Logistic Regression

Lecturer:
Jiří Matas

Authors:
Ondřej Drbohlav, Jiří Matas

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

Lecture date: 30.10.2015 & 02.11.2015

Last update: 02.11.2015, 2pm

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 (1)

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

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

For a binary (2-class) classification, this can equivalently be expressed as follows. Define the **log odds** $a(x)$ as the log of the ratio of the posteriors:

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

Then,

$$a(x) > 0 \qquad \Rightarrow \qquad q(x) = 1\,, \tag{3}$$

$$a(x) < 0 \qquad \Rightarrow \qquad q(x) = 2\,. \tag{4}$$

Logistic Regression, Motivation (2)

Now, let us check the functional form of the log odds in the following problems:

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

Normal distributions with equal variances

$$p(x|1) = \mathcal{N}(x|\mu_1, \sigma) \tag{5}$$

$$p(x|2) = \mathcal{N}(x|\mu_2, \sigma) \tag{6}$$

$$a(x) = \ln \frac{p(1|x)}{p(2|x)} = \ln \frac{p(x|1)\,p(1)}{p(x|2)\,p(2)} = \left\{ -\frac{1}{2\sigma^2}\left((x-\mu_1)^2 - (x-\mu_2)^2\right) \right\} + \ln \frac{p(1)}{p(2)} \tag{7}$$

$$= \frac{1}{\sigma^2}(\mu_1 - \mu_2)x + \text{const} = w_1 x + w_0 \qquad (w_1, w_0 \in \mathbb{R}) \tag{8}$$

Logistic Regression, Motivation (3)

Independent features with binary outcomes (D = number of features)

$$p(x|1) = \prod_{i=1}^{D} \pi_i^{x_i}(1-\pi_i)^{1-x_i} \tag{9}$$

$$p(x|2) = \prod_{i=1}^{D} \kappa_i^{x_i}(1-\kappa_i)^{1-x_i} \qquad (\pi_i, \kappa_i \in \mathbb{R}, x_i \in \{0,1\}) \tag{10}$$

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.

The log odds are:

$$a(x) = \sum_{i=1}^{D} \{x_i \ln \pi_i + (1-x_i)\ln(1-\pi_i) - x_i \ln \kappa_i - (1-x_i)\ln(1-\kappa_i)\} + \ln \frac{p(1)}{p(2)} \tag{11}$$

$$= w \cdot x + w_0 \qquad (w \in \mathbb{R}^D, w_0 \in \mathbb{R}) \tag{12}$$

Logistic Regression, Motivation (4)

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 "histogram" $\{x_i, i = 1, 2, .., D\}$ are

$$p(x|1) = \frac{n!}{\prod_{i=1}^{D} x_i!} \prod_{i=1}^{D} \pi_i^{x_i} \tag{13}$$

$$p(x|2) = \frac{n!}{\prod_{i=1}^{D} x_i!} \prod_{i=1}^{D} \kappa_i^{x_i} \qquad (x_i \in \mathbb{N}_0, \pi_i, \kappa_i \in \mathbb{R}, \sum_{i=1}^{D} x_i = n) \tag{14}$$

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

Summary

In many real-world problems, the log odds $a(x)$ are a linear function of the observations x.

Logistic Regression, Model

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

$$a(x) = \ln \frac{p(1|x)}{p(2|x)} = w \cdot x + w_0 \,, \qquad w = (w_1, w_2, ..., w_D) \in \mathbb{R}^D \,, \quad w_0 \in \mathbb{R} \,, \tag{15}$$

where w_0 is the bias term.

Let us rewrite this as

$$a(x) = w \cdot x + w_0 \cdot \underset{\substack{\uparrow \\ x_0}}{1} = [w_0, w_1, w_2, ..., w_D] \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_D \end{bmatrix} = w' \cdot x' . \tag{16}$$

Note: From now on, we will drop the dash sign and write again only 'x' or '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 $a(x)$ and the posterior probabilities $p(1|x)$ and $p(2|x)$.

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

$$a(x) = \ln \frac{p(1|x)}{p(2|x)} = w \cdot x \,. \tag{17}$$

From this, it follows that

$$\frac{p(1|x)}{p(2|x)} = \exp(a(x)) \qquad \Rightarrow \qquad p(2|x) = \frac{1}{1+\exp(a(x))} = \frac{1}{1+e^{w \cdot x}} \tag{18}$$

and

$$p(1|x) = 1 - p(2|x) = \frac{\exp(a(x))}{1+\exp(a(x))} = \frac{1}{1+\exp(-a(x))} = \frac{1}{1+e^{-w \cdot x}} \,. \tag{19}$$

Logistic Regression, Log Odds and Posteriors (2)

Again,

$$a(x) = \ln \frac{p(1|x)}{p(2|x)} = w \cdot x \tag{20}$$

$$p(1|x) = \frac{1}{1 + e^{-w \cdot x}} = \sigma(w \cdot x) \tag{21}$$

$$p(2|x) = \frac{1}{1 + e^{w \cdot x}} = \sigma(-w \cdot x) \tag{22}$$

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 (21, 22) as

$$p(k|x) = \frac{1}{1 + e^{-kw \cdot x}}, \qquad k \in \{-1, 1\} \tag{23}$$

Finding w: Objective $E(w)$

$$p(k|x) = \frac{1}{1+e^{-kw\cdot x}}\,, \qquad k \in \{-1, 1\} \tag{24}$$

How do we find w?

We adopt the Maximum Likelihood approach for finding w. Let us have the training set $\mathcal{T} = \{(x_1, k_1), (x_2, k_2), ..., (x_N, k_N)\}$. The optimal w^* is the one which maximizes the conditional log likelihood $l(w)$:

$$l(w) = \sum_{(x,k)\in\mathcal{T}} \ln p(k|x) = -\sum_{(x,k)\in\mathcal{T}} \ln(1+e^{-kw\cdot x}) \qquad \text{(conditional log likelihood)} \tag{25}$$

$$w^* = \operatorname*{argmax}_w l(w) \qquad \text{(optimal } w^*\text{)} \tag{26}$$

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

Finding w: Gradient of $E(w)$

$$E(w) = -\sum_{(x,k)\in\mathcal{T}} \ln p(k|x) = \sum_{(x,k)\in\mathcal{T}} \ln(1+e^{-kw\cdot x}) \tag{27}$$

$$\tag{28}$$

The gradient vector $g(w)$ of E is:

$$g(w) = \frac{\partial E(w)}{\partial w} = \sum_{(x,k)\in\mathcal{T}} \frac{e^{-kw\cdot x}}{1+e^{-kw\cdot x}}(-kx) = -\sum_{(x,k)\in\mathcal{T}} \underbrace{\frac{1}{1+e^{kw\cdot x}}}_{p(-k|x)} kx \tag{29}$$

$$= -\sum_{(x,k)\in\mathcal{T}} (1-p(k|x))\, kx\,. \tag{30}$$

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

Finding w: $E(w)$ is convex

The Hessian matrix $H(w)$ of the objective function E is

$$H(w) = \frac{\partial^2 E(x)}{\partial w^2} = \frac{\partial g(w)}{\partial w} = -\frac{\partial}{\partial w} \sum_{(x,k)\in\mathcal{T}} \frac{1}{1+e^{kw\cdot x}} kx \tag{31}$$

$$= \sum_{(x,k)\in\mathcal{T}} \underbrace{\frac{e^{kw\cdot x}}{(1+e^{kw\cdot x})^2} k^2}_{>0} \; xx^\top = \sum_{(x,k)\in\mathcal{T}} p(-1|x)p(1|x) \; xx^\top \tag{32}$$

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

Note 1. Can you show that $H(w)$ is positive definite?

Note 2. Strictly speaking, $H(w)$ is positive definite if the training set contains more than 1 distinct point (the outer product $xx^\top$ is positive semi-definite only). Due to the fact that the zero-index component x_0 is fixed to 1, x and x' are linearly independent if they are not identical.

