$.

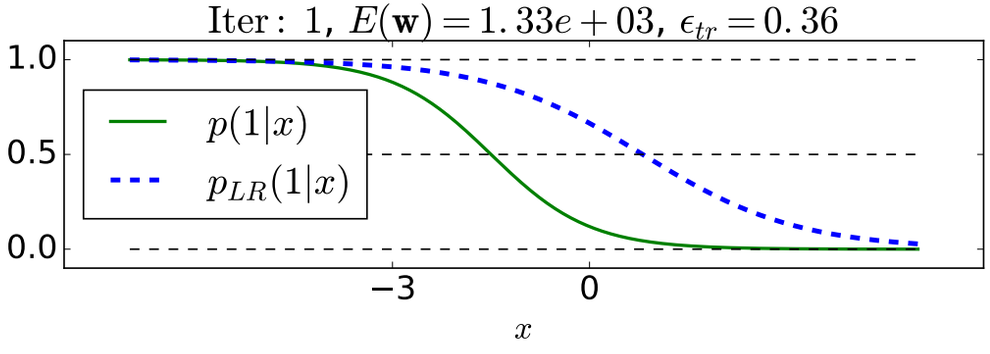
Example 1, Two Normal Distribs with Equal Variance (2)

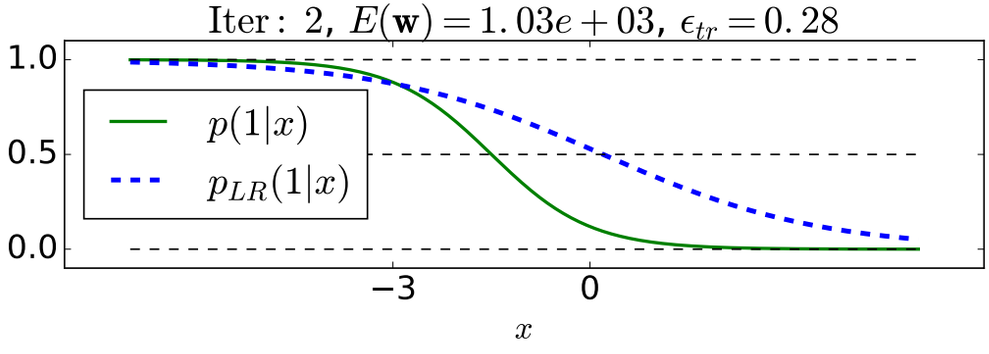
Initial state.
Training set: 1000 samples from each of the distributions.

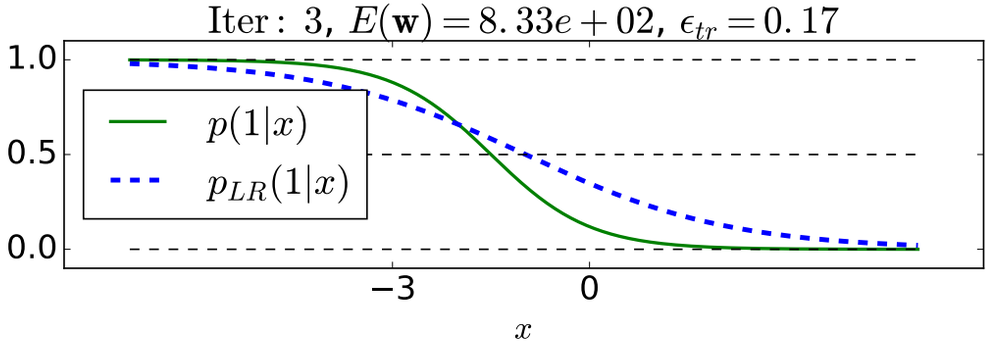


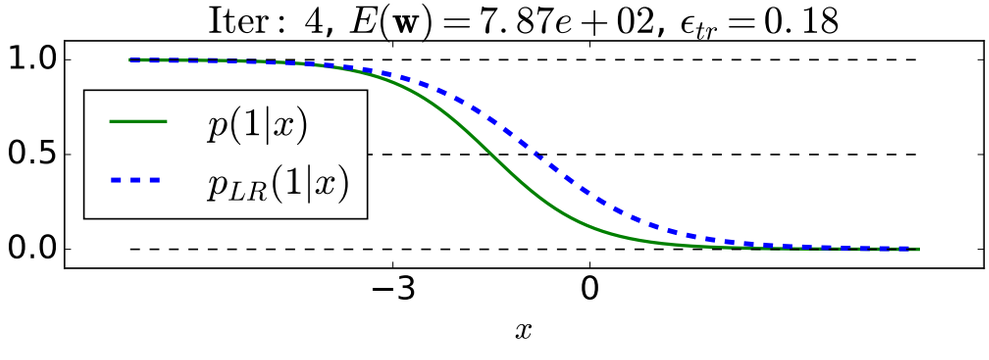
$p(1|x)$: The actual conditional for the 1st class.
$p_{LR}(1\,|\,x)$: The conditional for the 1st class predicted by logistic regression.
$E(\mathbf{w})$: the value of cross entropy.
ϵ_{tr} : the training error (error on the training set.)
(initial $w = [1, -1]^\top$)

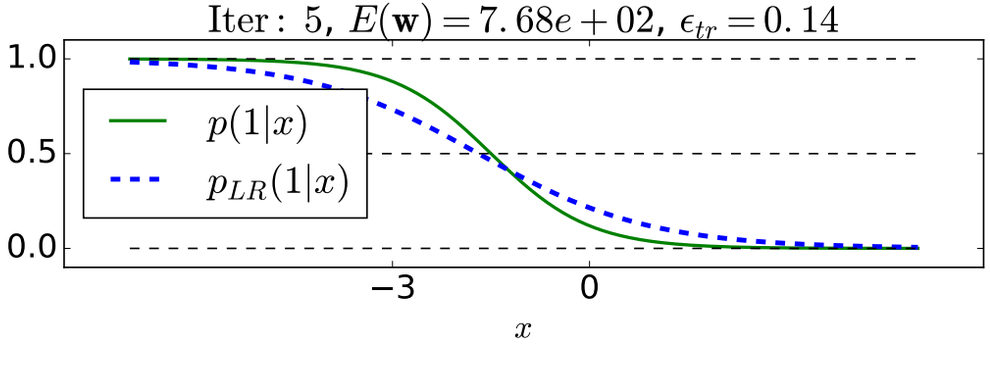
Example 1, Two Normal Distribs with Equal Variance (3)