Finding w: Gradient Descent

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

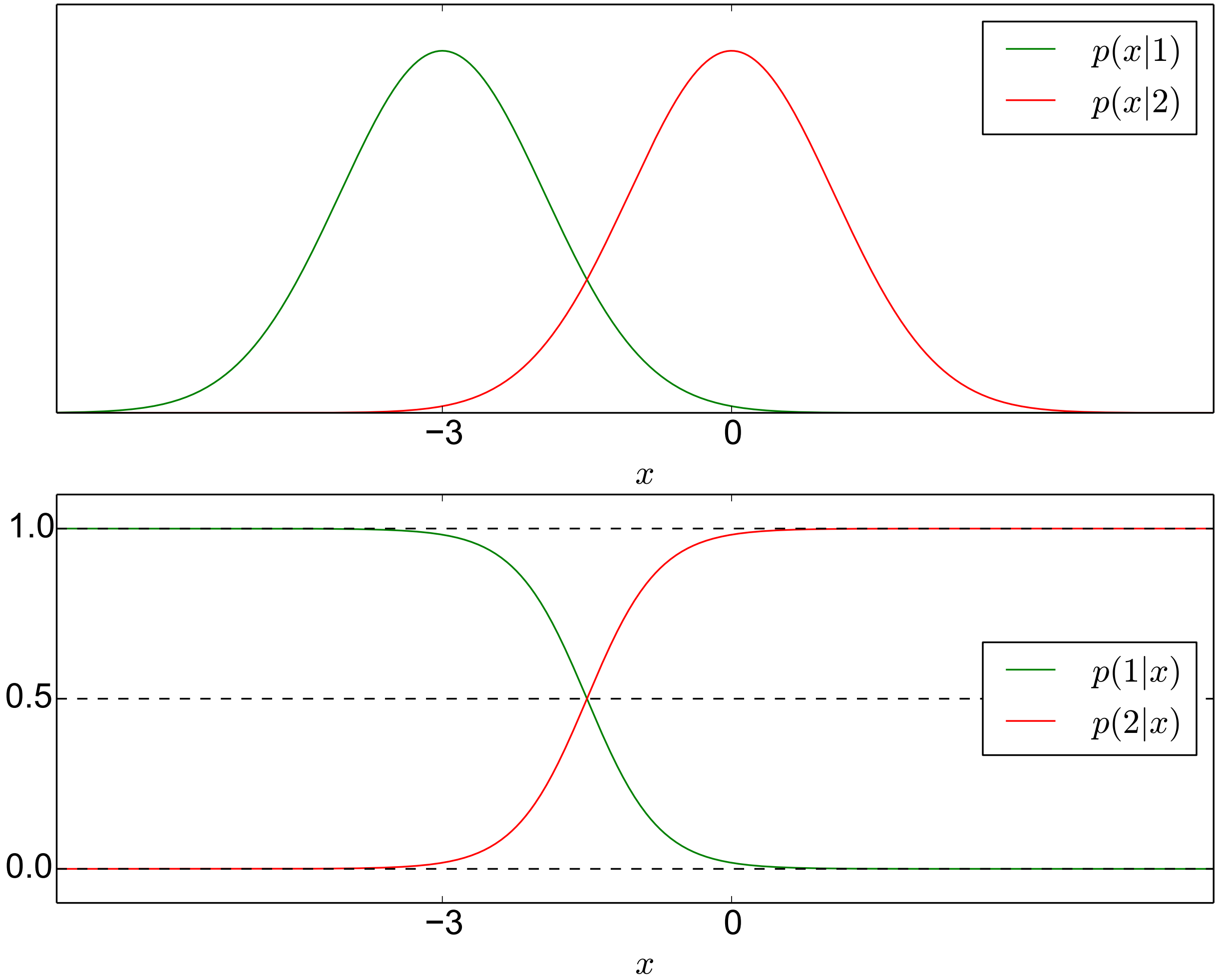
```
# input: x (observations), k (class labels), w_init (initial w)

# 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 $E(w)$ 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.

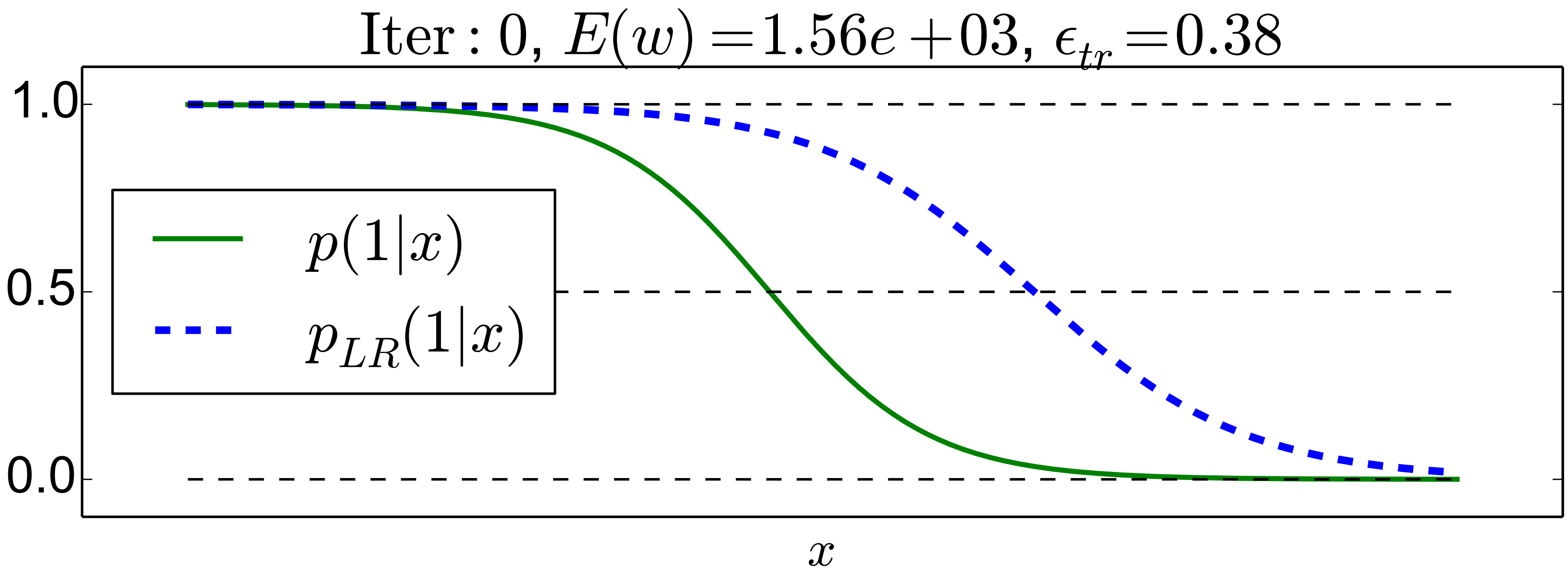
Example 1, Two Normal Distribs with Equal Variance (1)



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

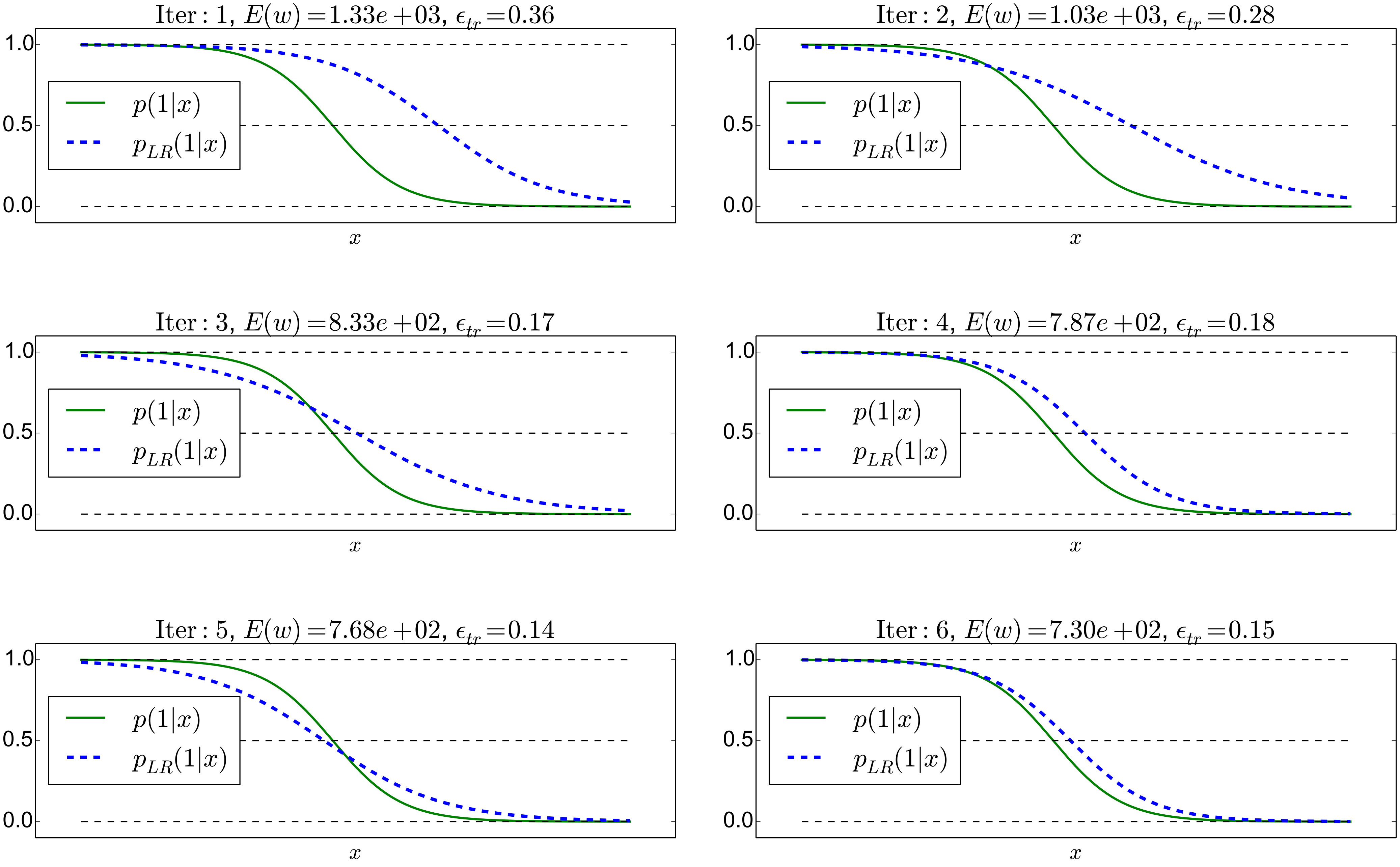
Example 1, Two Normal Distribs with Equal Variance (2)

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

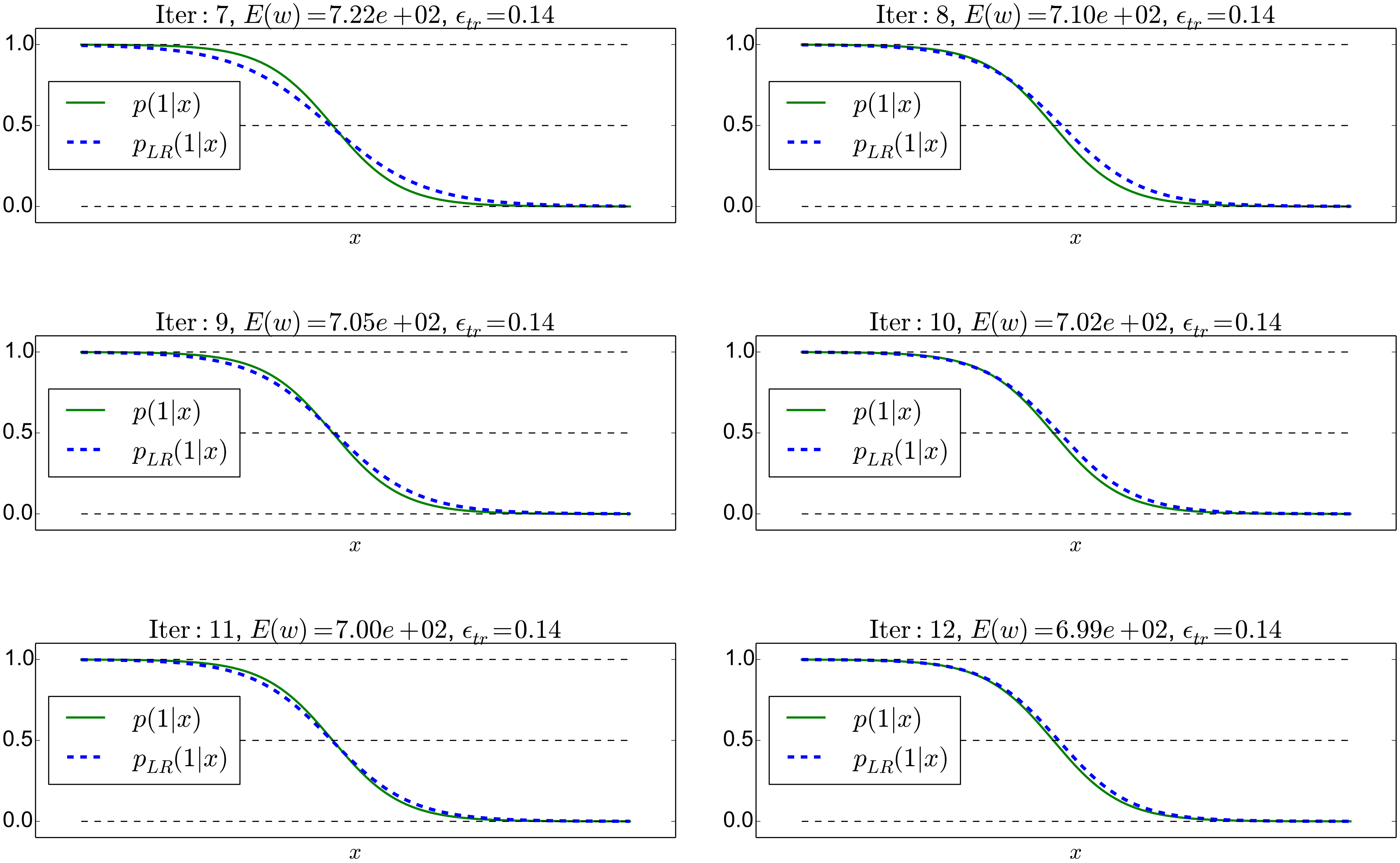


$p(1|x)$: The actual conditional for the 1st class.
$p_{LR}(1\,|\,x)$: The conditional for the 1st class predicted by logistic regression.
$E(w)$: the value of cross entropy.
ϵ_{tr} : the training error (error on the training set.)
(initial $w = [1, -1]^\top$)

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



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



$\mathrm{Iter}: 7,\ E(w) = 7.22e+02,\ \epsilon_{tr} = 0.14$

$\mathrm{Iter}: 8,\ E(w) = 7.10e+02,\ \epsilon_{tr} = 0.14$

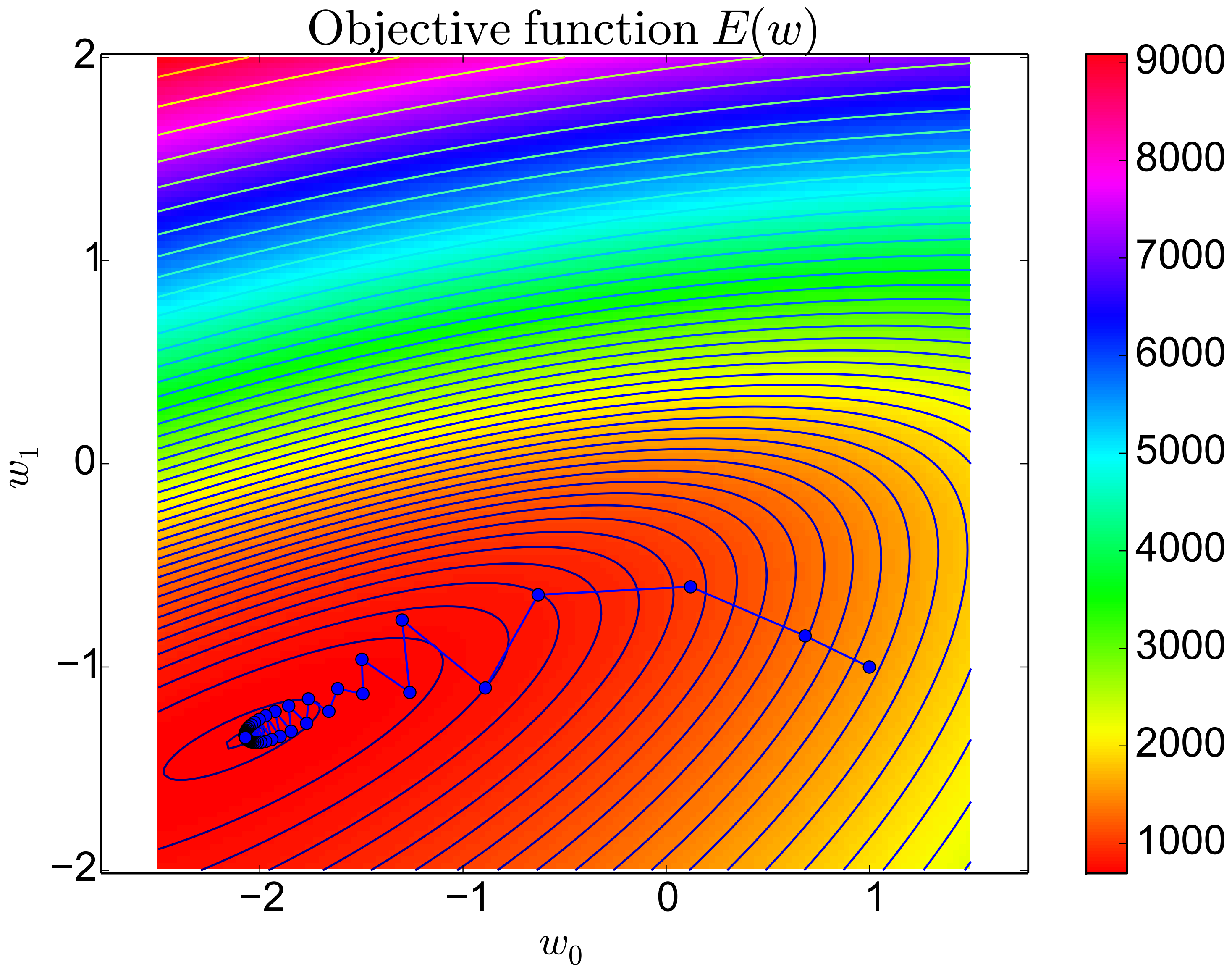
$\mathrm{Iter}: 9,\ E(w) = 7.05e+02,\ \epsilon_{tr} = 0.14$

$\mathrm{Iter}: 10,\ E(w) = 7.02e+02,\ \epsilon_{tr} = 0.14$

$\mathrm{Iter}: 11,\ E(w) = 7.00e+02,\ \epsilon_{tr} = 0.14$

$\mathrm{Iter}: 12,\ E(w) = 6.99e+02,\ \epsilon_{tr} = 0.14$

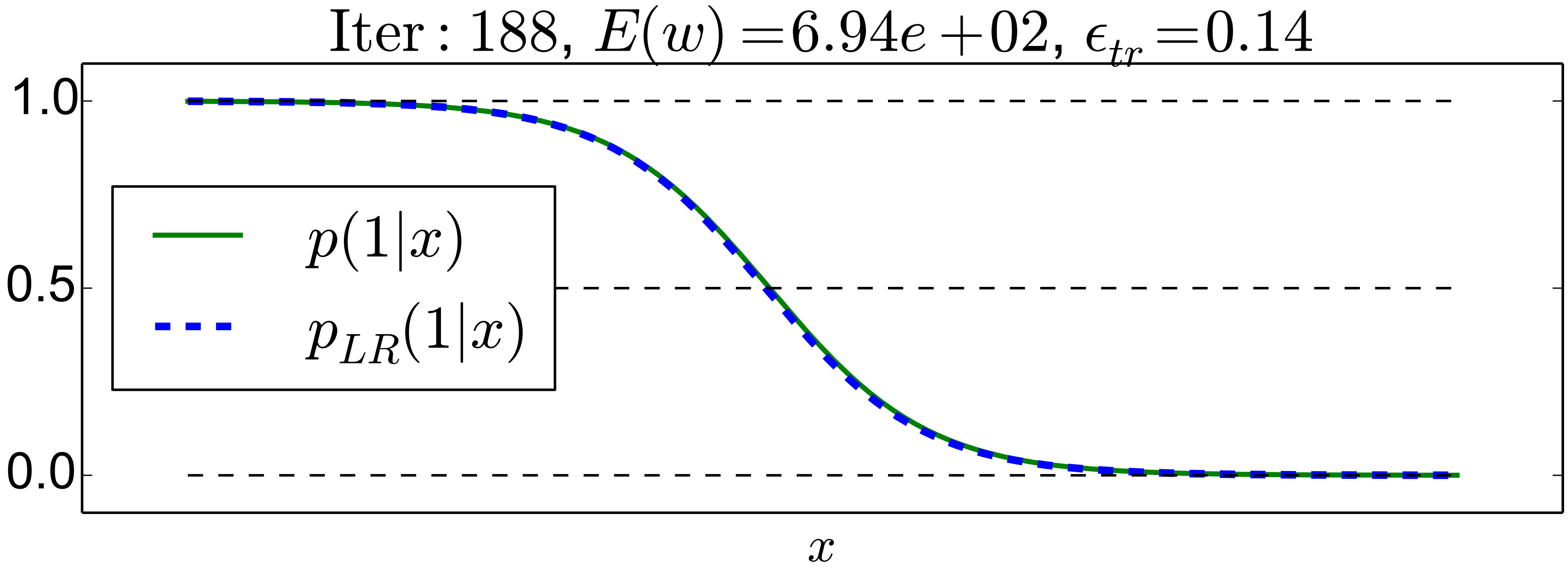
Example 1, Two Normal Distribs with Equal Variance (5)



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

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

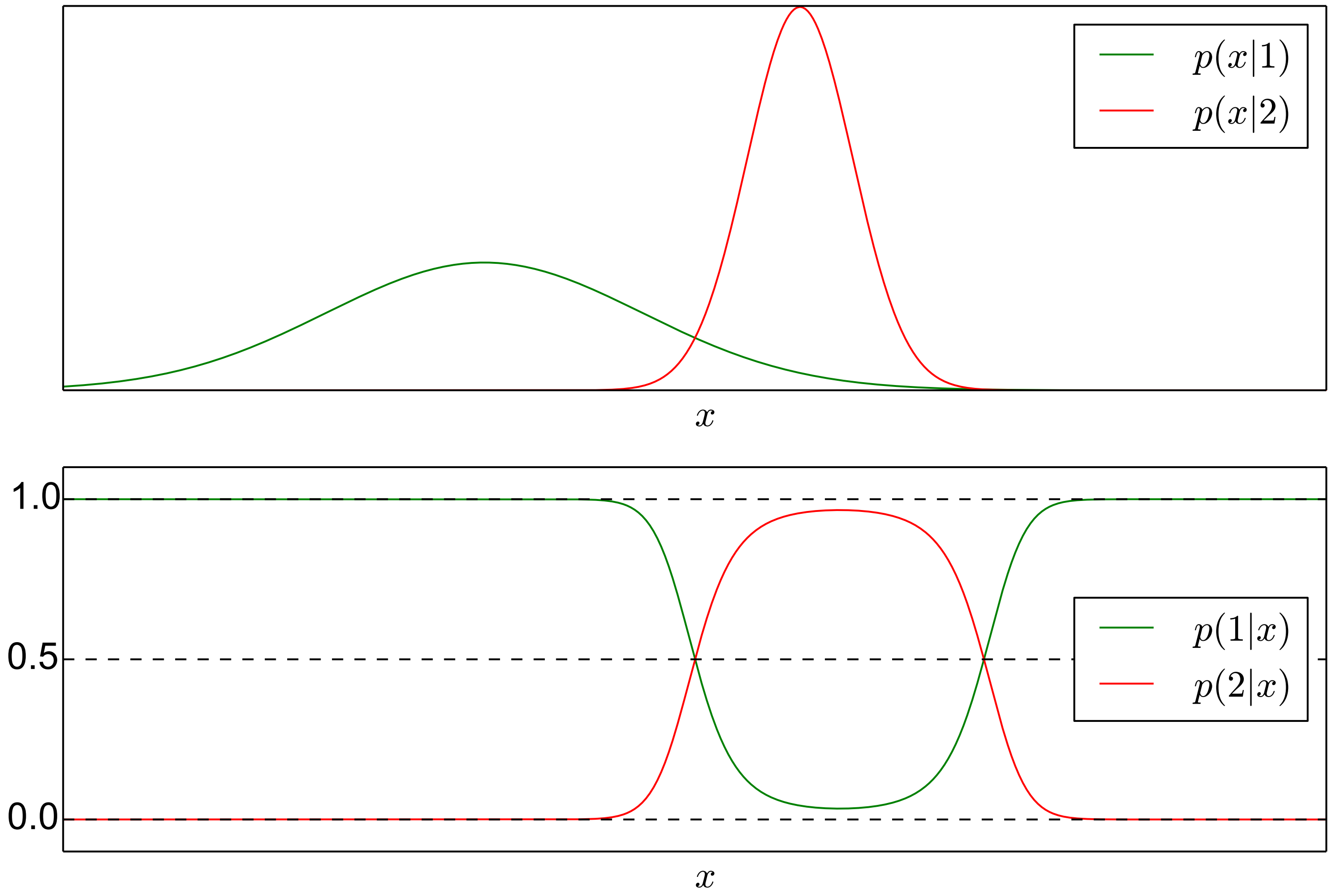
Converged state.
$w = [-2.07, -1.35]^\top$.



Things to note:

- ϵ_{tr} does not monotonically decrease with iterations. $E(w)$ 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.

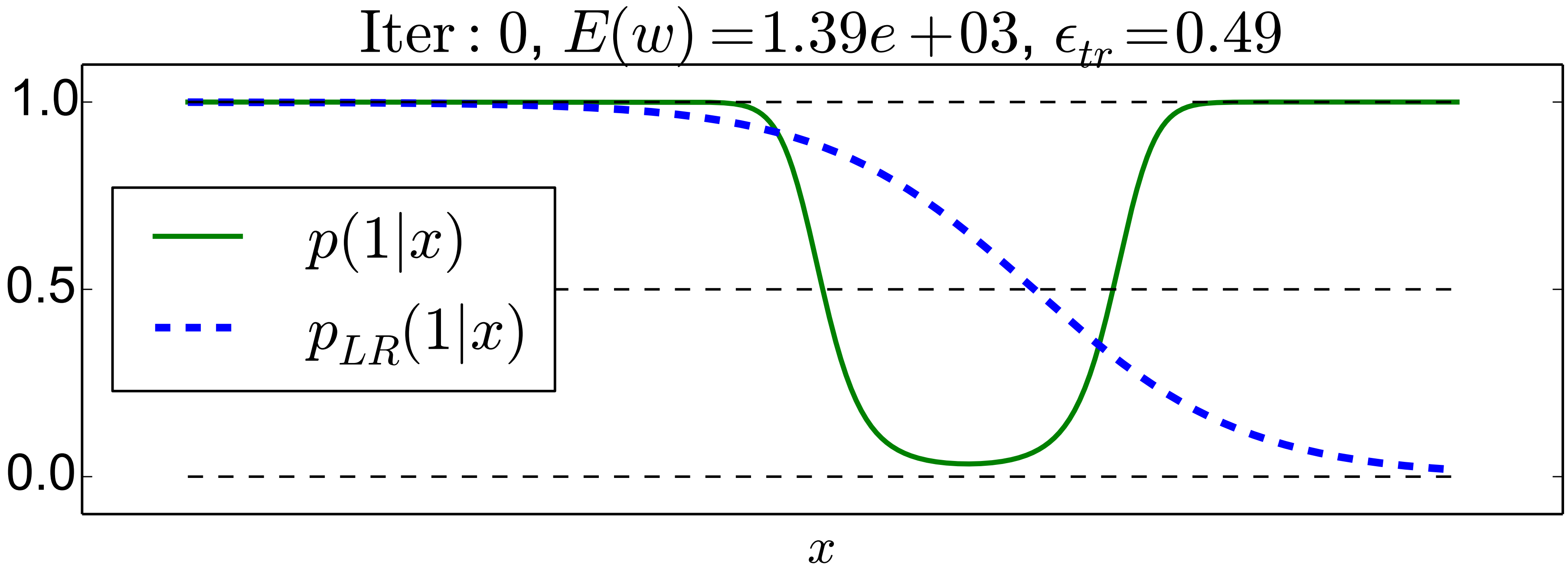
Example 2, Non-Equal Variance but Different Mean (1)



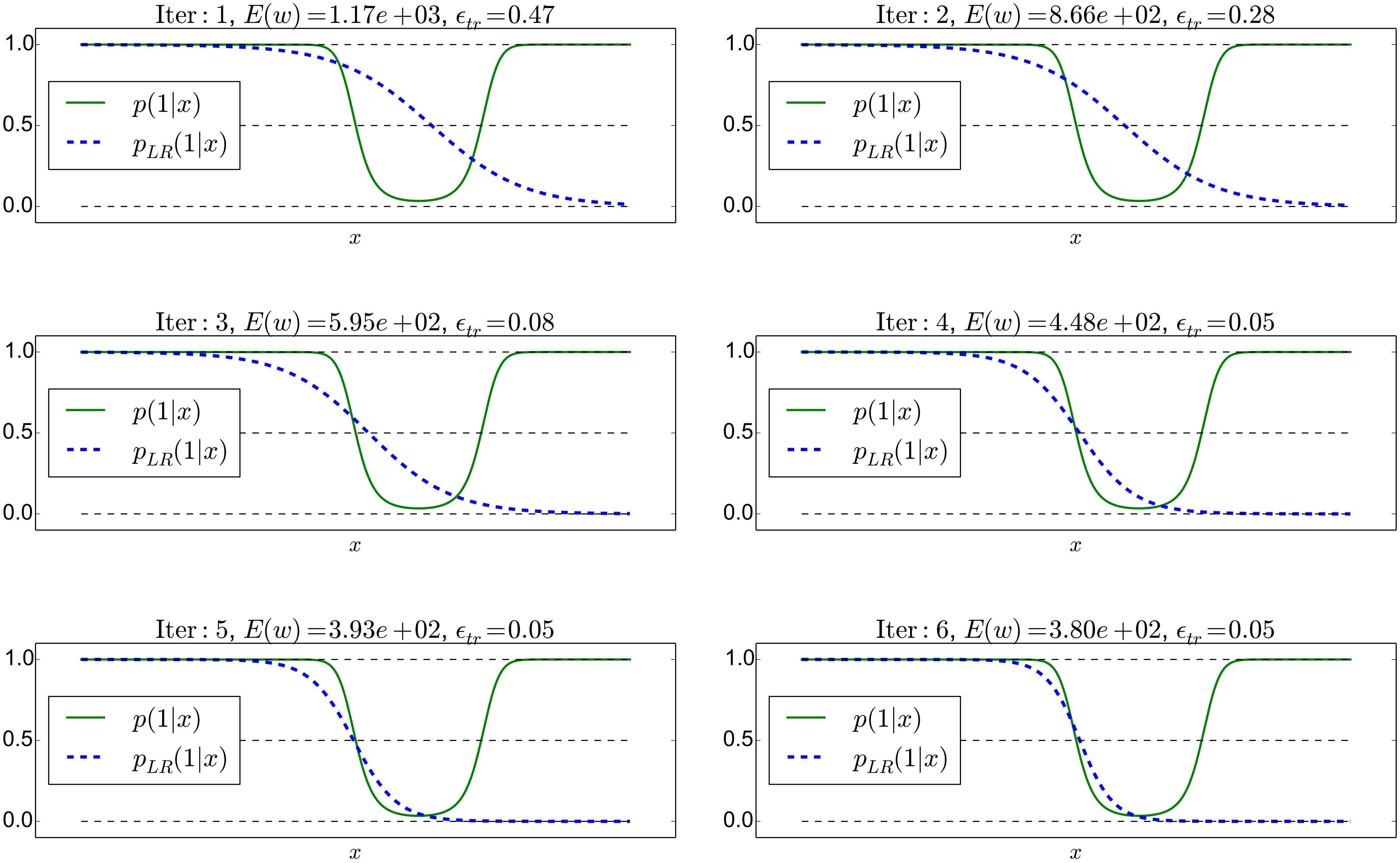
$p(x|1) = \mathcal{N}\left(x|\mu_1 = -3, \sigma_1 = 1.5\right)$
$p(x|2) = \mathcal{N}\left(x|\mu_2 = 0, \sigma_2 = 0.5\right)$
$p(1) = p(2) = 0.5$. **Bayesian error is** $\epsilon_B = 0.057$.