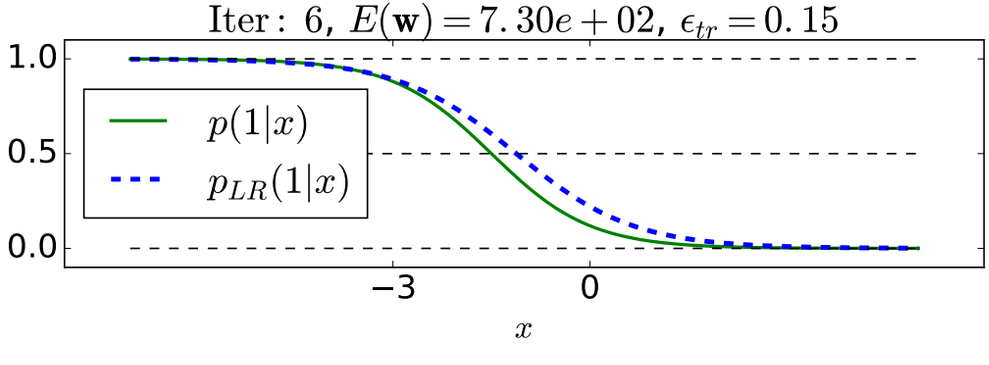




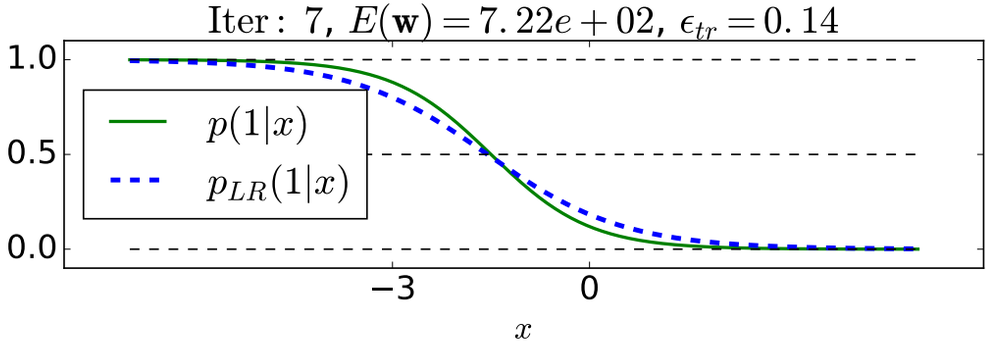


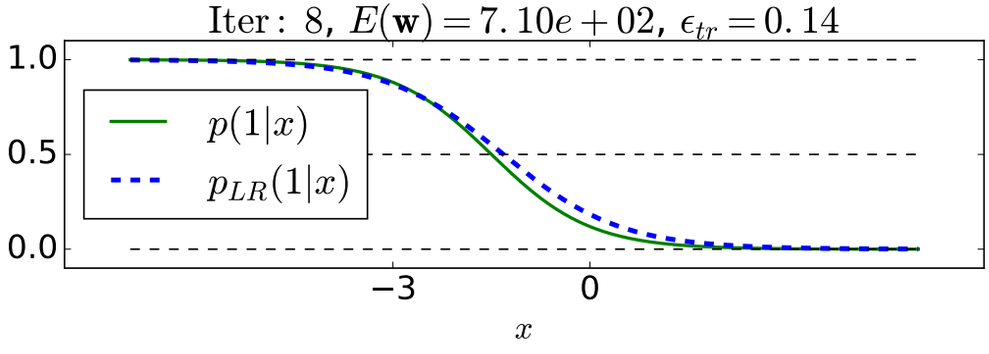


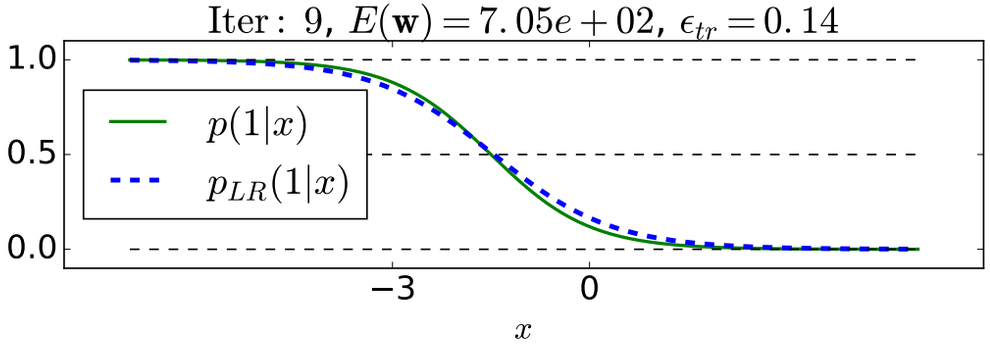


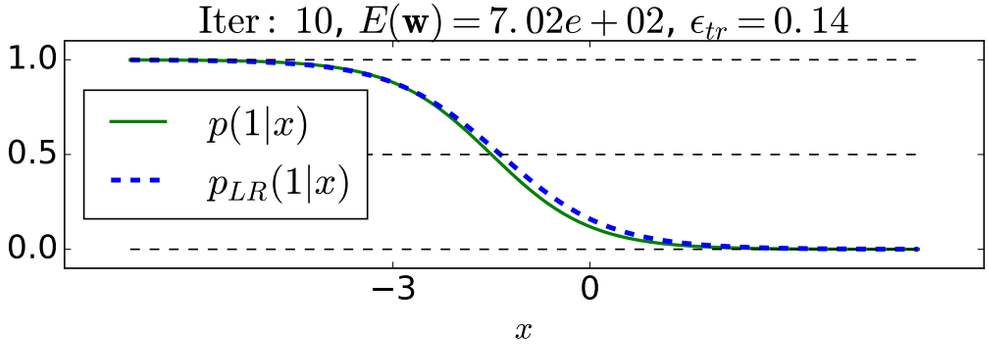


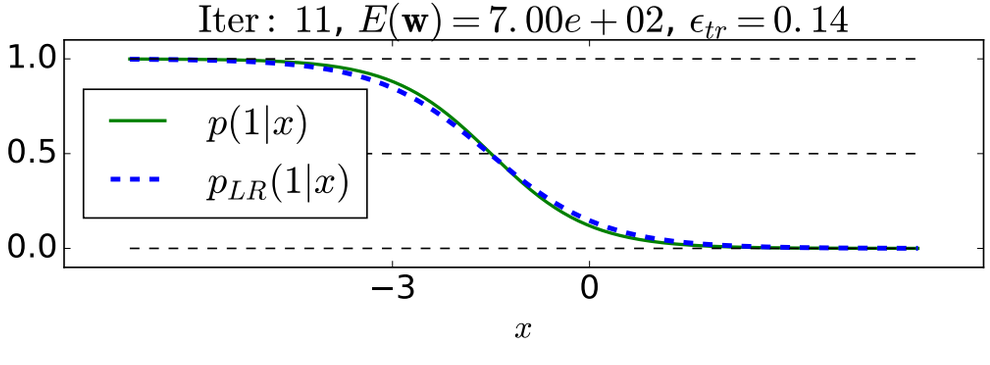
Example 1, Two Normal Distribs with Equal Variance (4)

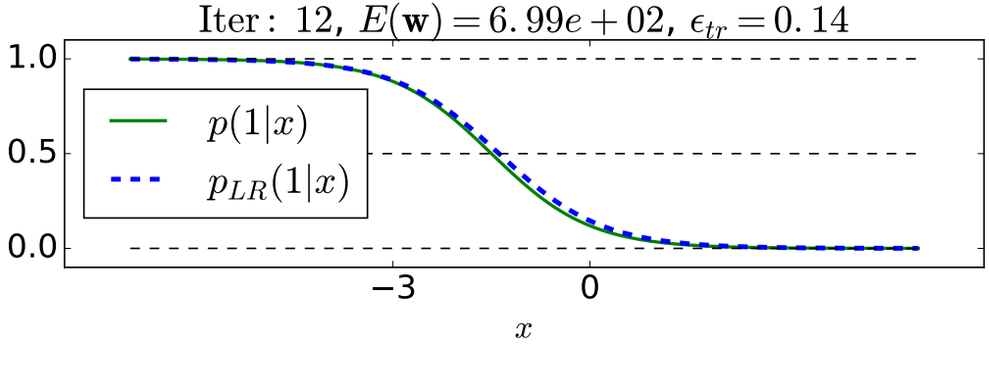




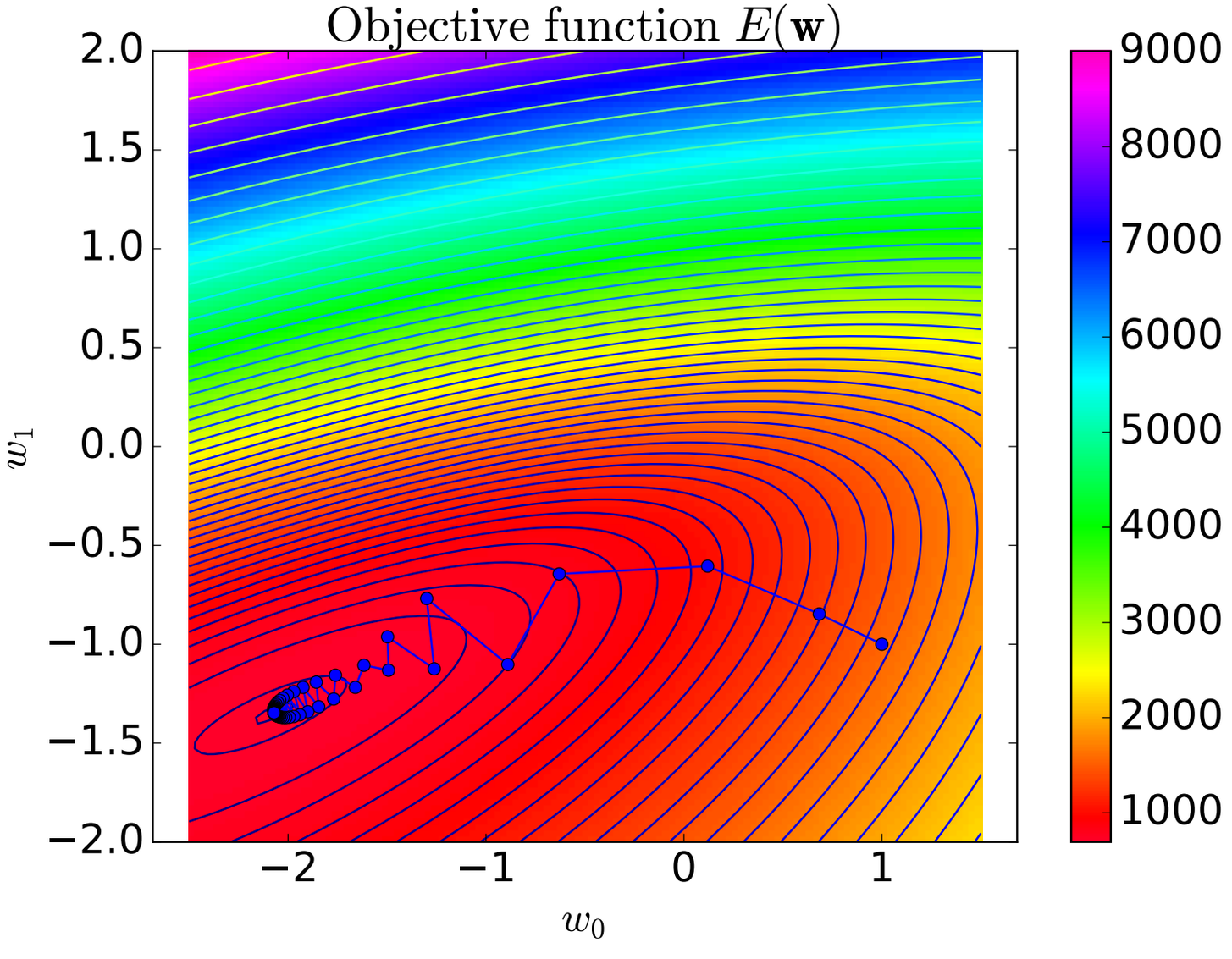






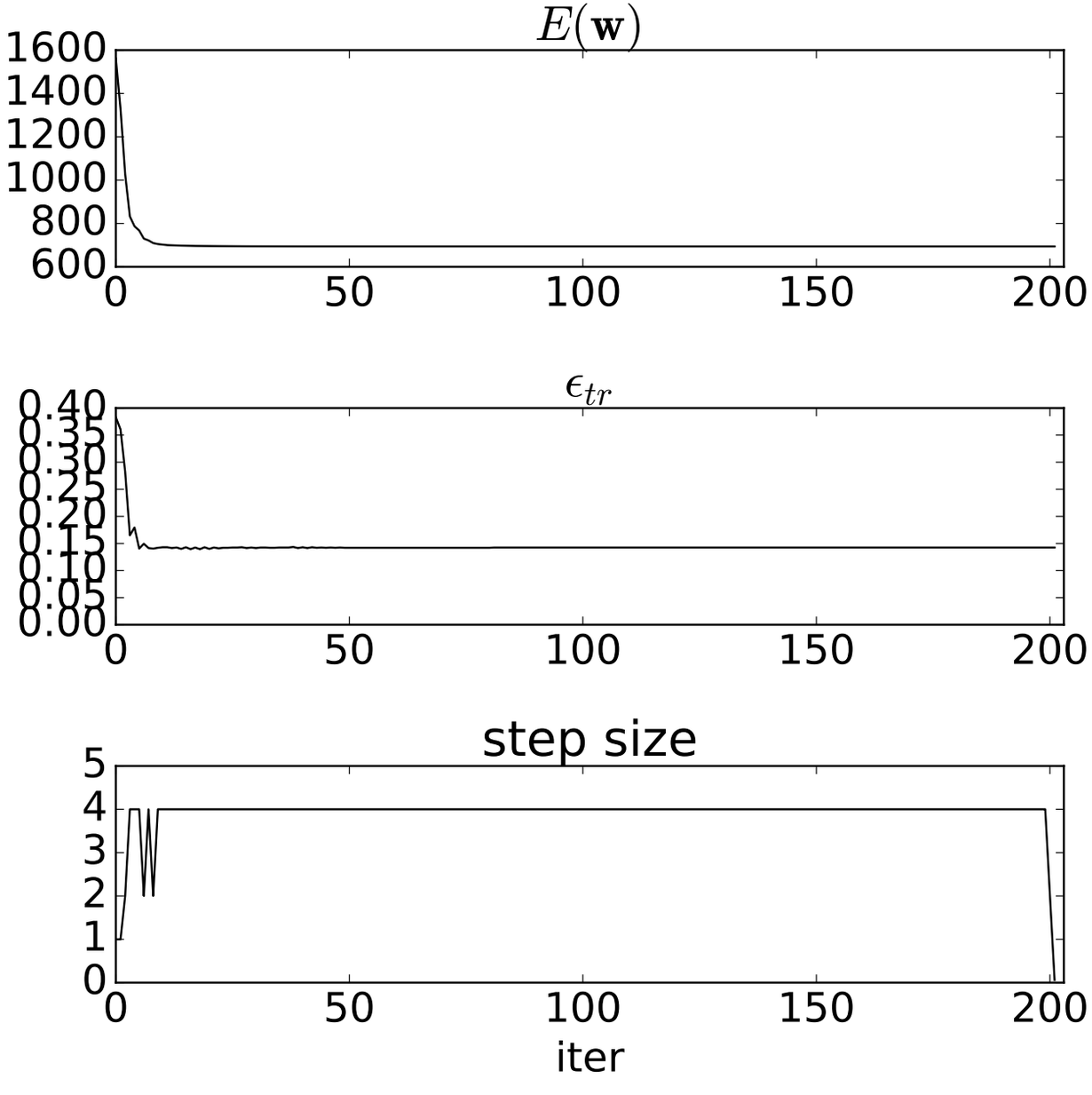


Example 1, Two Normal Distribs with Equal Variance (5)



The cross-entropy $E(\mathbf{w})$ and the progress of $\mathbf{w}$ with iterations.

Example 1, Two Normal Distribs with Equal Variance (6)

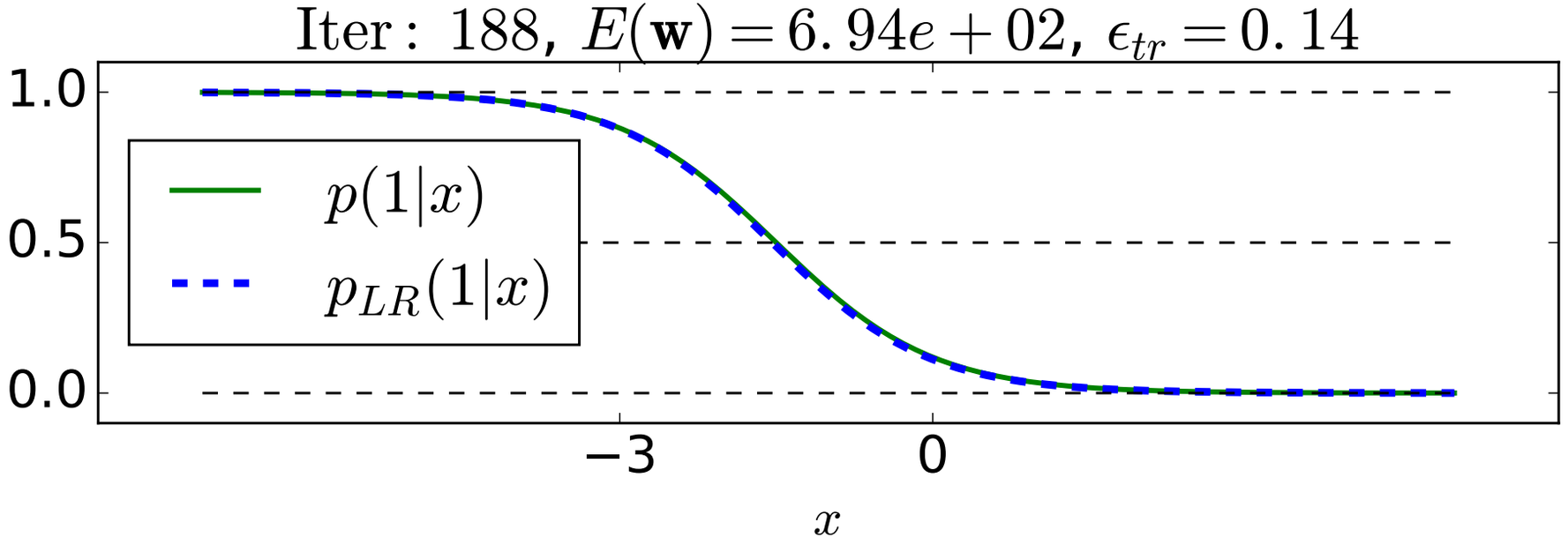


Example 1, Two Normal Distribs with Equal Variance (7)

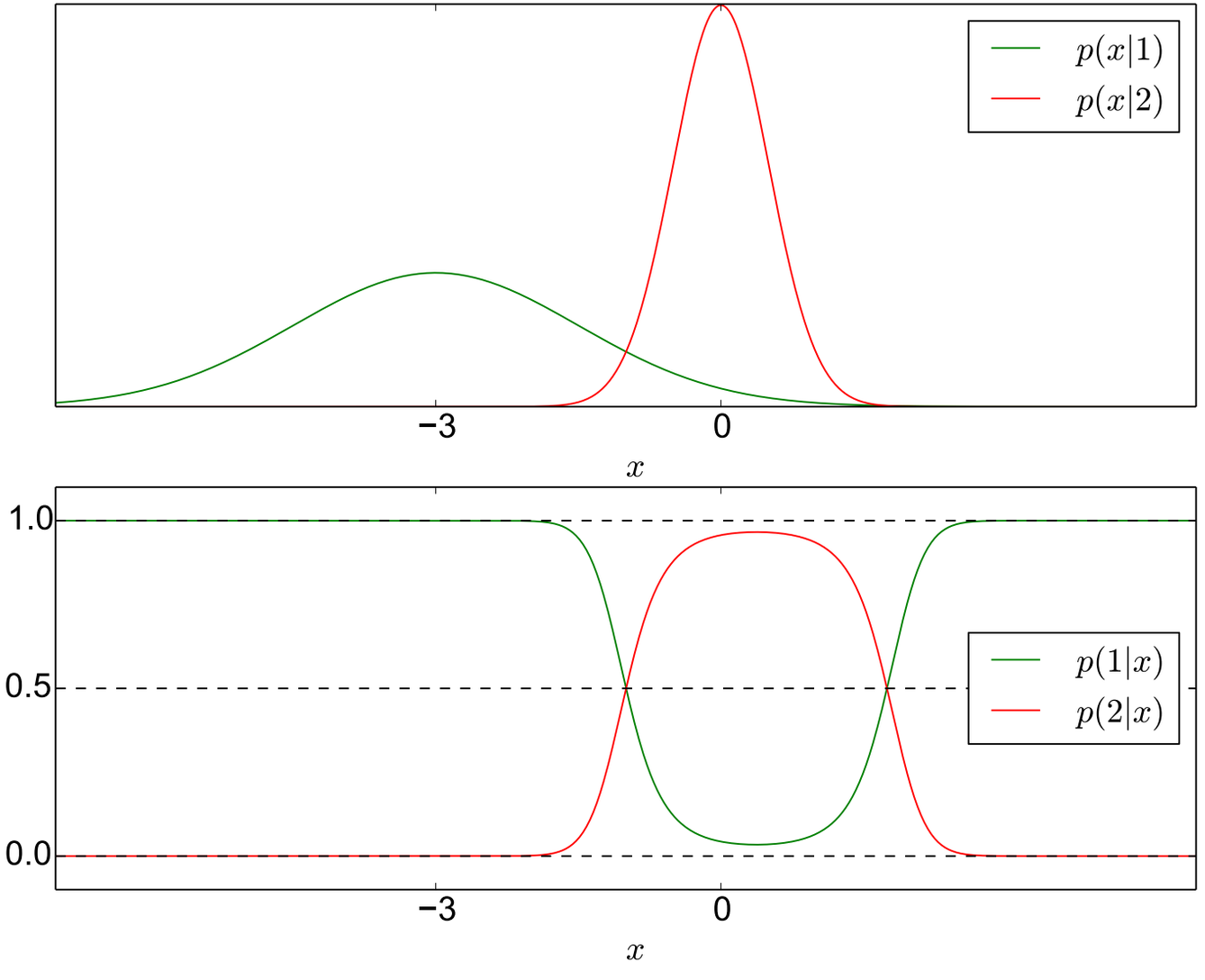
Things to note:

- ϵ_{tr} does not monotonically decrease with iterations. $E(\mathbf{w})$ does.
- Some intermediate ϵ_{tr}'s as well as the final one are lower than the Bayesian error $\epsilon_B = 0.16$. This is not a contradiction of the theory.

Converged state:
$w = [-2.07, -1.35]^\top$.