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)



Iter : 1, $E(w) = 1.17e+03$, $\epsilon_{tr} = 0.47$

Iter : 2, $E(w) = 8.66e+02$, $\epsilon_{tr} = 0.28$

Iter : 3, $E(w) = 5.95e+02$, $\epsilon_{tr} = 0.08$

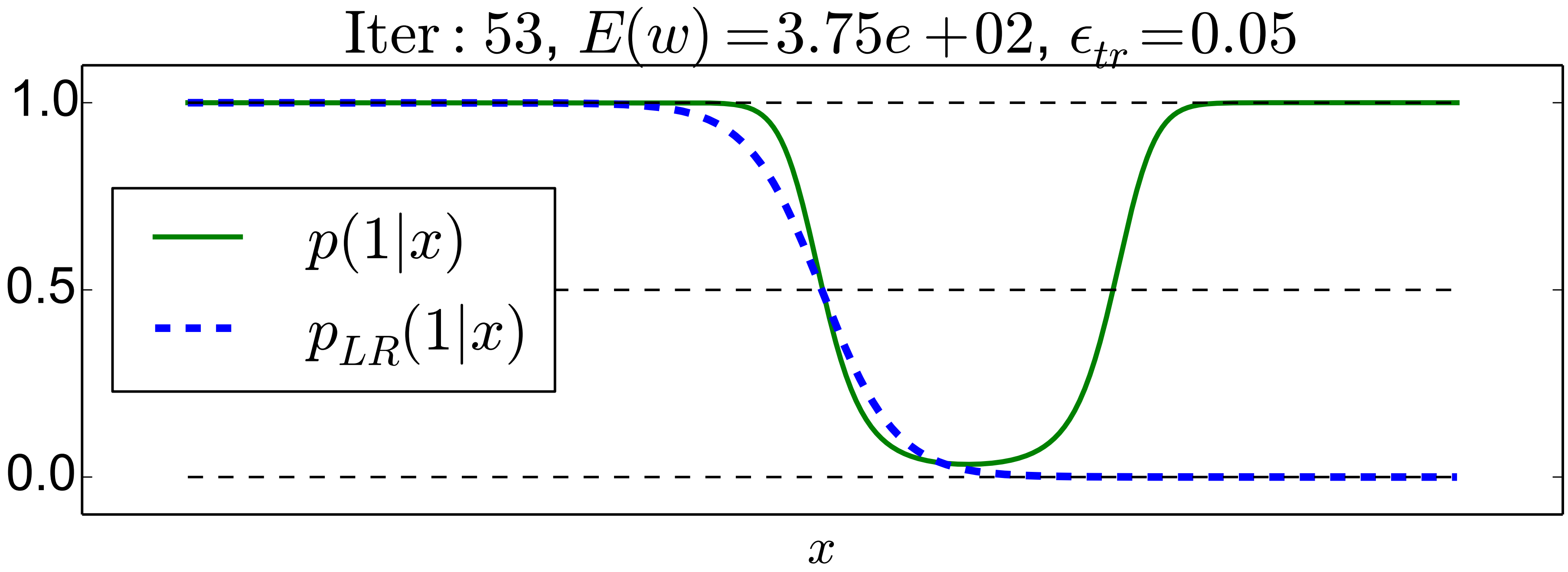
Iter : 4, $E(w) = 4.48e+02$, $\epsilon_{tr} = 0.05$

Iter : 5, $E(w) = 3.93e+02$, $\epsilon_{tr} = 0.05$

Iter : 6, $E(w) = 3.80e+02$, $\epsilon_{tr} = 0.05$

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

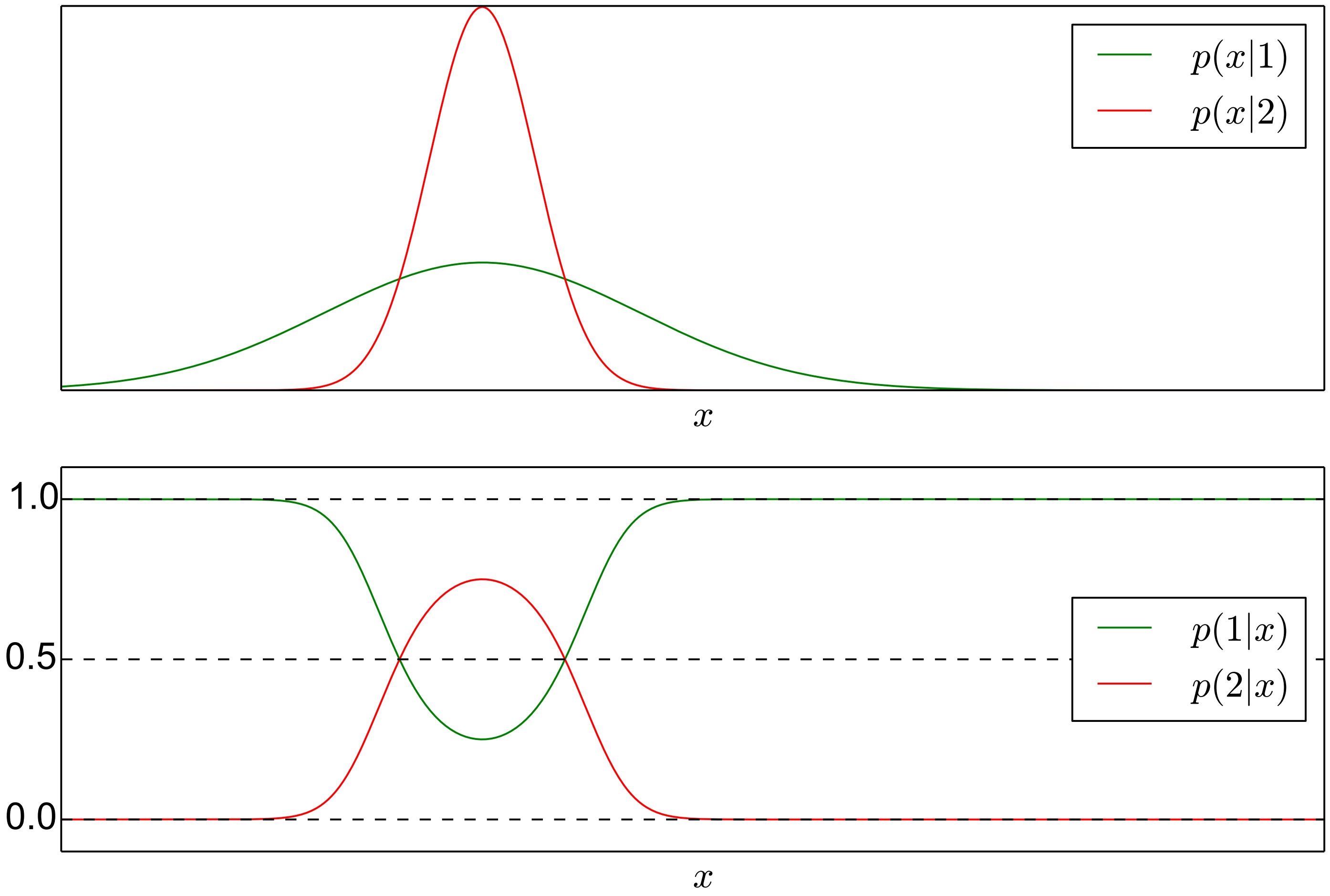
Converged state.
$w = [-2.88, -2.85]^\top$.



Things to note:

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

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

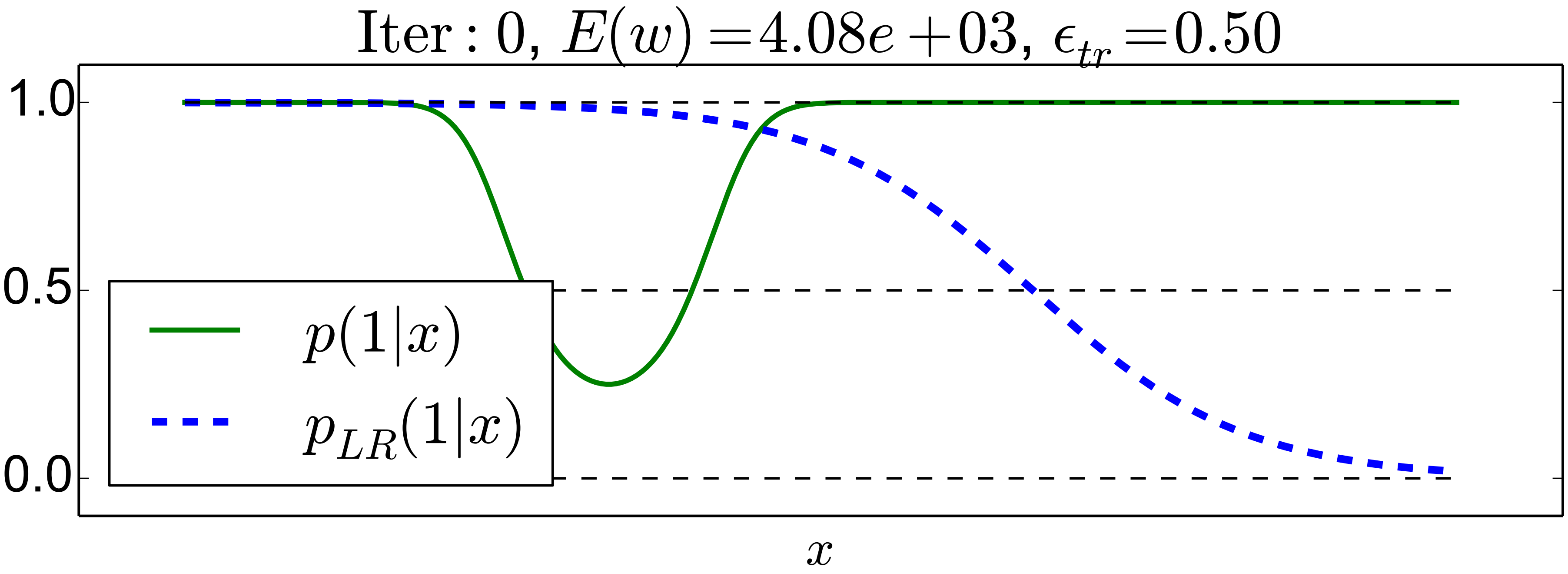


$$p(x|1) = \mathcal{N}\left(x|\mu_1 = -3, \sigma_1 = 1.5\right)$$
$$p(x|2) = \mathcal{N}\left(x|\mu_2 = -3, \sigma_2 = 0.5\right)$$

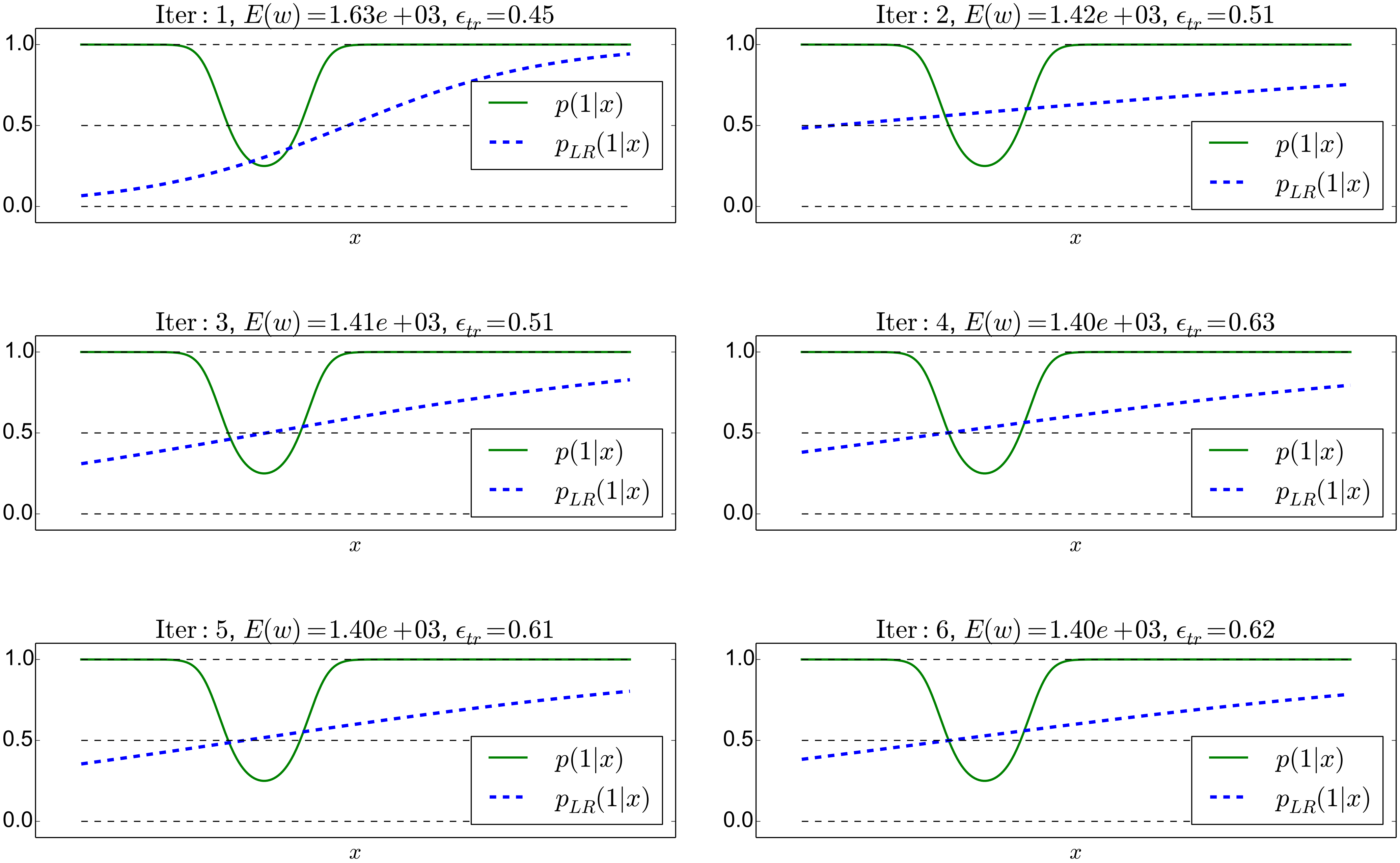
$p(1) = p(2) = 0.5$. **Bayesian error is** $\epsilon_B = 0.26$.

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.

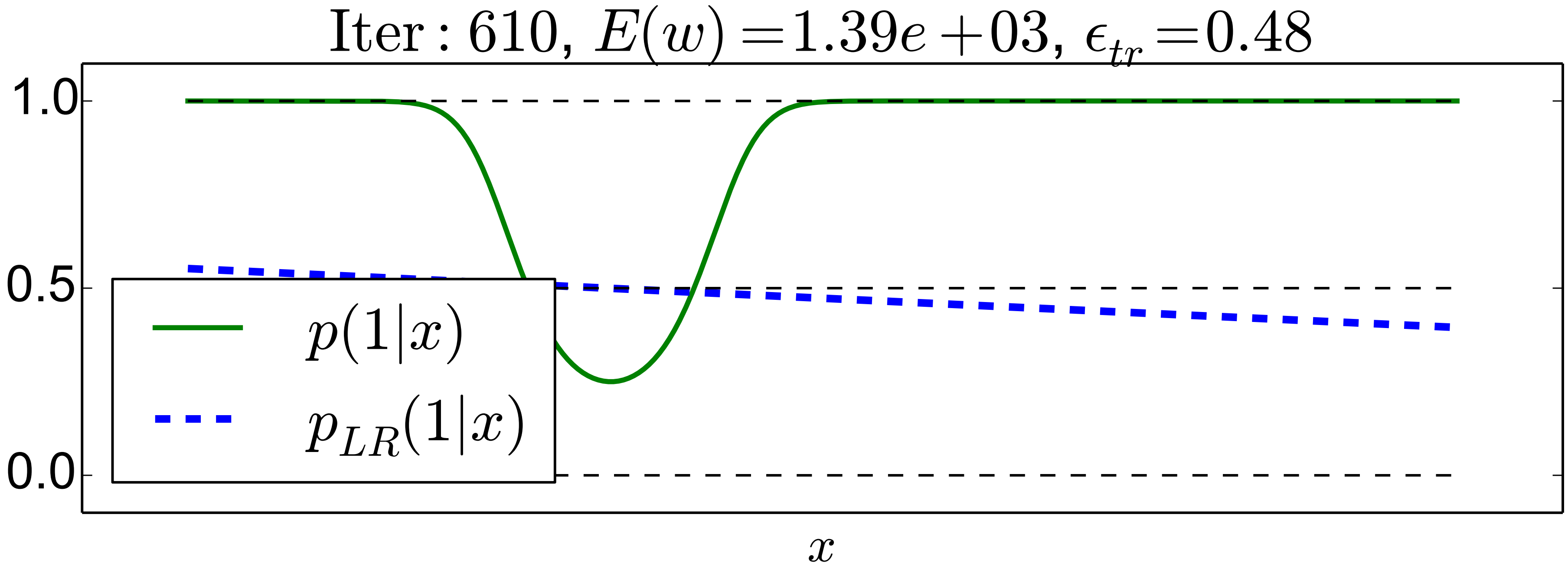


Example 3, Non-Equal Variance and the Same Mean (3)