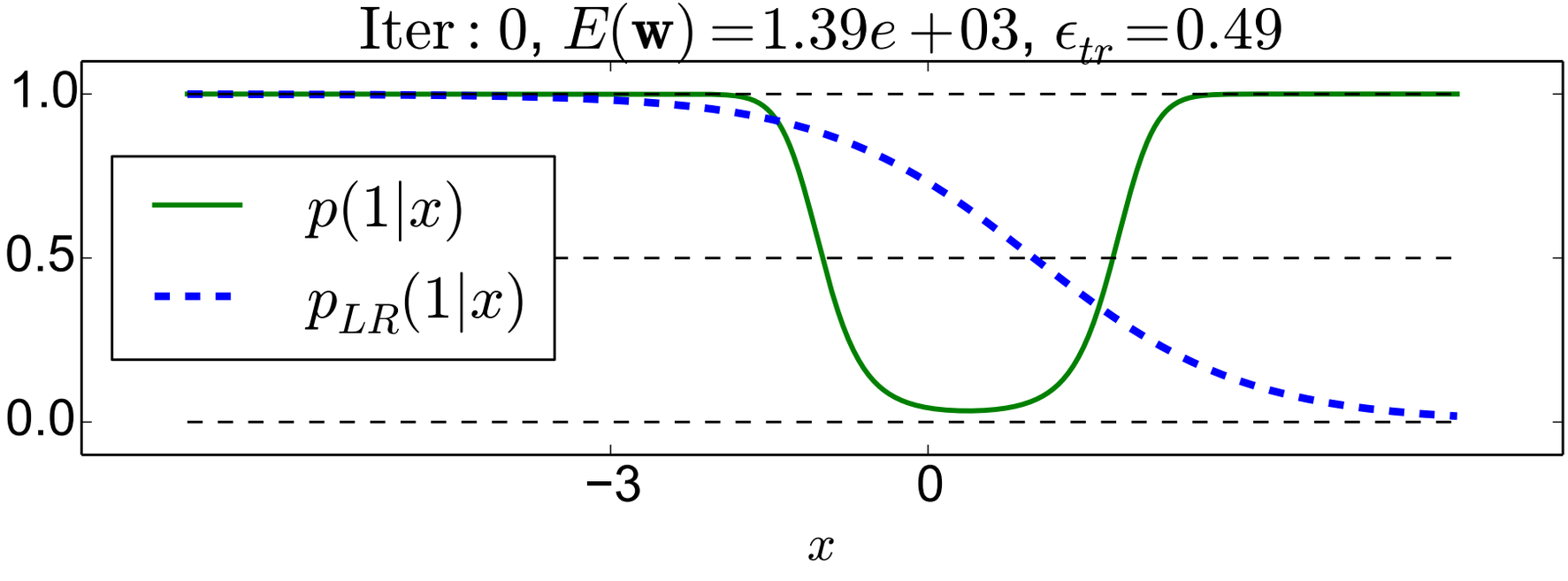
Example 2, Non-Equal Variance but Different Mean (1)



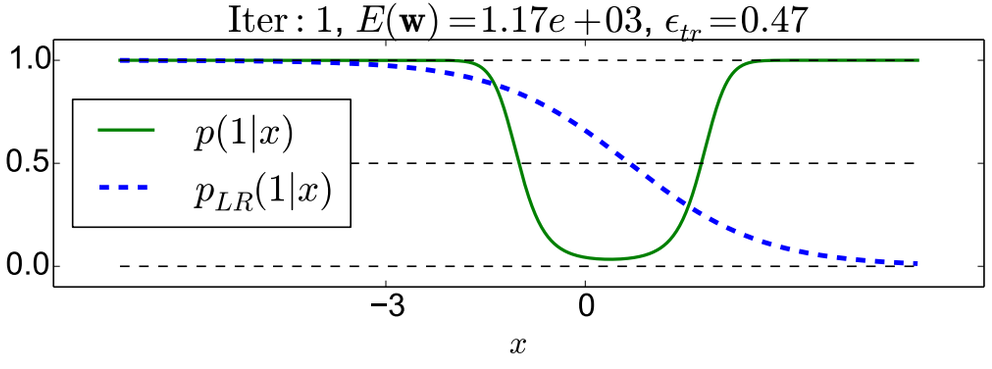
$p(x|1) = \mathcal{N}(x|\mu_1 = -3, \sigma_1 = 1.5)$
$p(x|2) = \mathcal{N}(x|\mu_2 = 0, \sigma_2 = 0.5)$
$p(1) = p(2) = 0.5$. **Bayesian error is** $\epsilon_B = 0.057$.

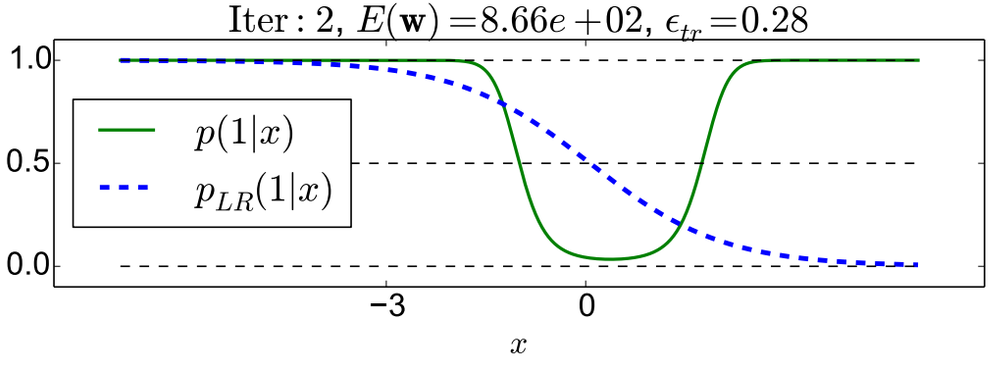
Example 2, Non-Equal Variance but Different Mean (2)

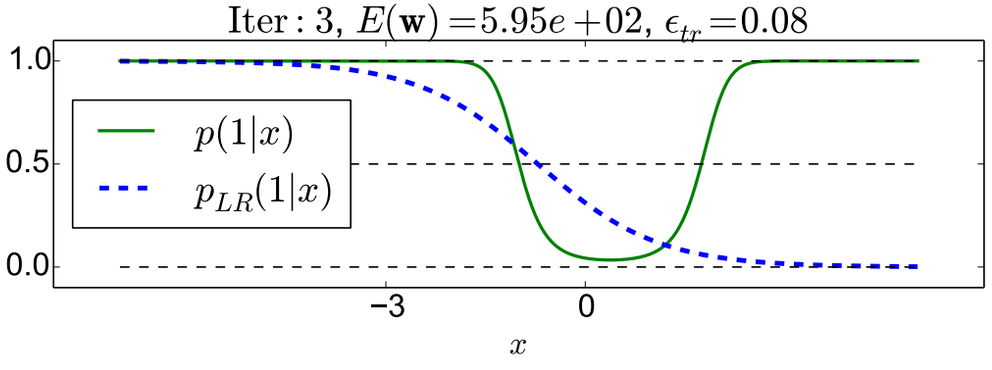
Initial state.
$w = [1, -1]^\top$.
Training set: 1000 samples from each of the distributions.

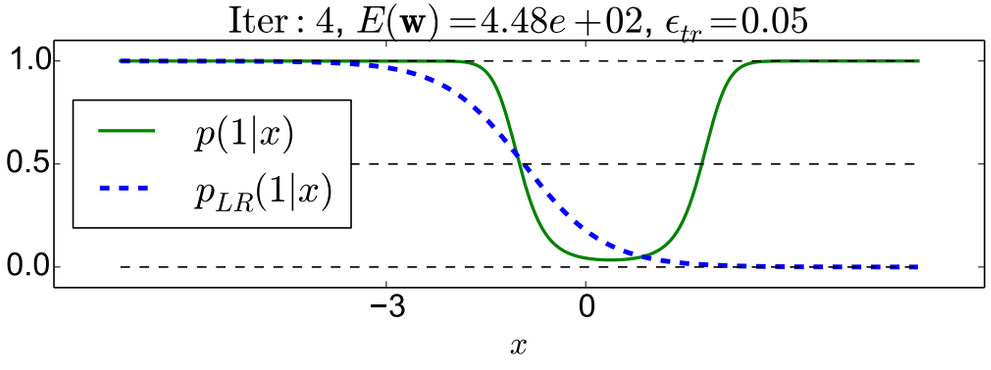


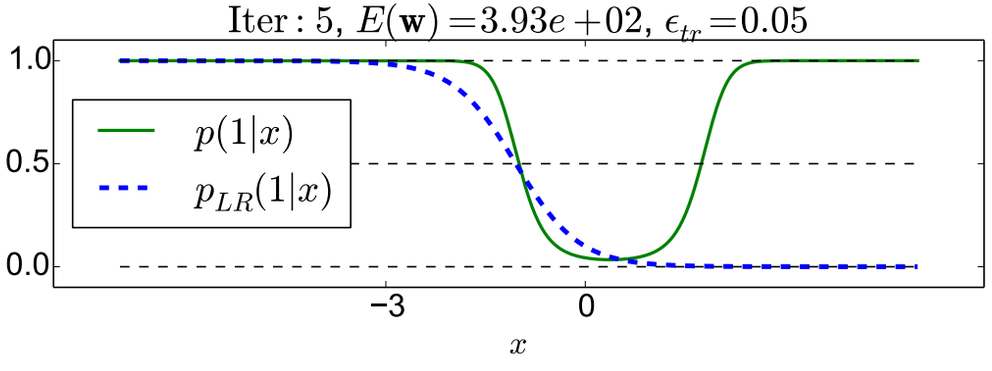
Example 2, Non-Equal Variance but Different Mean (3)

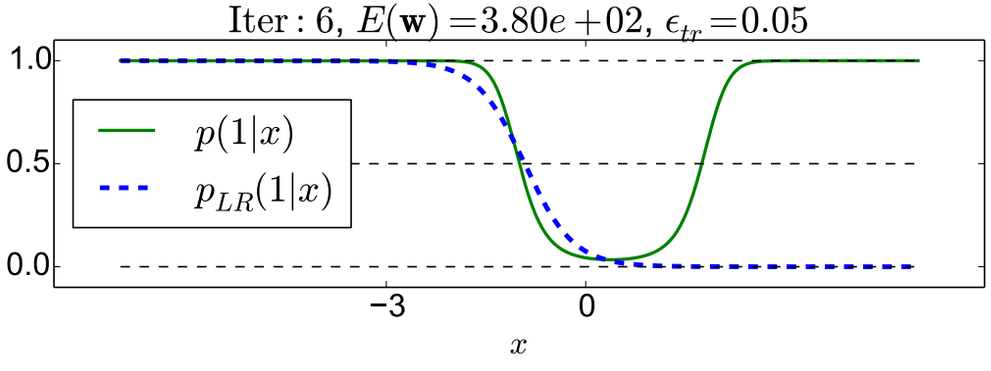




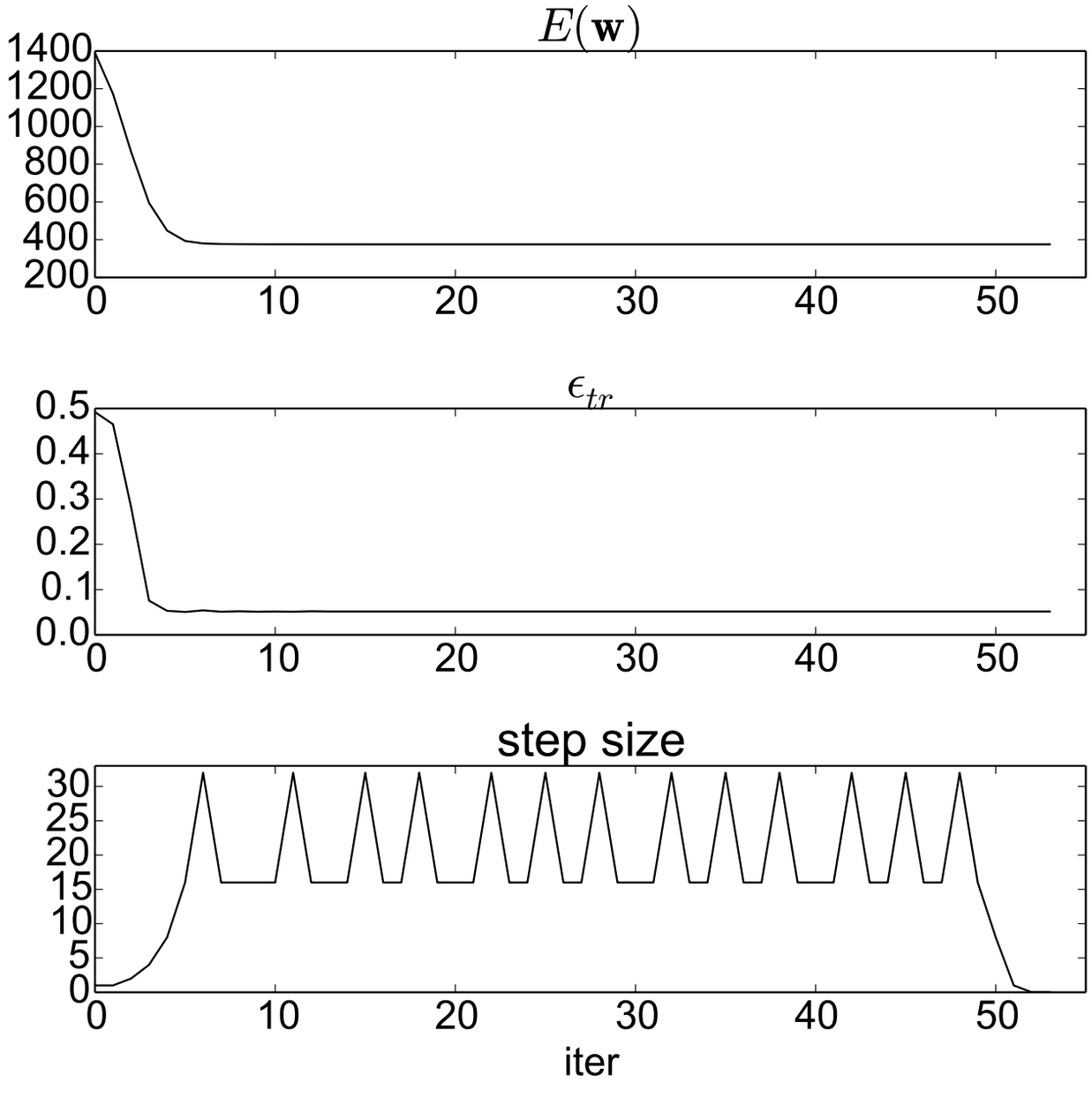








Example 2, Non-Equal Variance but Different Mean (4)

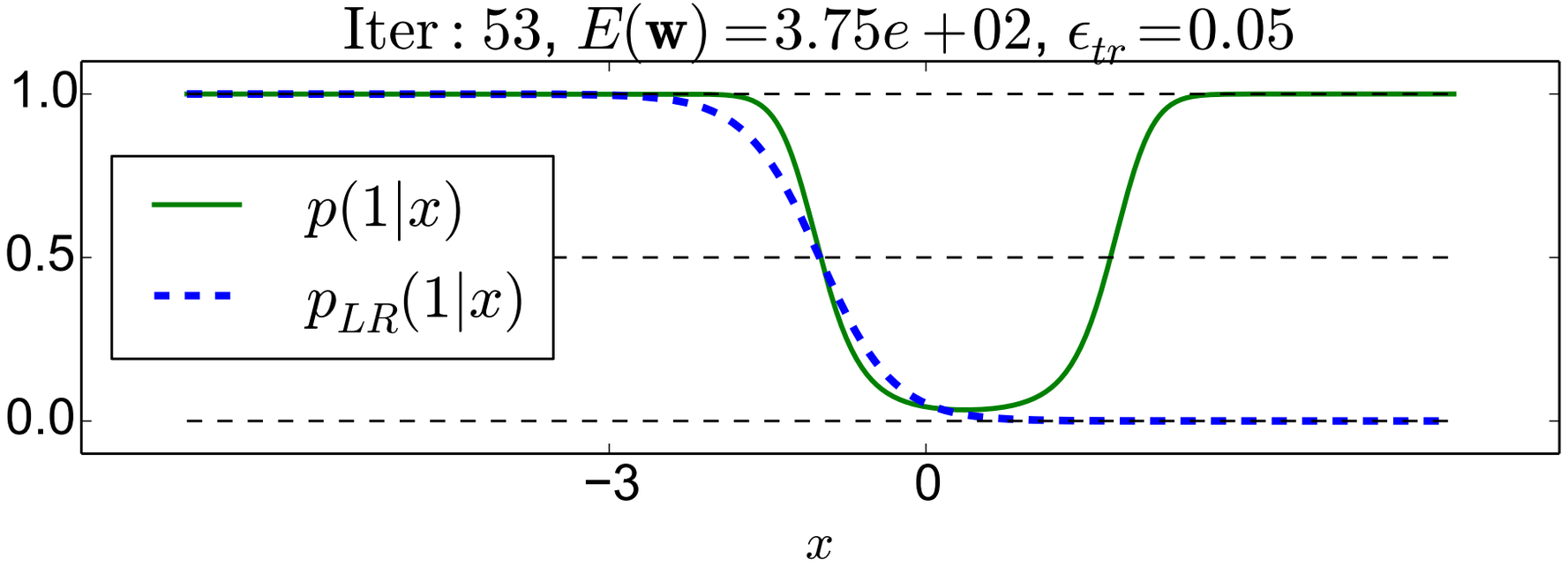


Example 2, Non-Equal Variance but Different Mean (5)

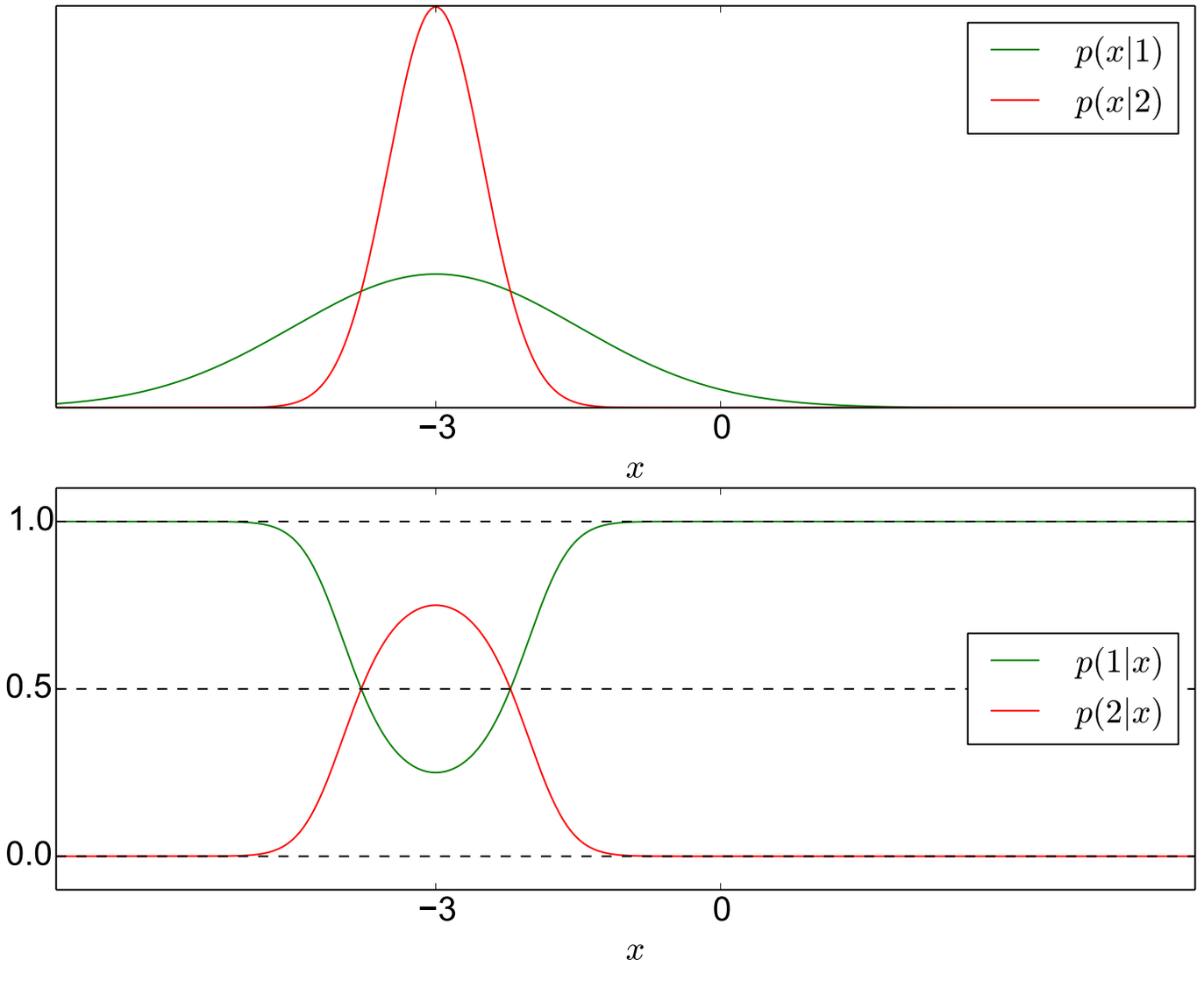
Things to note:

- The logistic regression cannot provide the two thresholds the optimal decision strategy requires. But it provides the one threshold which matters more in reducing the classification error (here the left one.)