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

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



Things to note:

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

Notes: Logistic Regression with Multiple classes

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

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

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

$$p(k|x) = \frac{e^{w_k x}}{e^{w_1 x} + e^{w_2 x} + ... + e^{w_K x}} \,. \tag{33}$$

Things to note:

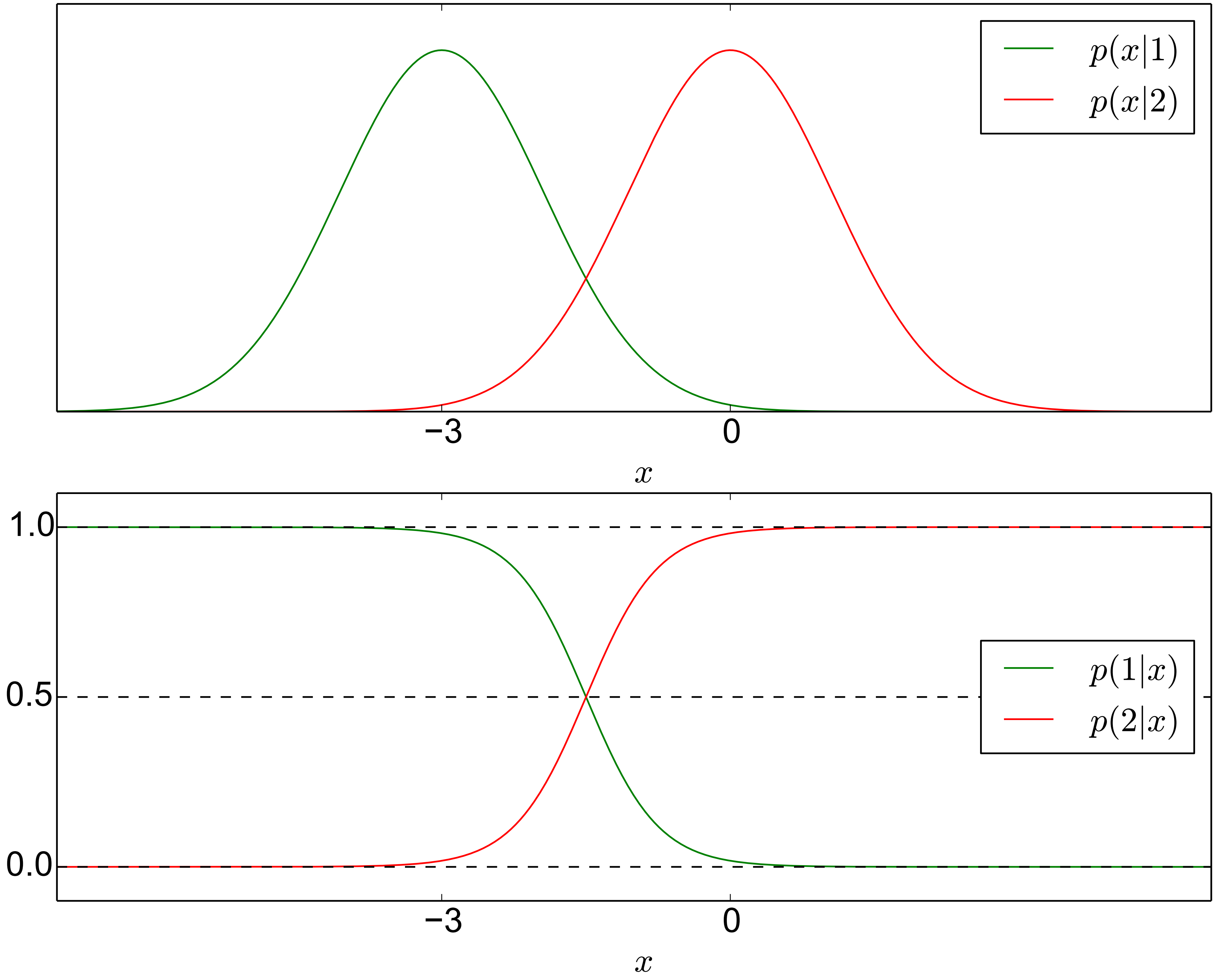
- The above term indeed sums to 1 (summing over k).
- For two classes only, we get the same terms as previously, with $w = w_1 - w_2$:

$$p(1|x) = \frac{e^{w_1 x}}{e^{w_1 x} + e^{w_2 x}} = \frac{1}{1 + e^{-(w_1 - w_2)x}} = \sigma((w_1 - w_2)x) \tag{34}$$

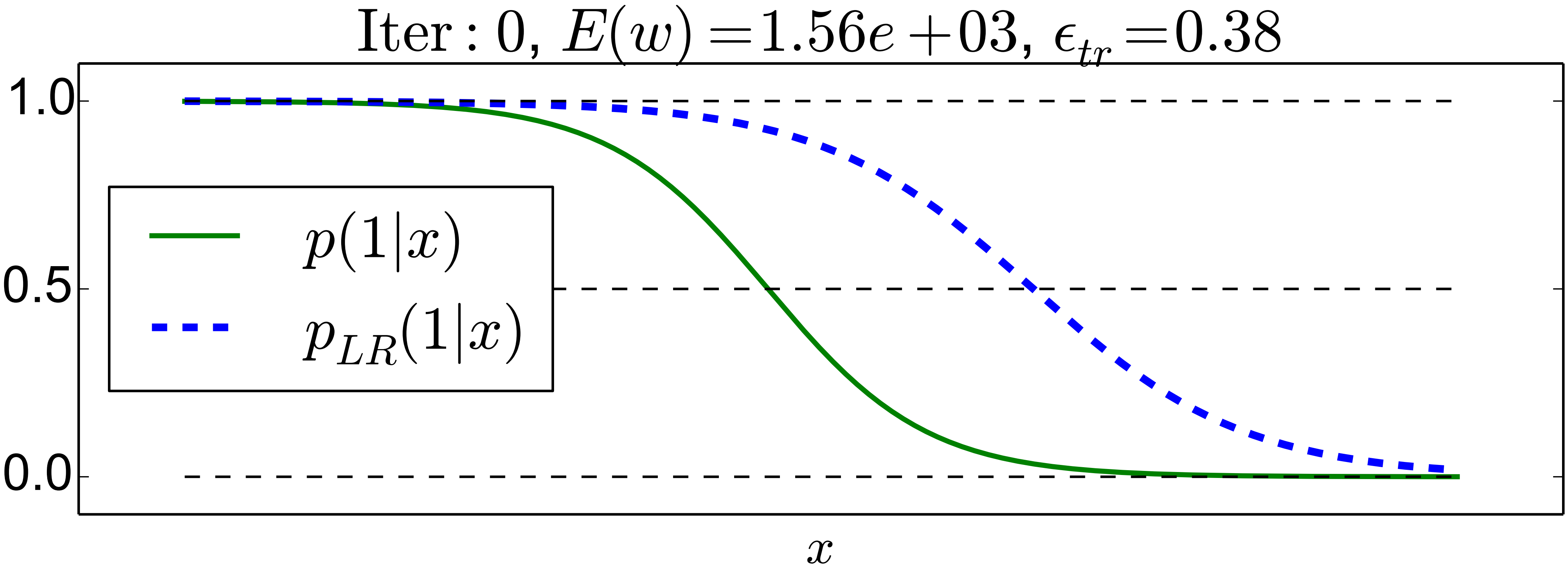
$$p(2|x) = \frac{e^{w_2 x}}{e^{w_1 x} + e^{w_2 x}} = \frac{1}{1 + e^{-(w_1 - w_2)x}} = \sigma(-(w_1 - w_2)\,x) \tag{35}$$

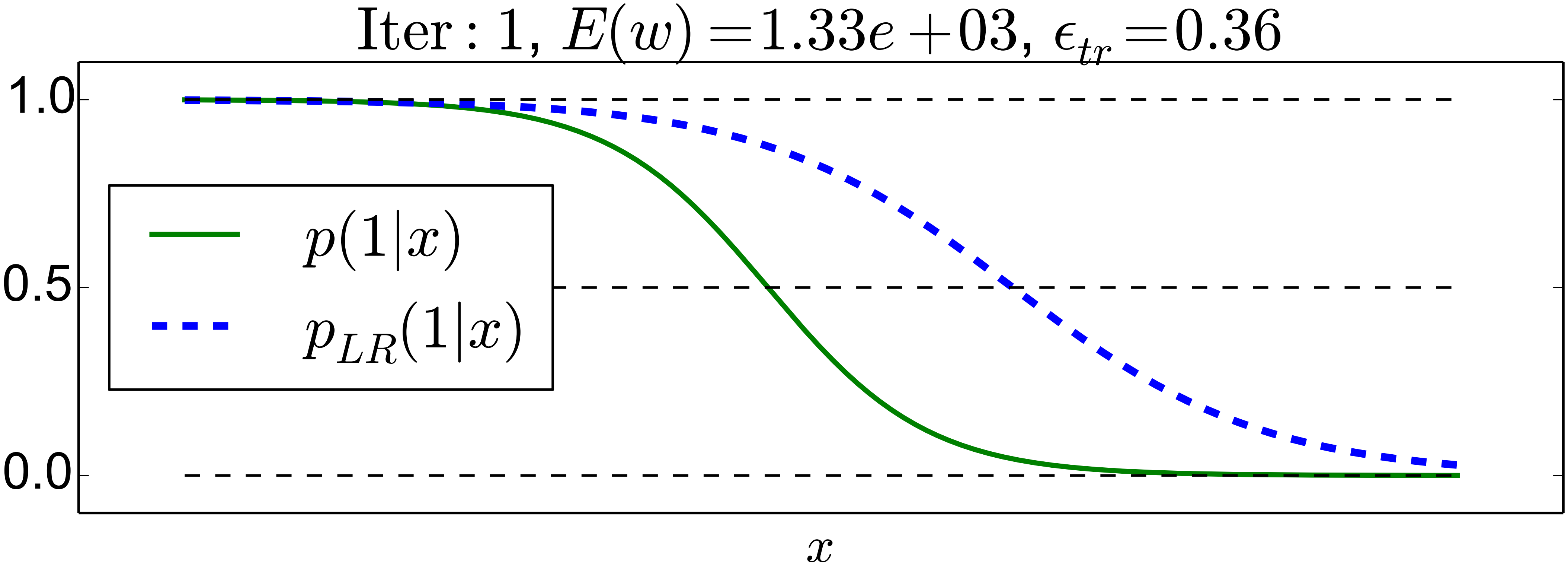
(36)

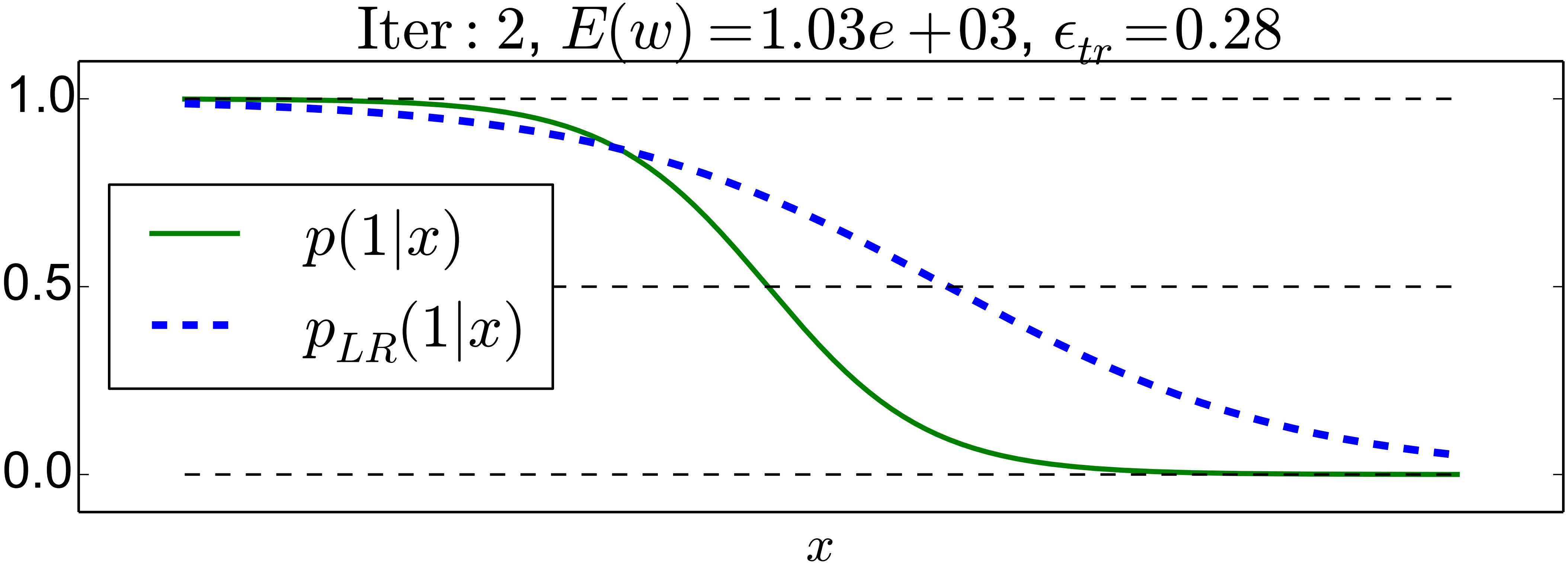
m
p

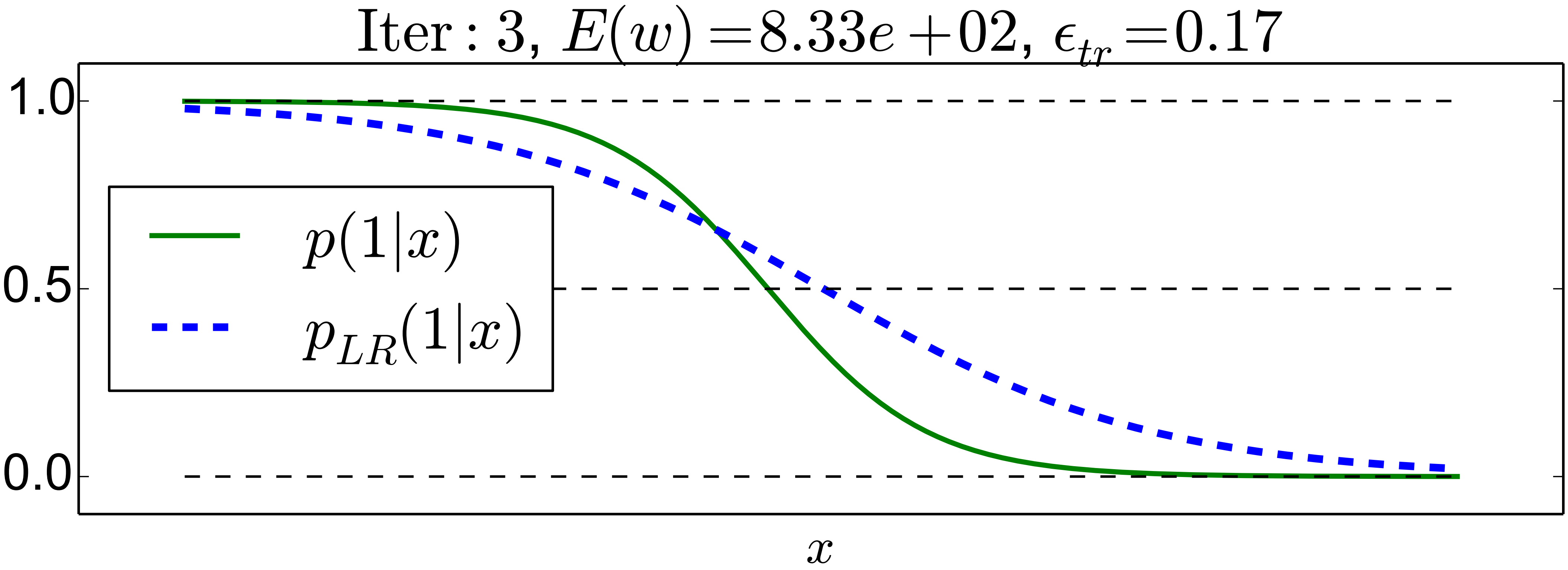