Converged state.
$w = [-2.88, -2.85]^\top$.



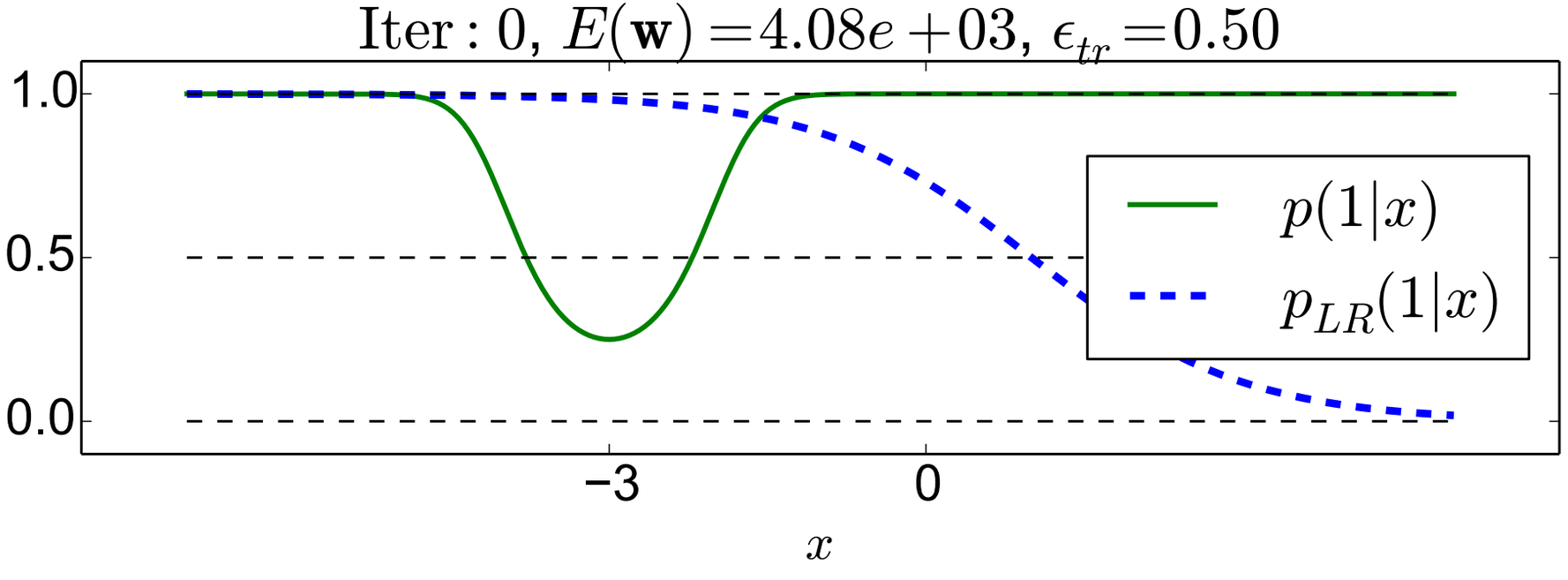
Example 3, Non-Equal Variance and the Same Mean (1)



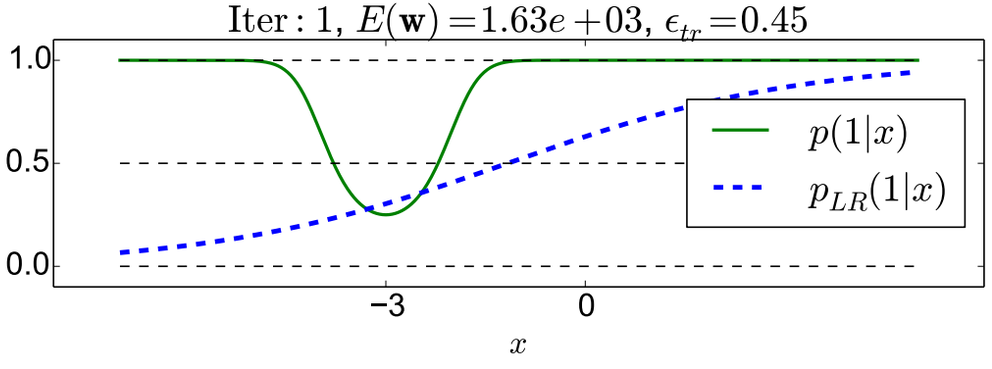
$p(x|1) = \mathcal{N}(x|\mu_1 = -3, \sigma_1 = 1.5)$
$p(x|2) = \mathcal{N}(x|\mu_2 = -3, \sigma_2 = 0.5)$
$p(1) = p(2) = 0.5$. **Bayesian error is** $\epsilon_B = 0.26$.

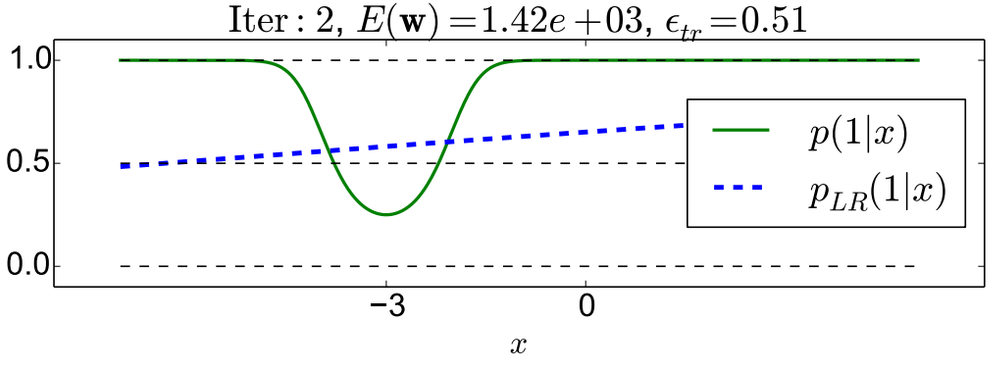
Example 3, Non-Equal Variance and the Same Mean (2)

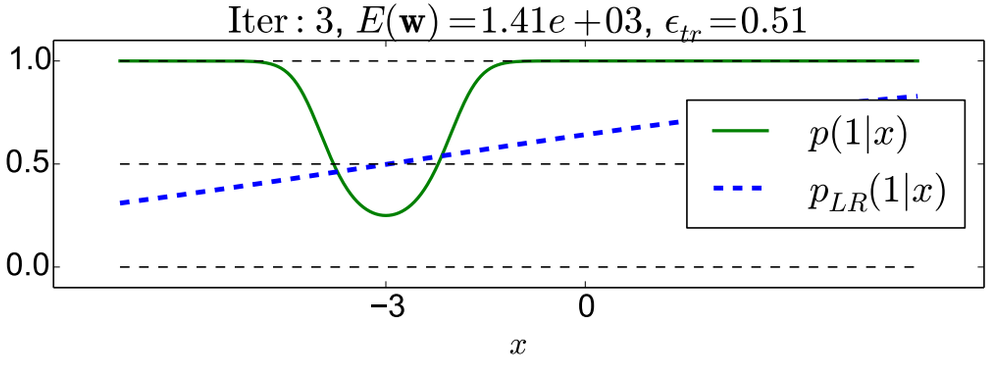
Initial state.
$w = [1, -1]^\top$.
Training set: 1000 samples from each of the distributions.

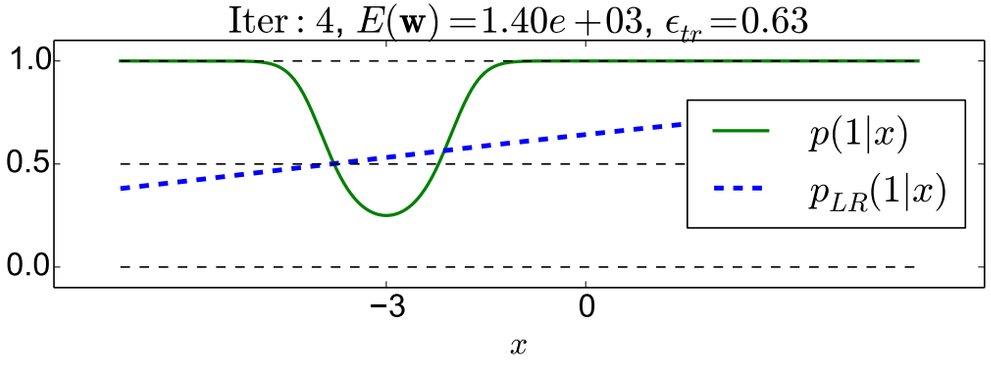


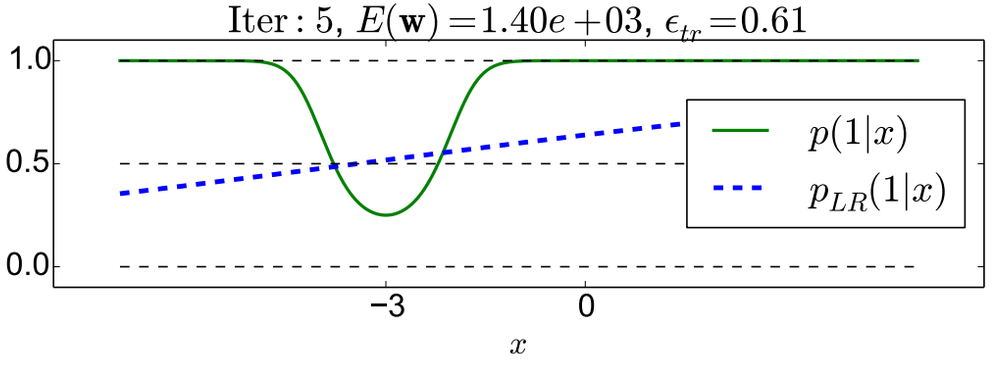
Example 3, Non-Equal Variance and the Same Mean (3)

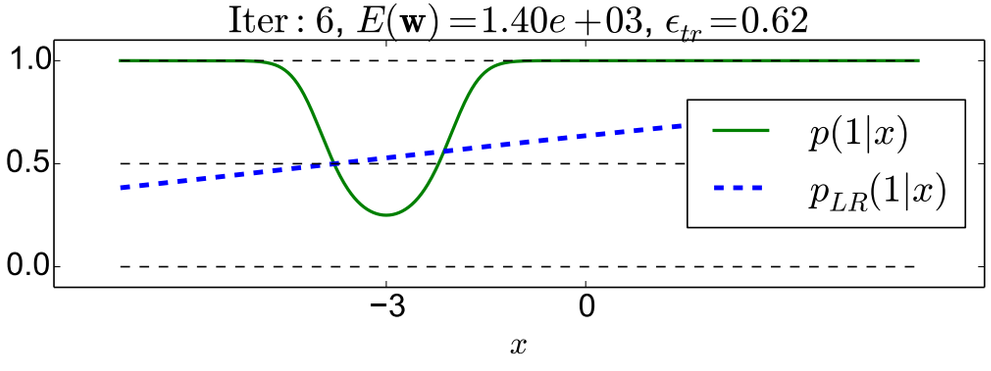




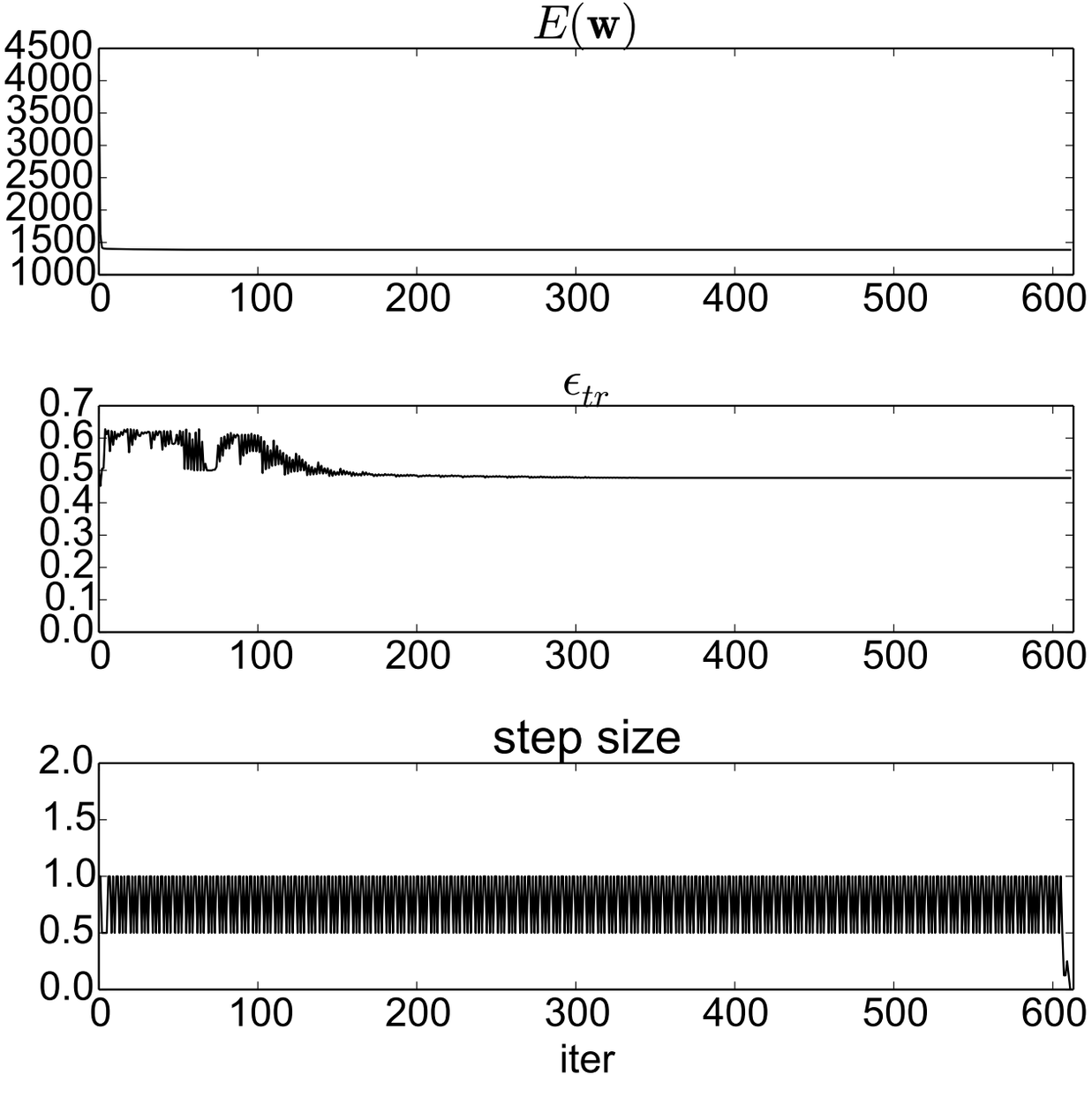








Example 3, Non-Equal Variance and the Same Mean (4)

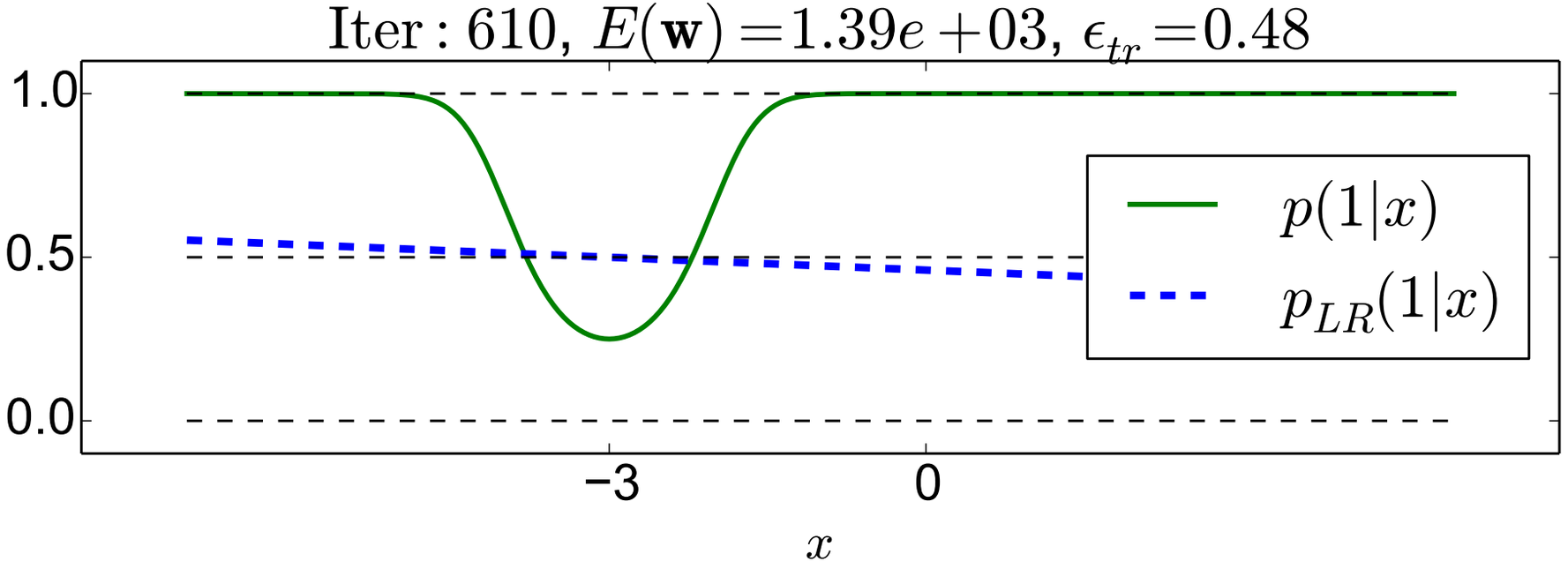


Example 3, Non-Equal Variance and the Same Mean (5)

Things to note:

- Failure case. The logistic regression cannot provide a good fit to the log odds in this case.

Final state:
$w = [-0.161, -0.053]^\top$.



Logistic Regression with Multiple Classes (1)

The logistic regression can be generalized to multiple classes as follows.

Each class $k \in \{1, 2, .., K\}$ has an associated weight vector $\mathbf{w}_k$.