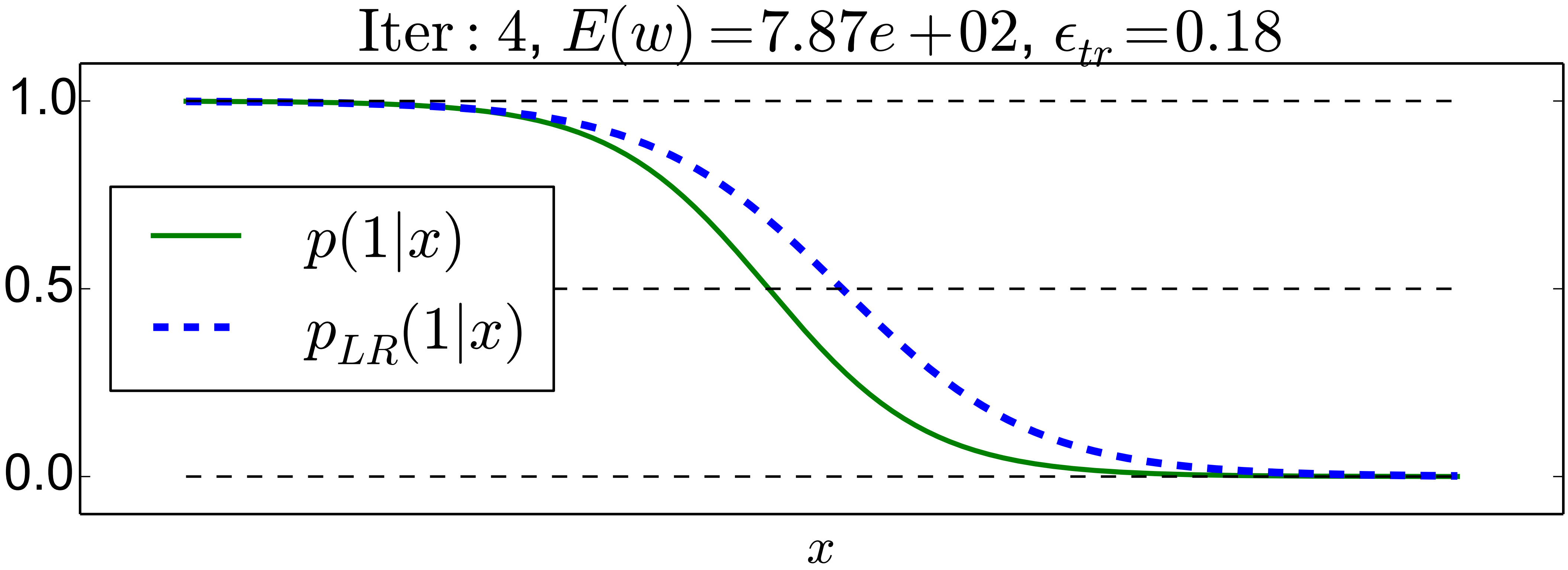
$p(x|1)$
$p(x|2)$
-3
0
x
1.0
0.5
0.0
$p(1|x)$
$p(2|x)$
-3
0
x


$\mathrm{Iter}: 0,\ E(w) = 1.56e+03,\ \epsilon_{tr} = 0.38$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x

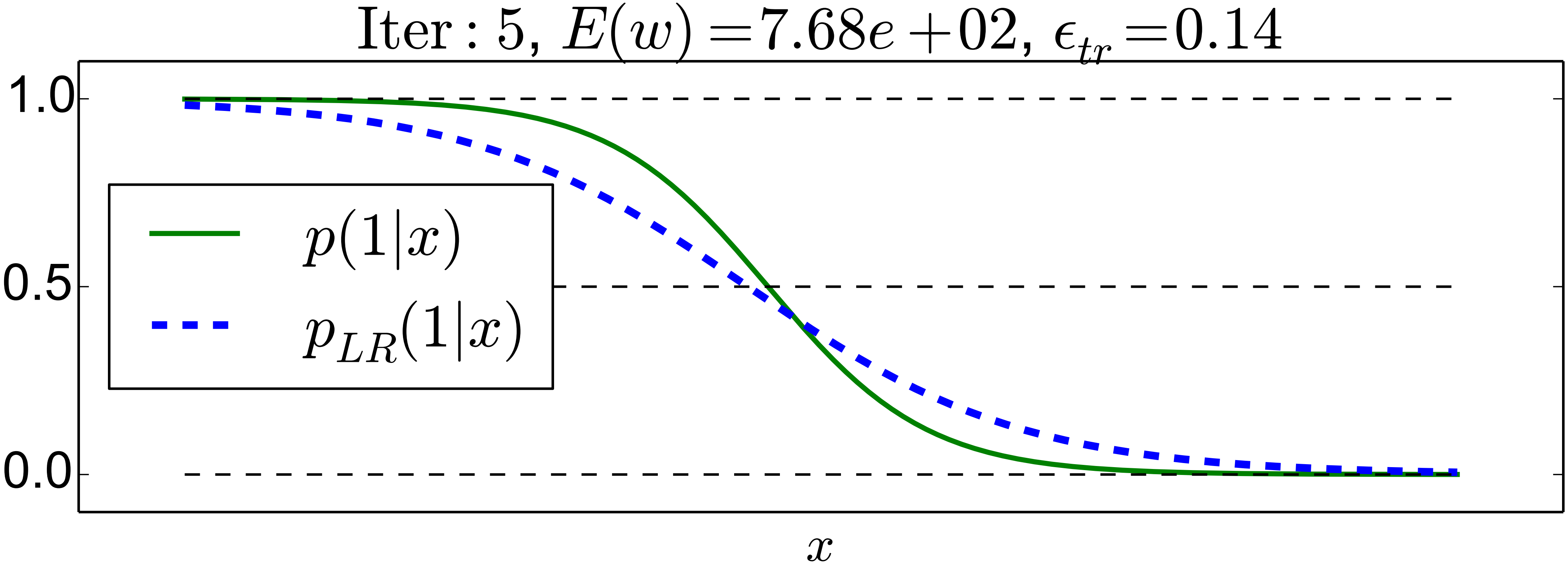

$\mathrm{Iter: 1},\ E(w) = 1.33e+03,\ \epsilon_{tr} = 0.36$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x

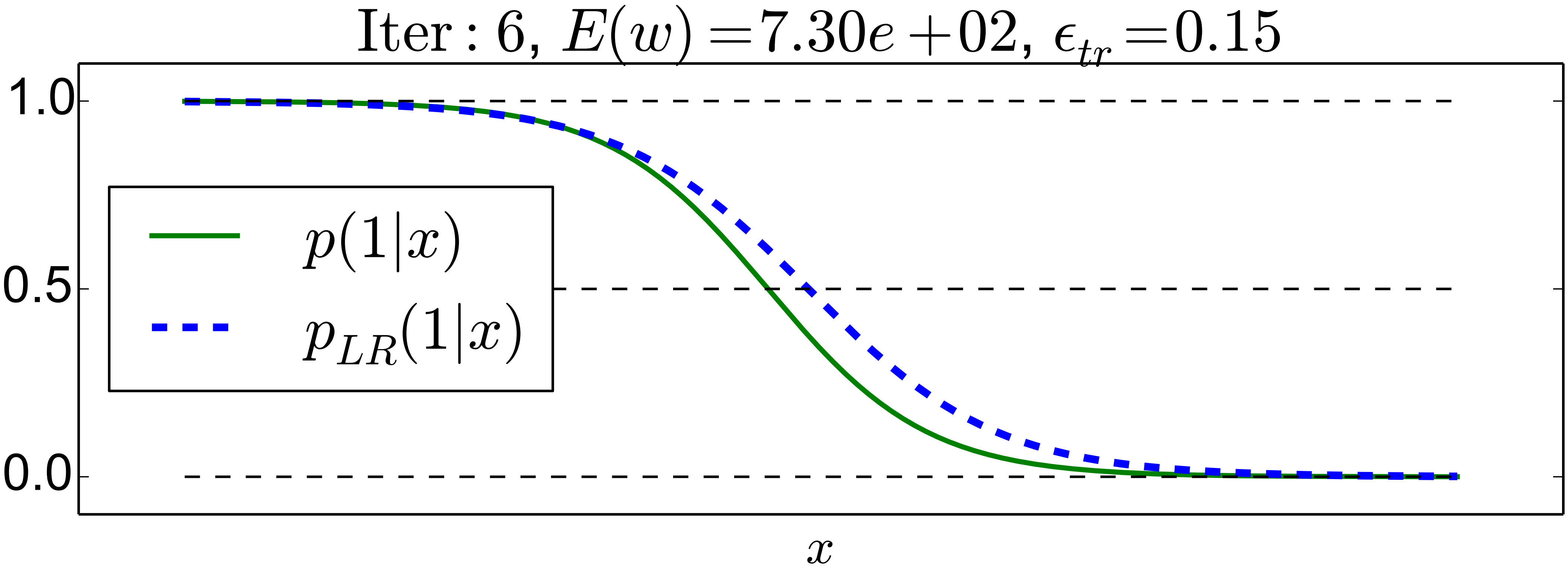

$\mathrm{Iter}: 2,\ E(w) = 1.03e+03,\ \epsilon_{tr} = 0.28$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x

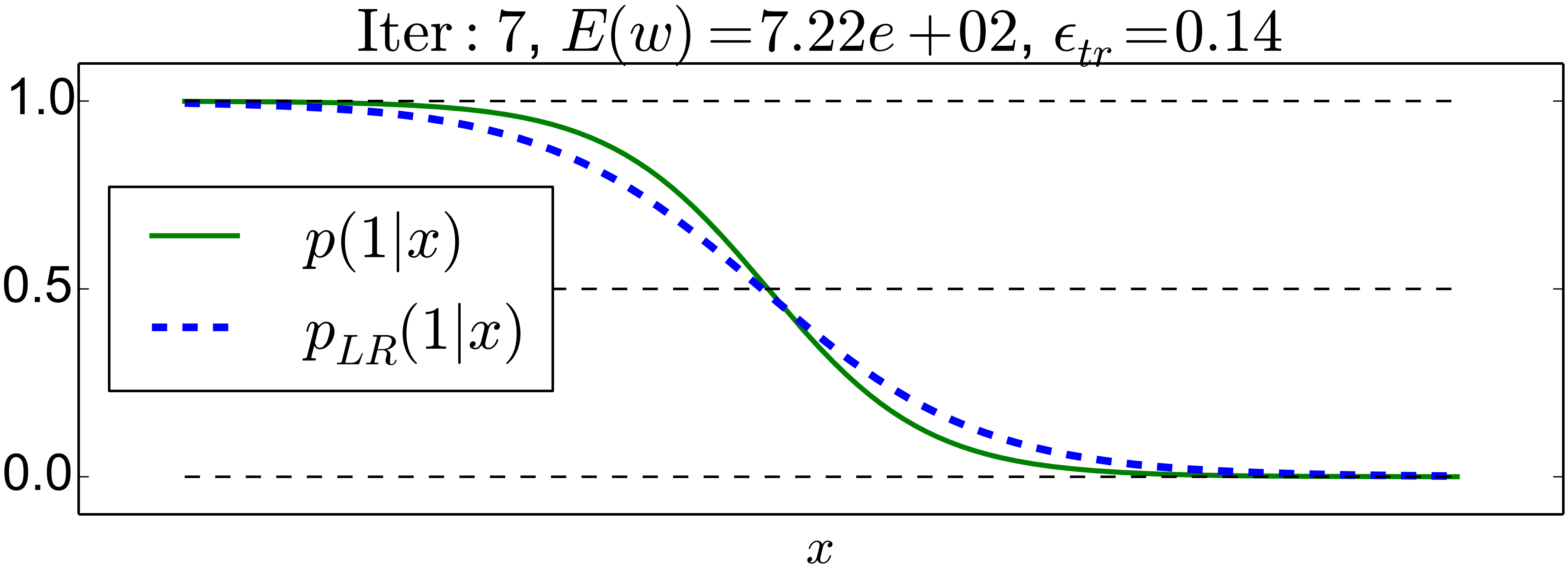

$\mathrm{Iter}: 3,\ E(w) = 8.33e+02,\ \epsilon_{tr} = 0.17$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x

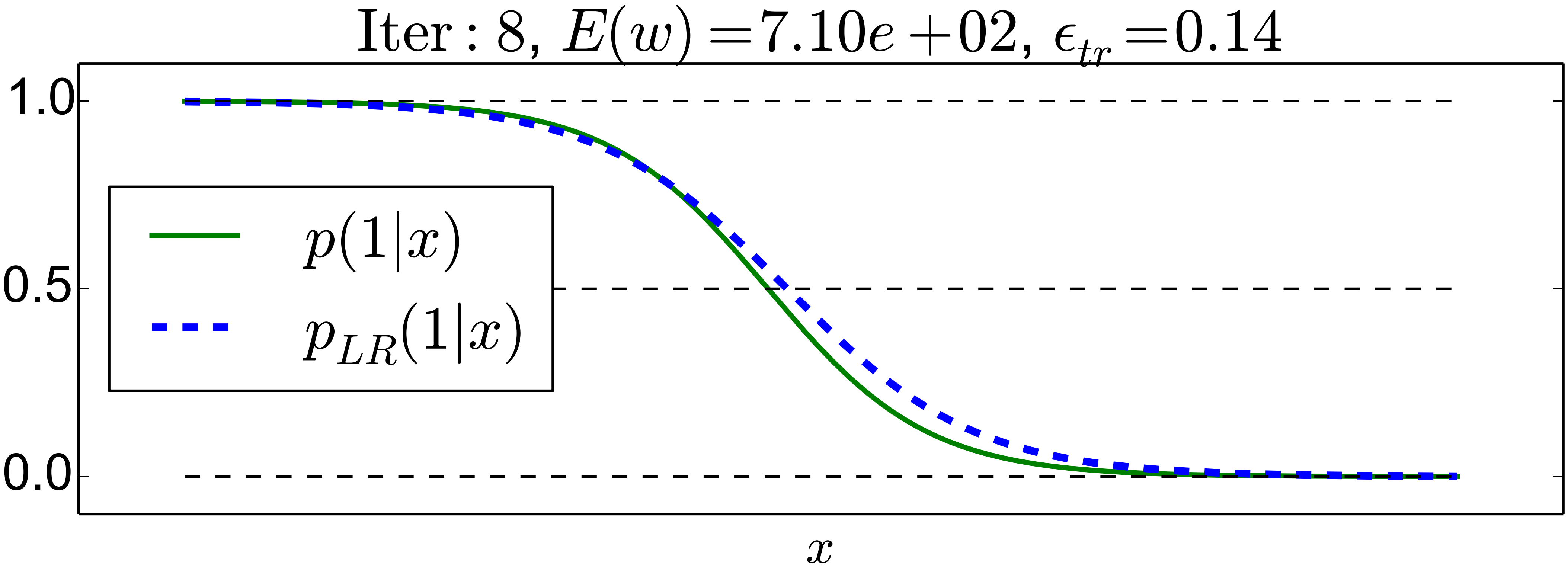


$\mathrm{Iter}: 4,\ E(w) = 7.87e+02,\ \epsilon_{tr} = 0.18$

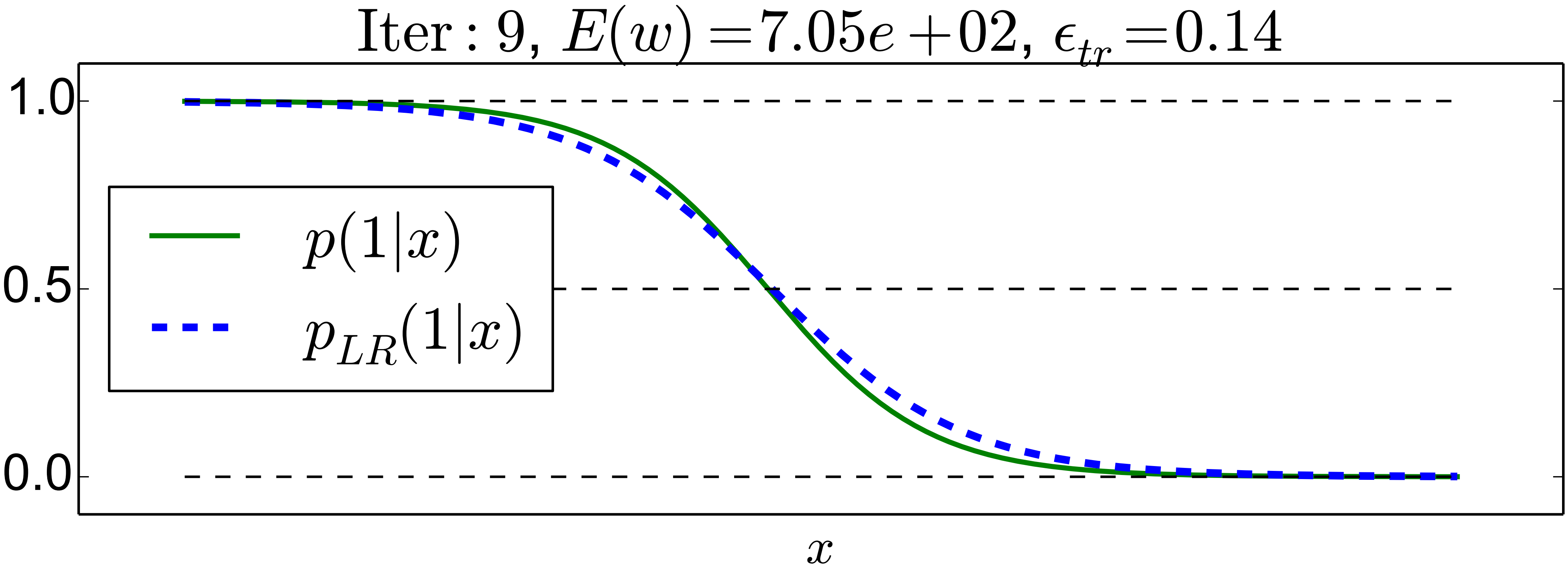

$\mathrm{Iter}: 5,\ E(w) = 7.68e+02,\ \epsilon_{tr} = 0.14$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x

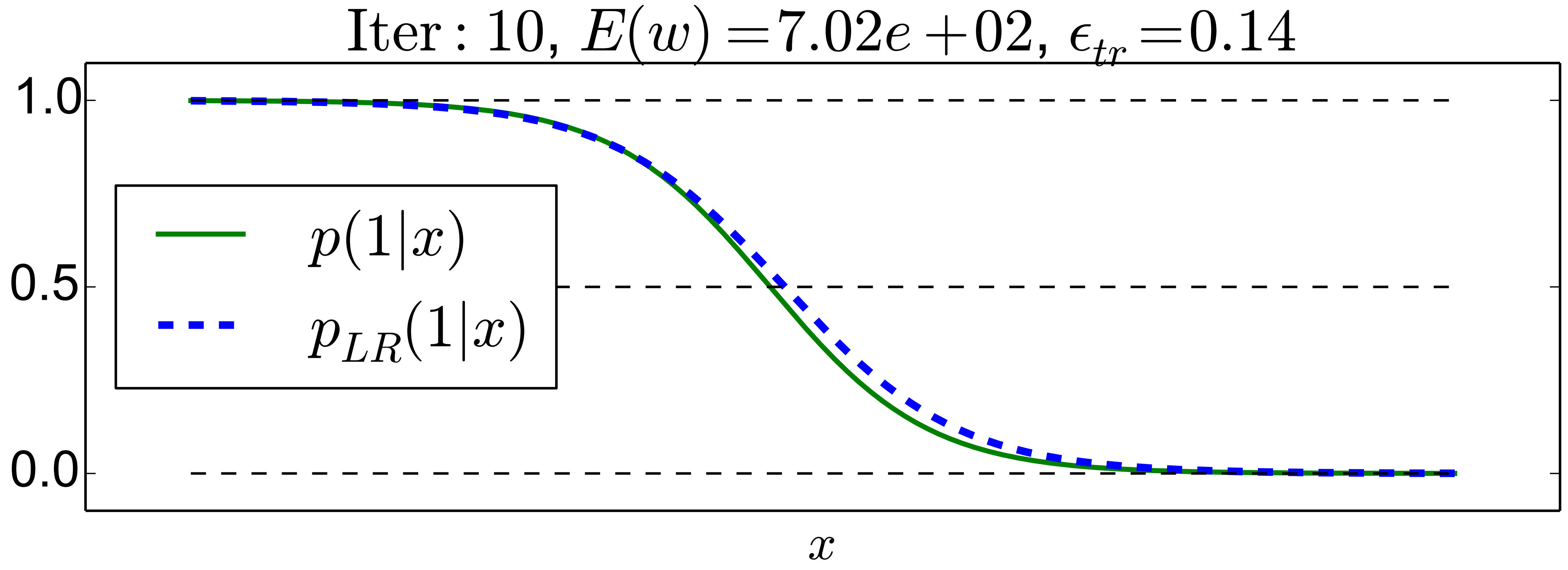

$\mathrm{Iter:6},\ E(w) = 7.30e+02,\ \epsilon_{tr} = 0.15$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x