The conditional probability for the k-th function is computed using the **softmax** function:

$$p(k|\mathbf{x}) = \frac{e^{\mathbf{w}_k \cdot \mathbf{x}}}{e^{\mathbf{w}_1 \cdot \mathbf{x}} + e^{\mathbf{w}_2 \cdot \mathbf{x}} + ... + e^{\mathbf{w}_K \cdot \mathbf{x}}} . \tag{54}$$

Things to note:

- The above term sums to 1 (summing over k.)
- For two classes only, we get the same terms as previously, with $\mathbf{w} = \mathbf{w}_1 - \mathbf{w}_2$:

$$p(1|\mathbf{x}) = \frac{e^{\mathbf{w}_1 \cdot \mathbf{x}}}{e^{\mathbf{w}_1 \cdot \mathbf{x}} + e^{\mathbf{w}_2 \cdot \mathbf{x}}} = \frac{1}{1 + e^{-(\mathbf{w}_1 - \mathbf{w}_2) \cdot \mathbf{x}}} = \sigma((\mathbf{w}_1 - \mathbf{w}_2) \cdot \mathbf{x}) \tag{55}$$

$$p(2|\mathbf{x}) = \frac{e^{\mathbf{w}_2 \cdot \mathbf{x}}}{e^{\mathbf{w}_1 \cdot \mathbf{x}} + e^{\mathbf{w}_2 \cdot \mathbf{x}}} = \frac{1}{1 + e^{(\mathbf{w}_1 - \mathbf{w}_2) \cdot \mathbf{x}}} = \sigma(-(\mathbf{w}_1 - \mathbf{w}_2) \cdot \mathbf{x}) \tag{56}$$

Logistic Regression with Multiple Classes (2)

Conditional probabilities:

$$p(k|\mathbf{x}) = \frac{e^{\mathbf{w}_k \cdot \mathbf{x}}}{e^{\mathbf{w}_1 \cdot \mathbf{x}} + e^{\mathbf{w}_2 \cdot \mathbf{x}} + ... + e^{\mathbf{w}_K \cdot \mathbf{x}}}. \tag{57}$$

The training set $\mathcal{T} = \{(\mathbf{x}_1, k_1), (\mathbf{x}_2, k_2), ..., (\mathbf{x}_N, k_N)\}$ (as before);
The set of parameters to find: $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_K]$;

The conditional log likelihood $l'(\mathbf{W})$:

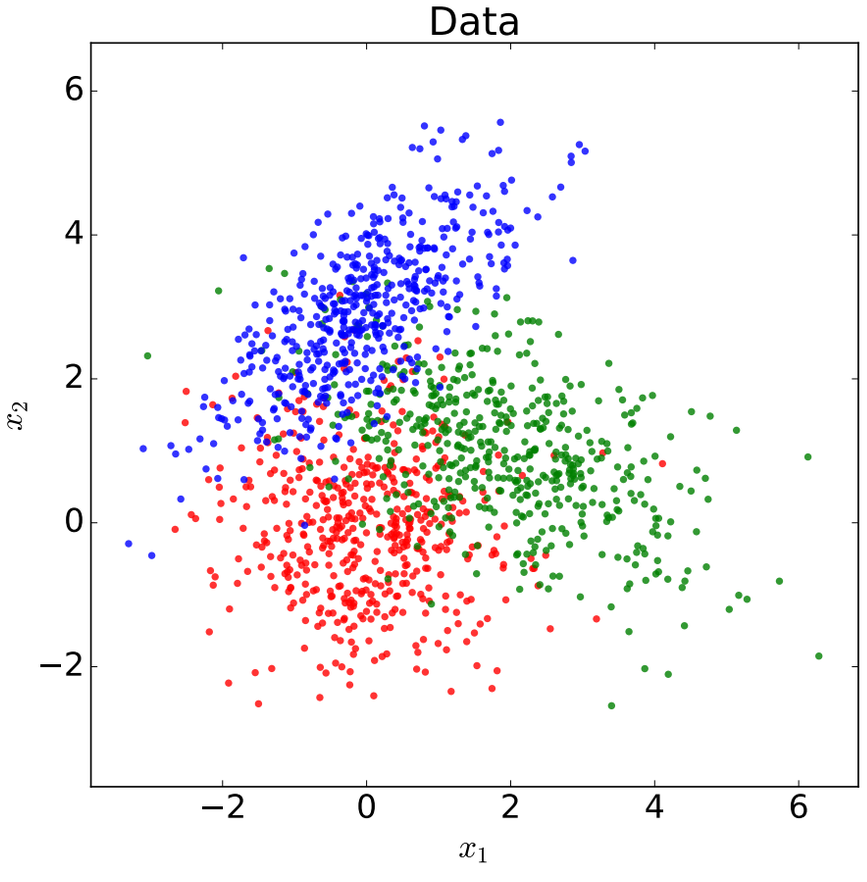
$$l'(\mathbf{W}) = \sum_{(\mathbf{x},k)\in\mathcal{T}} \ln p(k|\mathbf{x}) = \sum_{(\mathbf{x},k)\in\mathcal{T}} \mathbf{w}_k \cdot \mathbf{x} - \sum_{(\mathbf{x},k)\in\mathcal{T}} \ln(e^{\mathbf{w}_1 \cdot \mathbf{x}} + e^{\mathbf{w}_2 \cdot \mathbf{x}} + ... + e^{\mathbf{w}_K \cdot \mathbf{x}}) \tag{58}$$

Optimal parameters $\mathbf{W}^*$:

$$\mathbf{W}^* = \underset{\mathbf{W}}{\text{argmax}}\ l'(\mathbf{W}) = \underset{\mathbf{W}}{\text{argmin}}\ E(\mathbf{W}) \tag{59}$$

($E(\mathbf{W})$ is cross entropy, as before)

Logistic Regression with Multiple Classes, Example (1)



Each class: 500 samples from multivariate normal distribution.

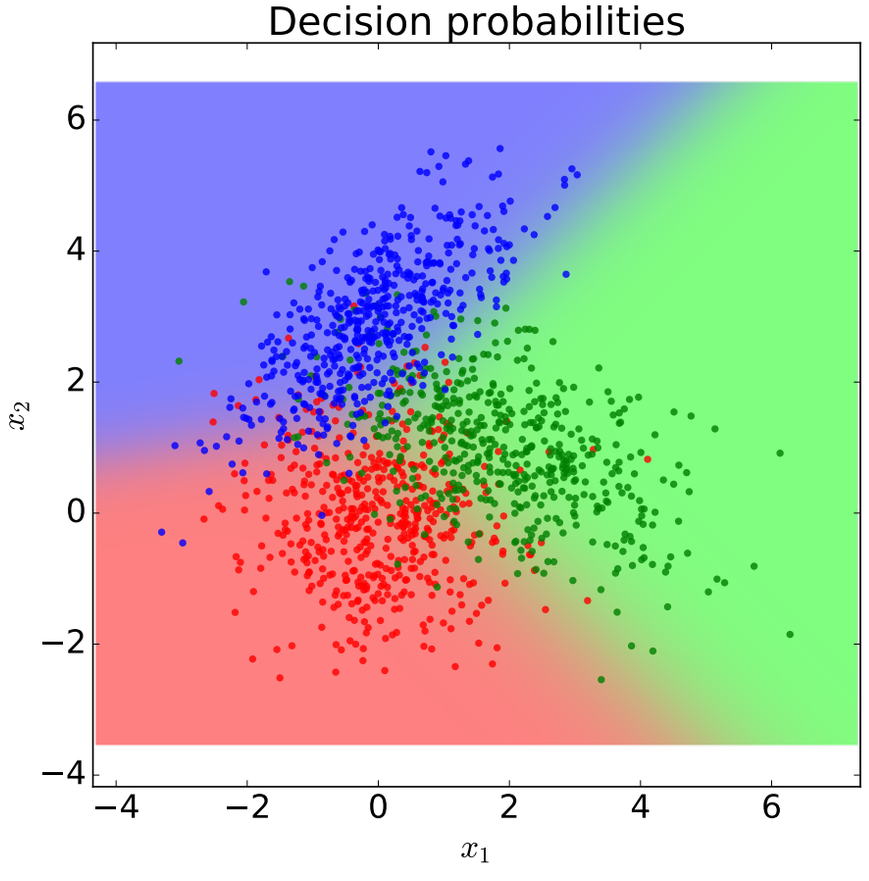
Centers:

$(0, 0)$, $(2, 1)$, $(0, 3)$.

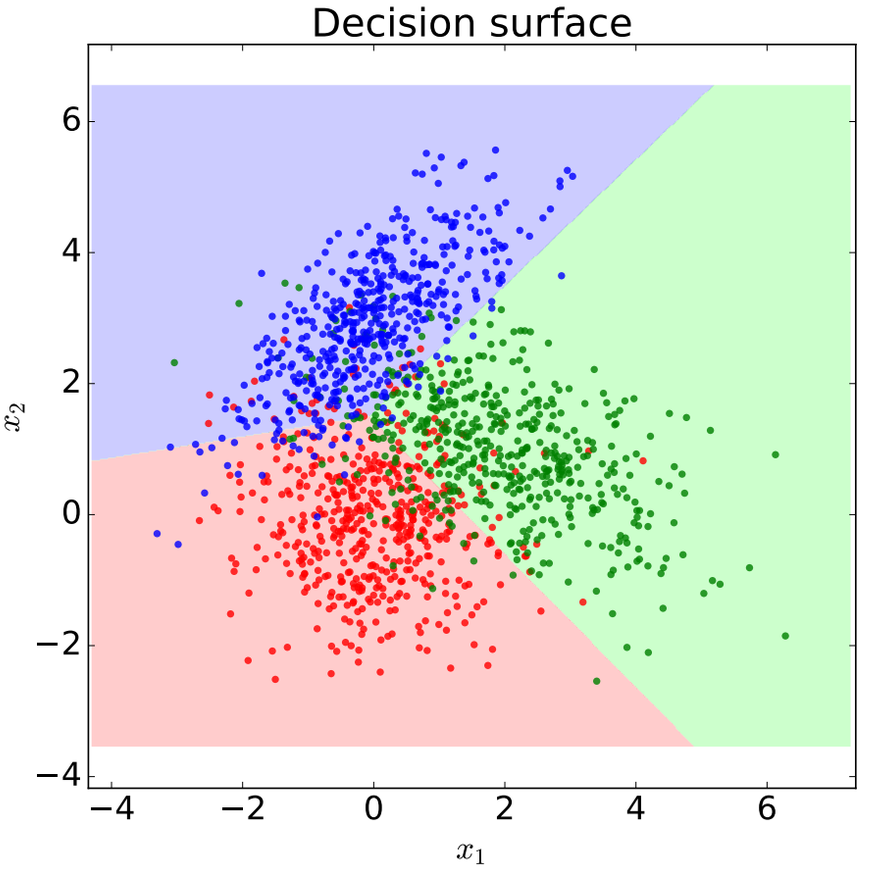
Cov. matrices:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 2 & -0.7 \\ -0.7 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}.$$

Logistic Regression with Multiple Classes, Example (2)



Conditional probabilities $p(k|\mathbf{x})$ (coded by color intensity)



Decisions, coded by class' colors. Note the linearity of decision boundaries.

Notes

- Logistic regression is a discriminative classifier.
- Output of logistic regression are posterior probabilities and it is trained using gradient optimization of cross entropy objective. We will talk later about Neural Nets (NNs) and see the similarities.
- The learnt weights are interpretable; they tell how their corresponding features contribute to the decision.