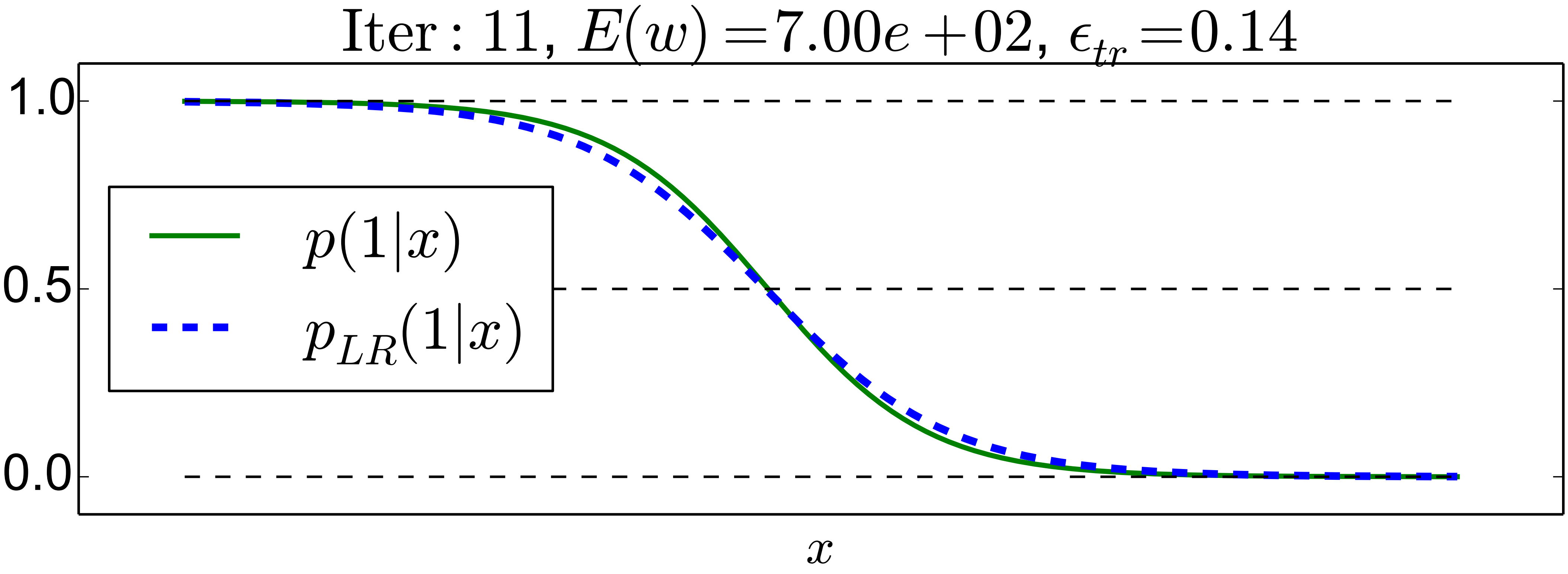

$\mathrm{Iter}: 7,\ E(w) = 7.22e+02,\ \epsilon_{tr} = 0.14$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x



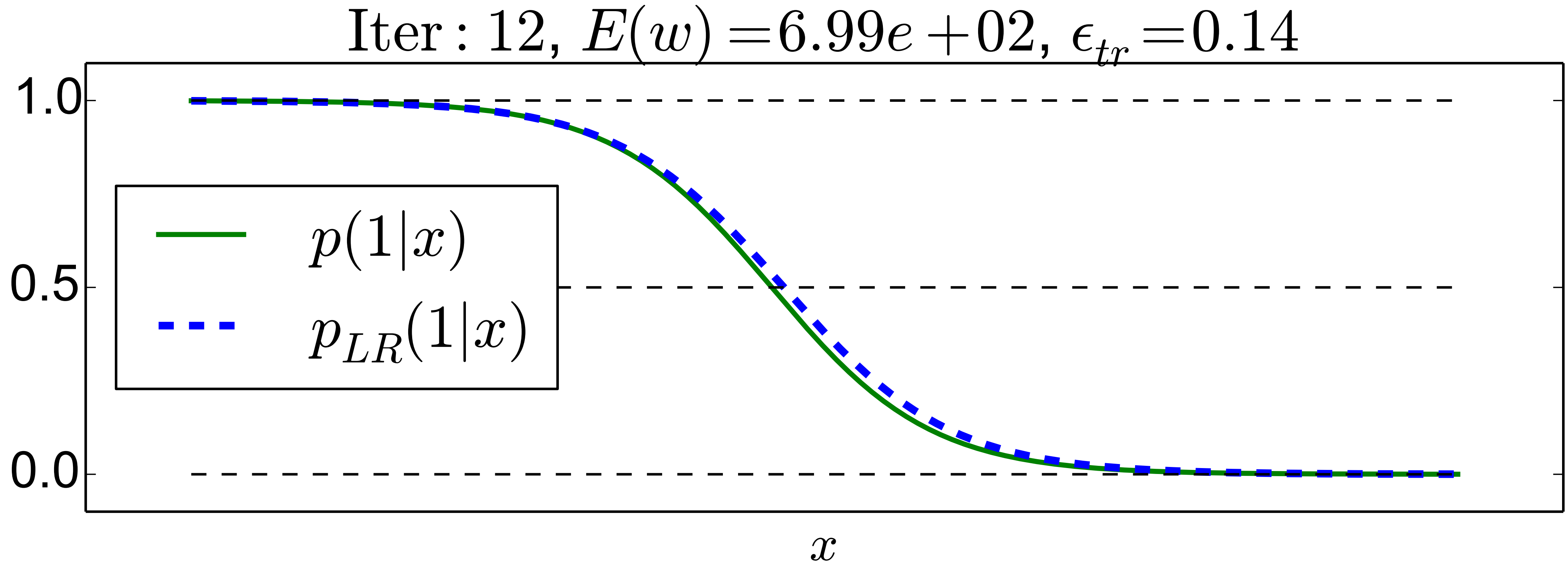
$\mathrm{Iter:}\,8,\ E(w) = 7.10e+02,\ \epsilon_{tr} = 0.14$


Iter : 9, $E(w)$ =7.05e +02, ϵ_{tr} =0.14
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x

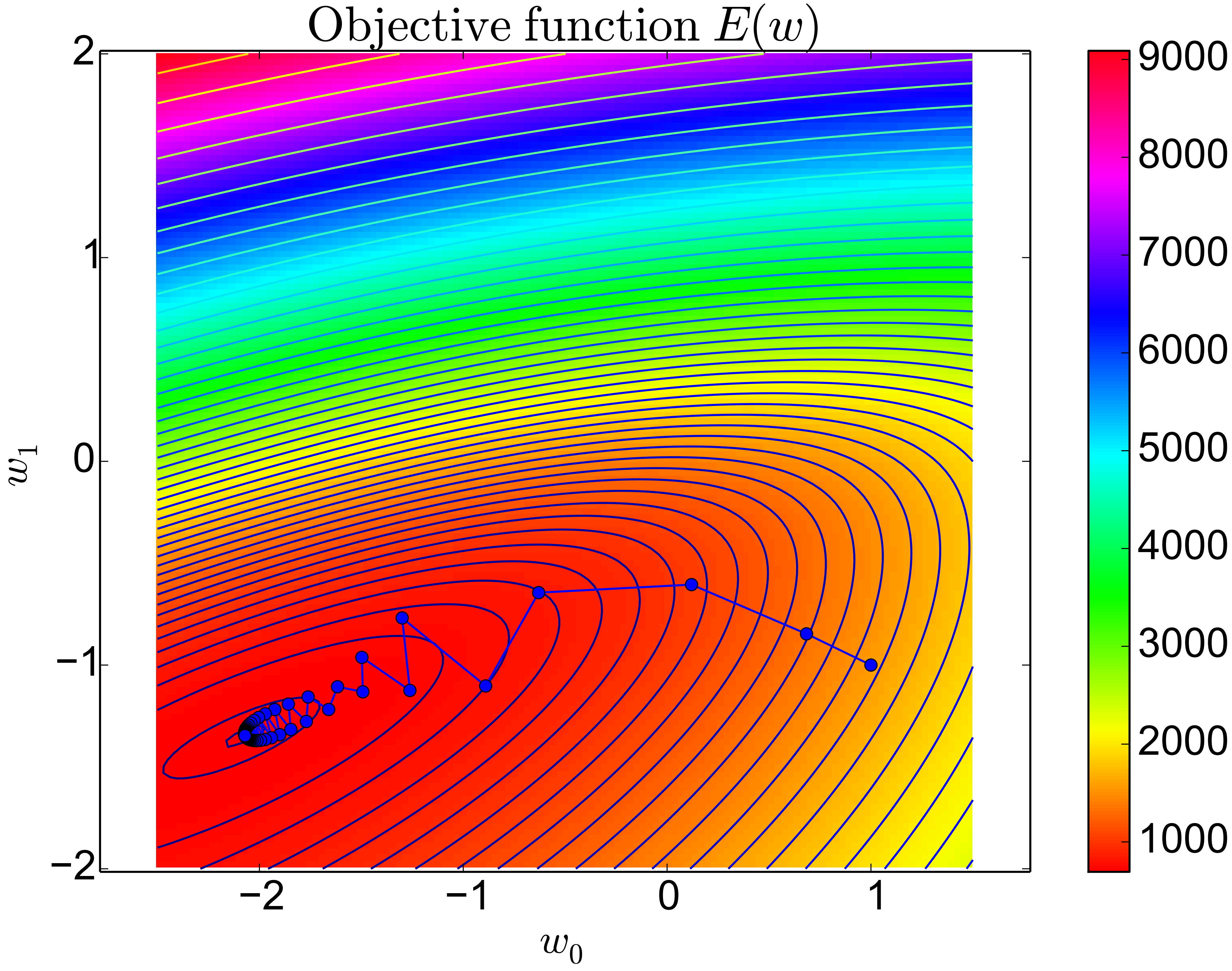

$\mathrm{Iter}: 10,\ E(w) = 7.02e+02,\ \epsilon_{tr} = 0.14$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x

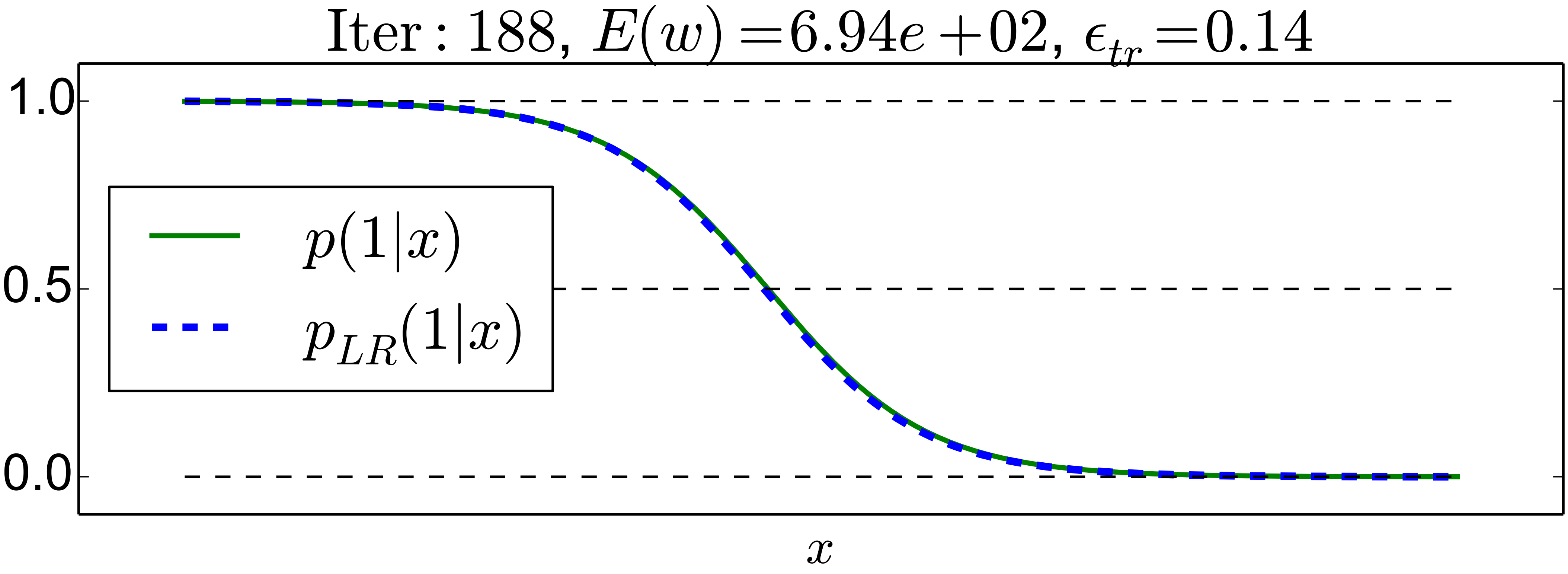


$\mathrm{Iter}: 11,\ E(w)=7.00e+02,\ \epsilon_{tr}=0.14$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x

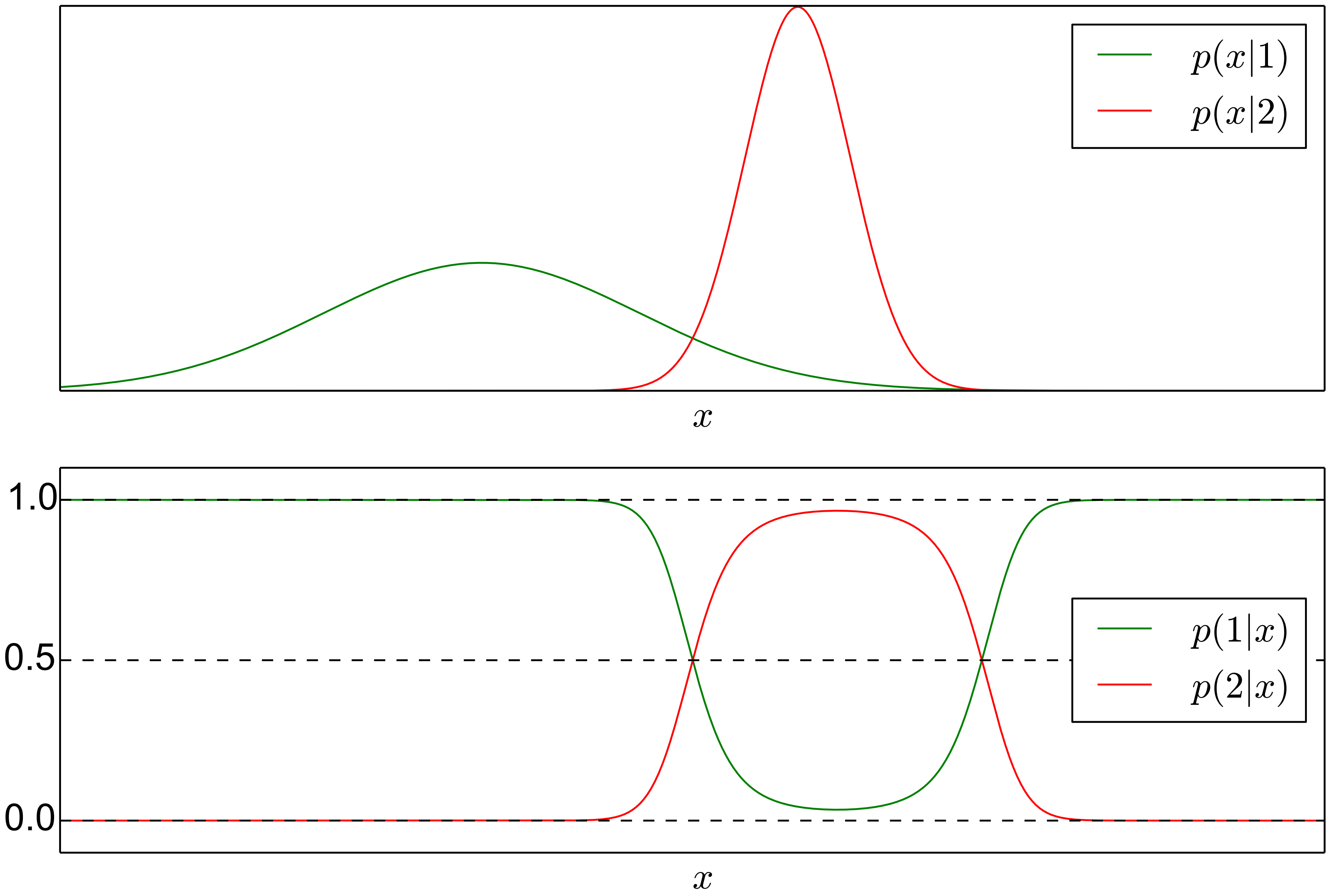


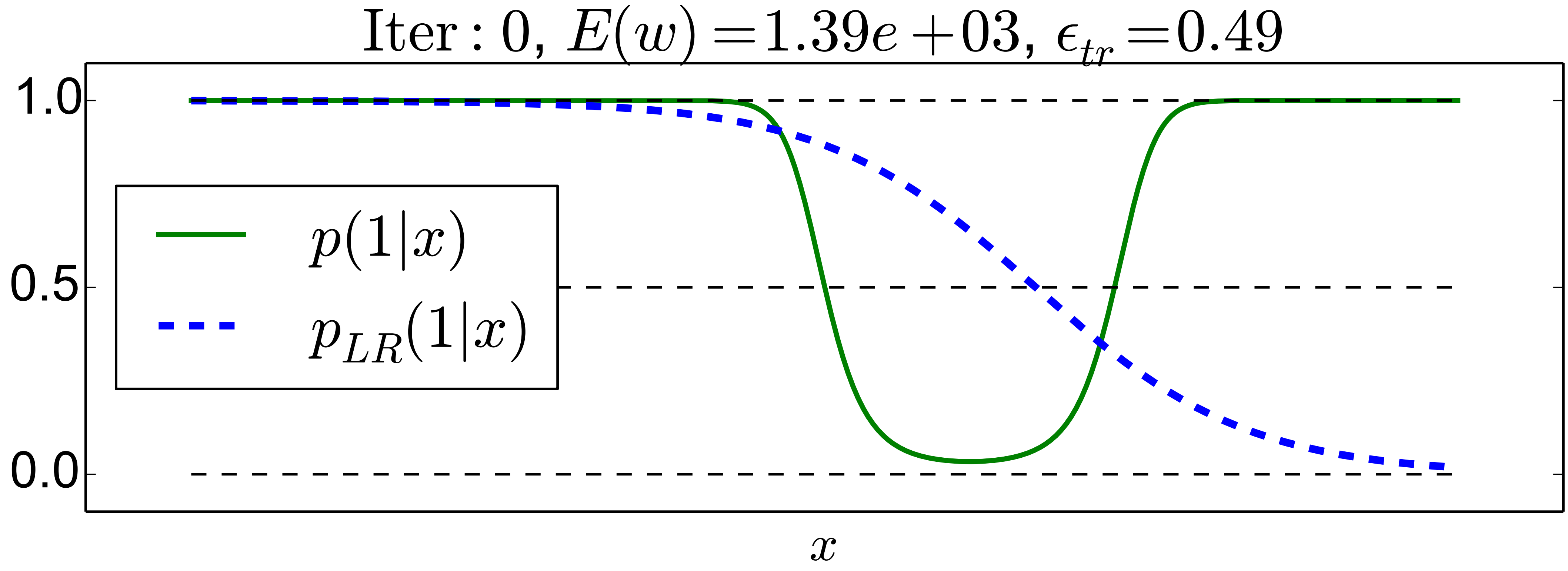
$\mathrm{Iter}: 12,\ E(w) = 6.99e+02,\ \epsilon_{tr} = 0.14$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x

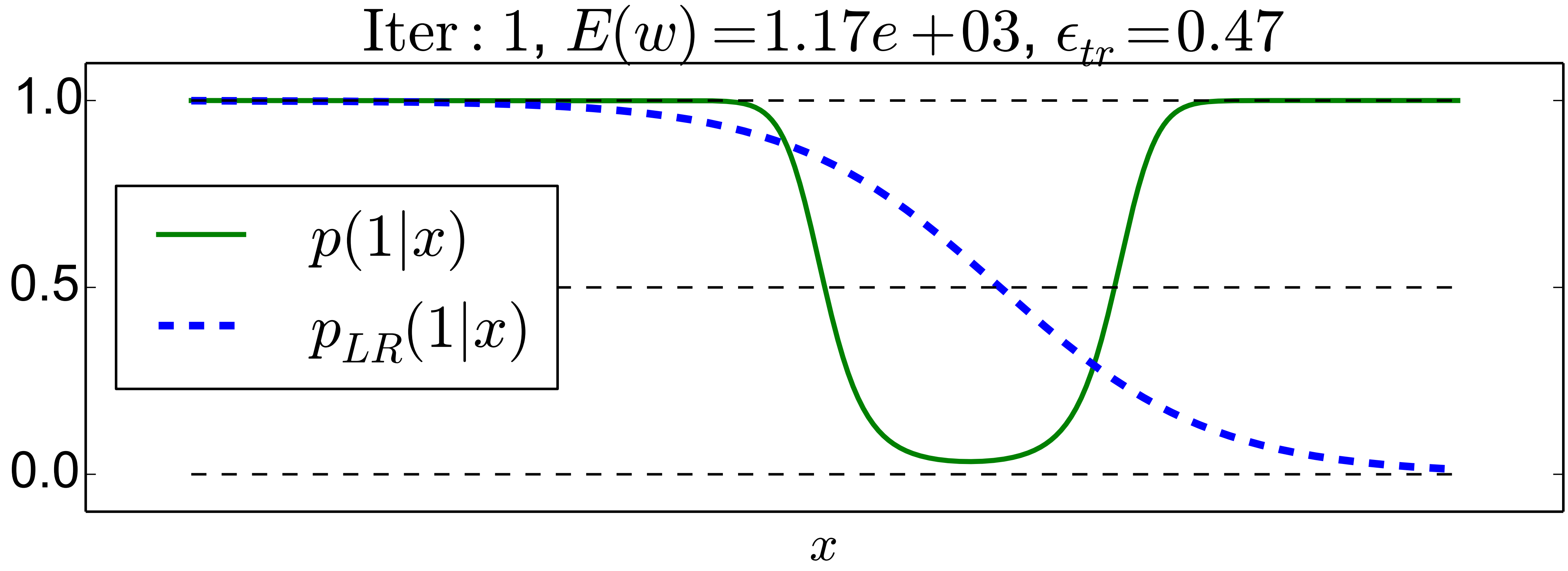

Objective function $E(w)$
2
1
0
−1
−2
w_1
−2
−1
0
1
w_0
9000
8000
7000
6000
5000
4000
3000
2000
1000



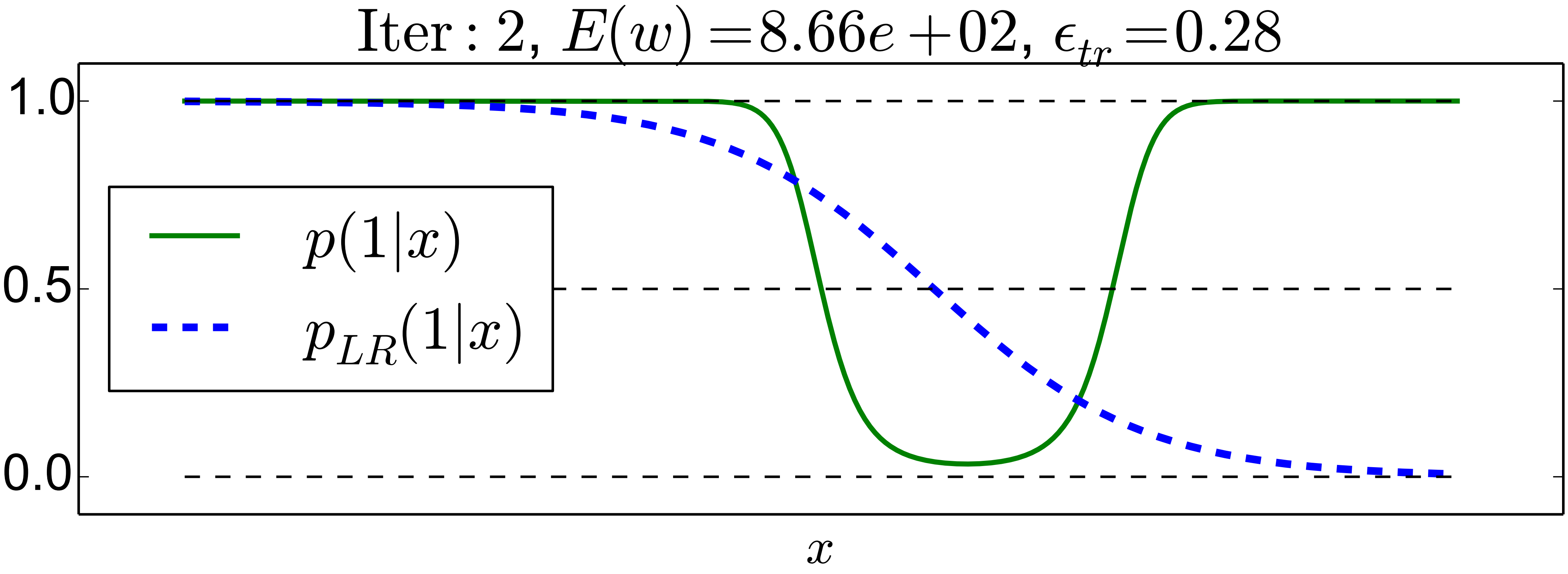
$\mathrm{Iter}: 188,\ E(w) = 6.94e+02,\ \epsilon_{tr} = 0.14$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x


$p(x|1)$
$p(x|2)$
x
1.0
0.5
0.0
$p(1|x)$
$p(2|x)$
x

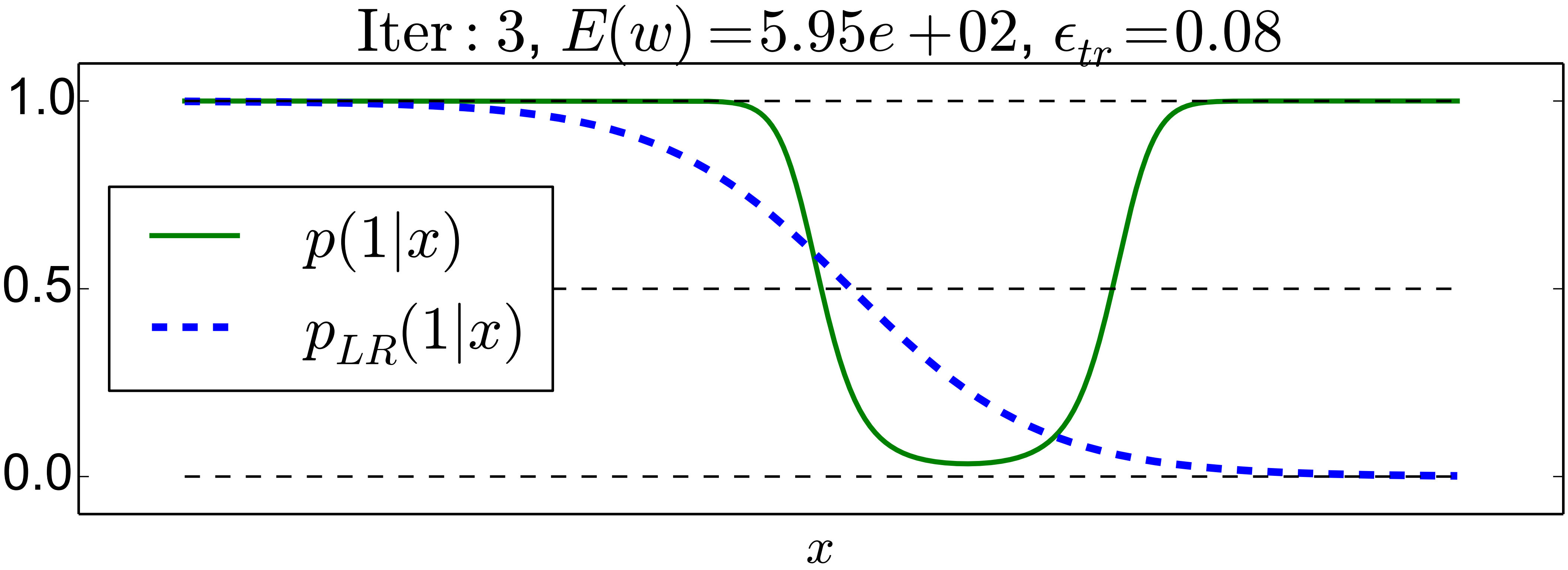

$\mathrm{Iter}: 0,\ E(w) = 1.39e+03,\ \epsilon_{tr} = 0.49$
1.0
0.5
0.0
x
$p(1|x)$
$p_{LR}(1|x)$

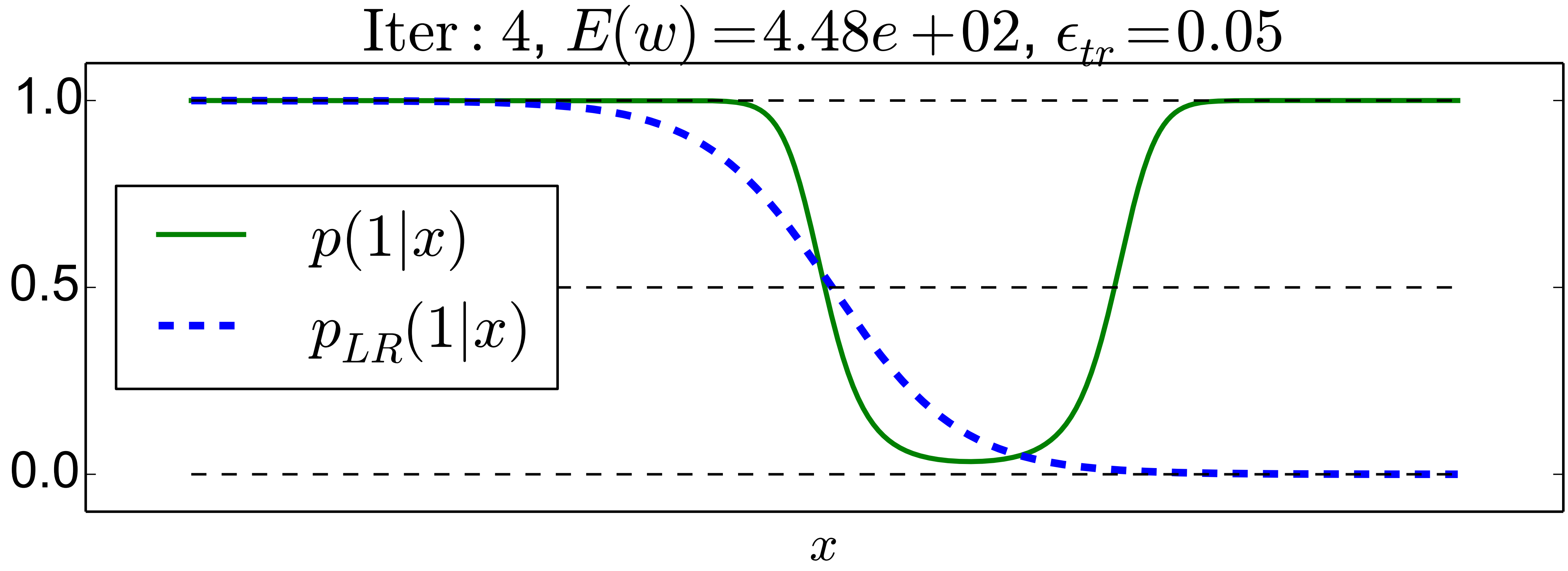


$\mathrm{Iter:}\,1,\ E(w) = 1.17e+03,\ \epsilon_{tr} = 0.47$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x

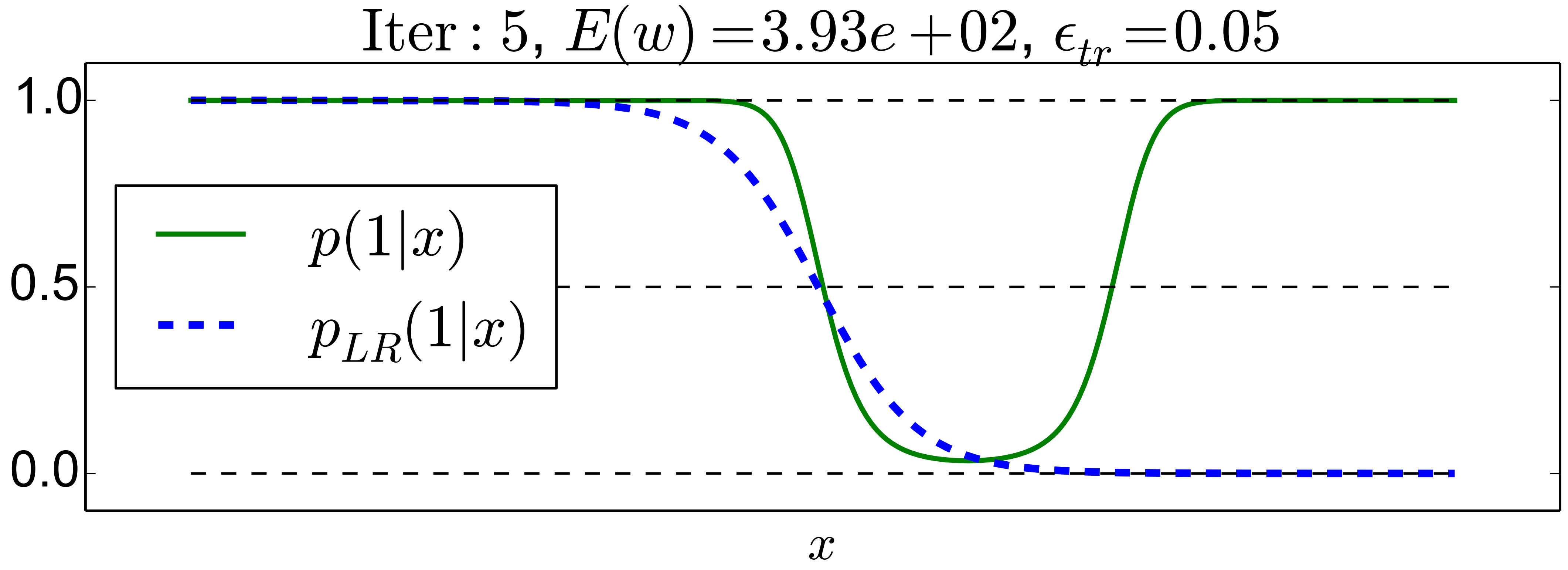


$\mathrm{Iter}: 2,\ E(w) = 8.66e+02,\ \epsilon_{tr} = 0.28$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x

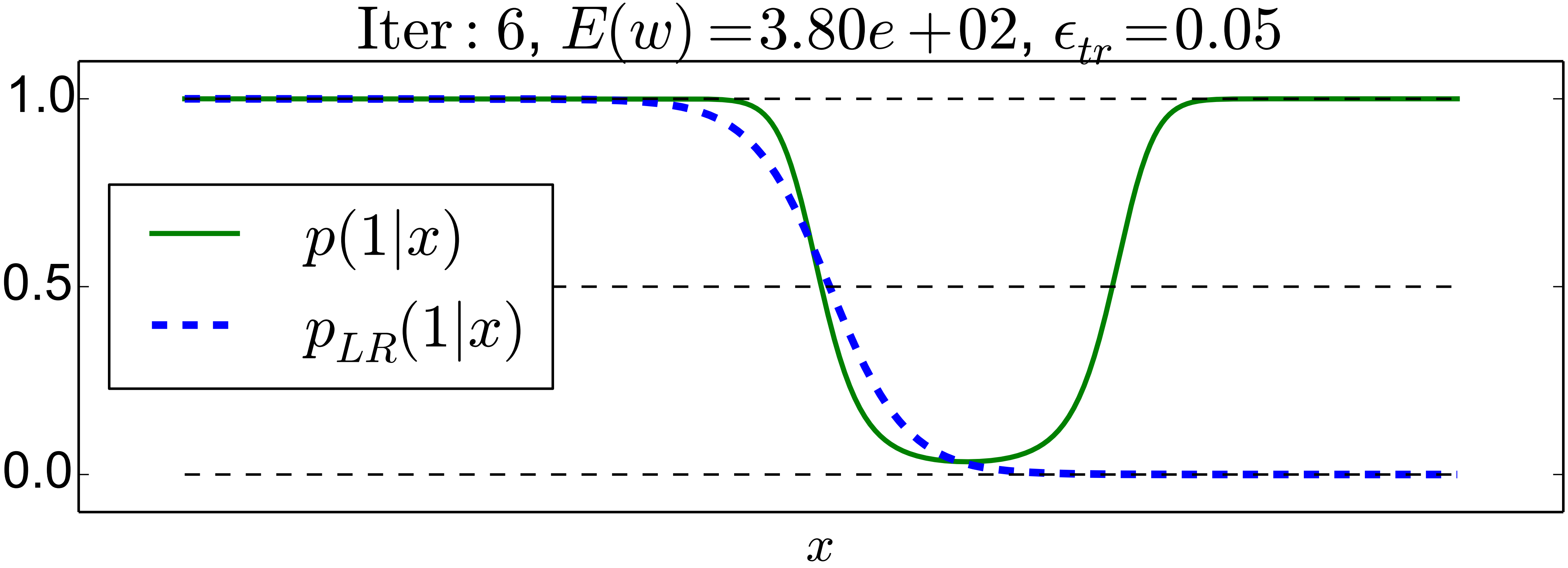

$\mathrm{Iter}: 3,\ E(w) = 5.95e+02,\ \epsilon_{tr} = 0.08$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x



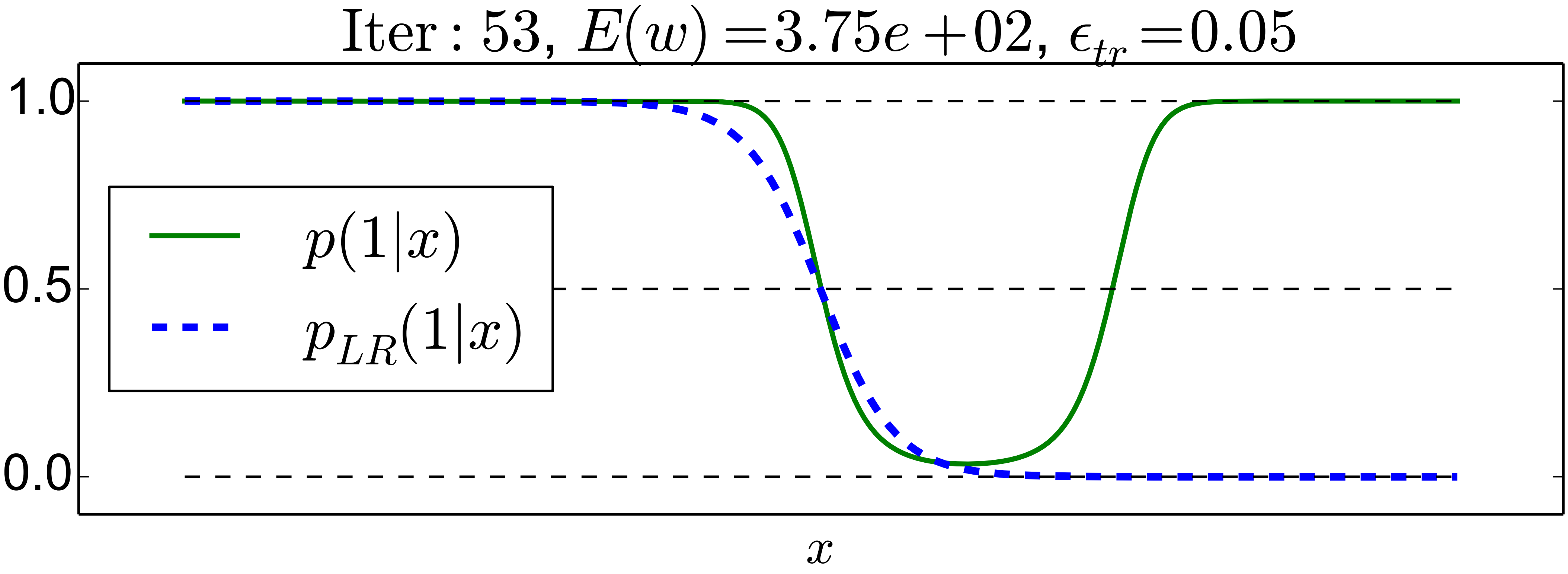
$\mathrm{Iter}: 4,\ E(w) = 4.48e+02,\ \epsilon_{tr} = 0.05$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x



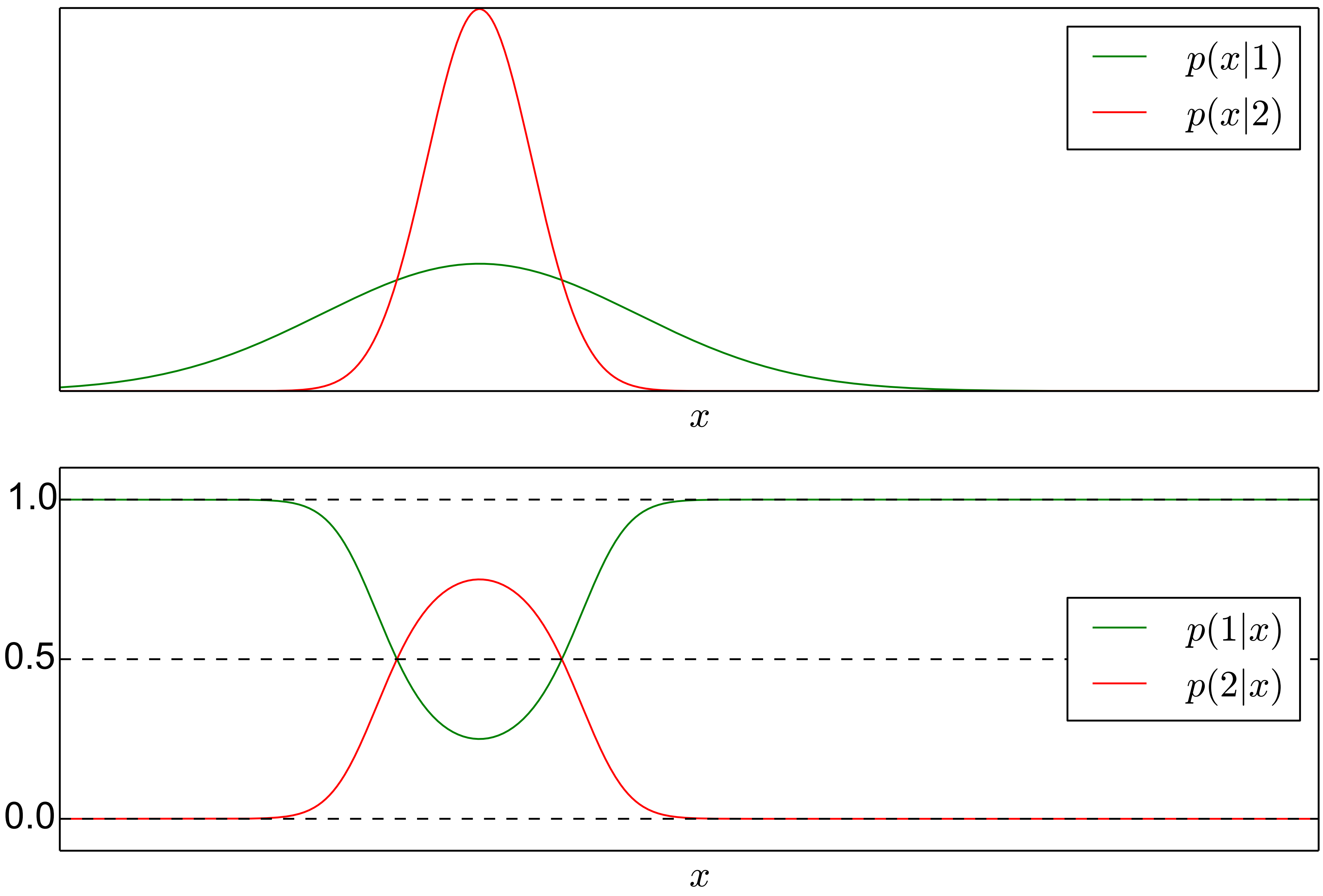
$\mathrm{Iter}: 5,\ E(w) = 3.93e+02,\ \epsilon_{tr} = 0.05$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x

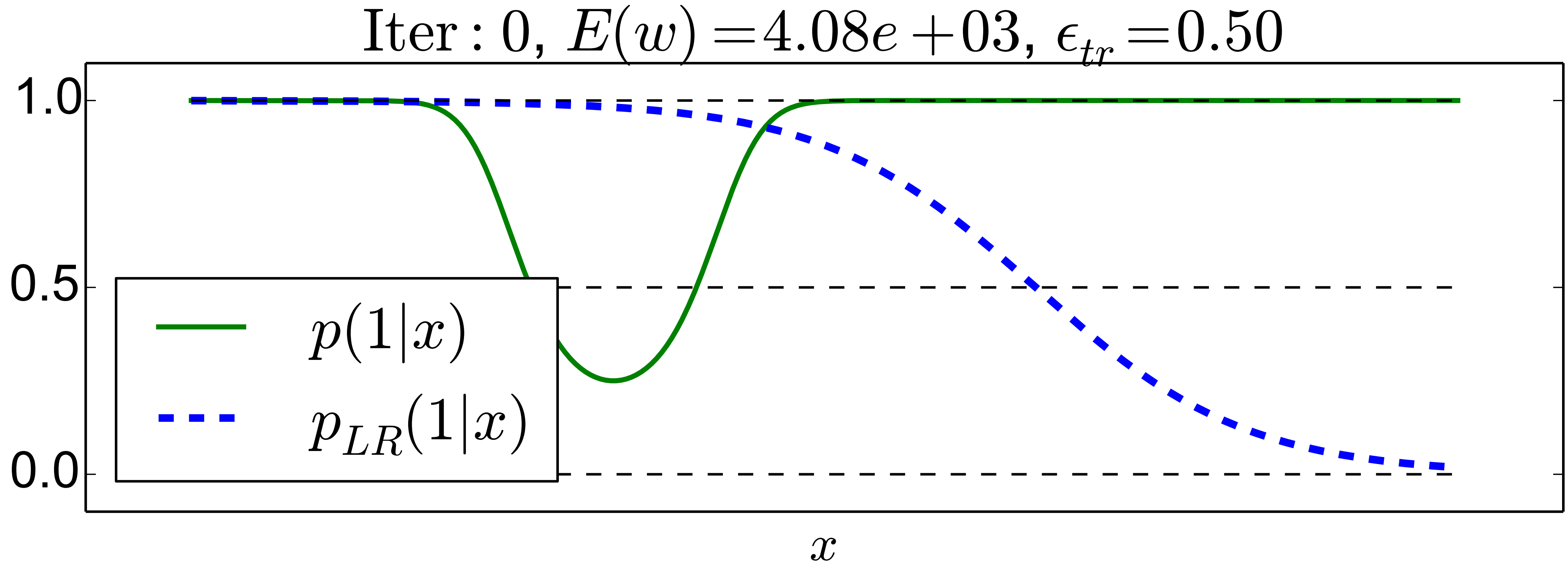


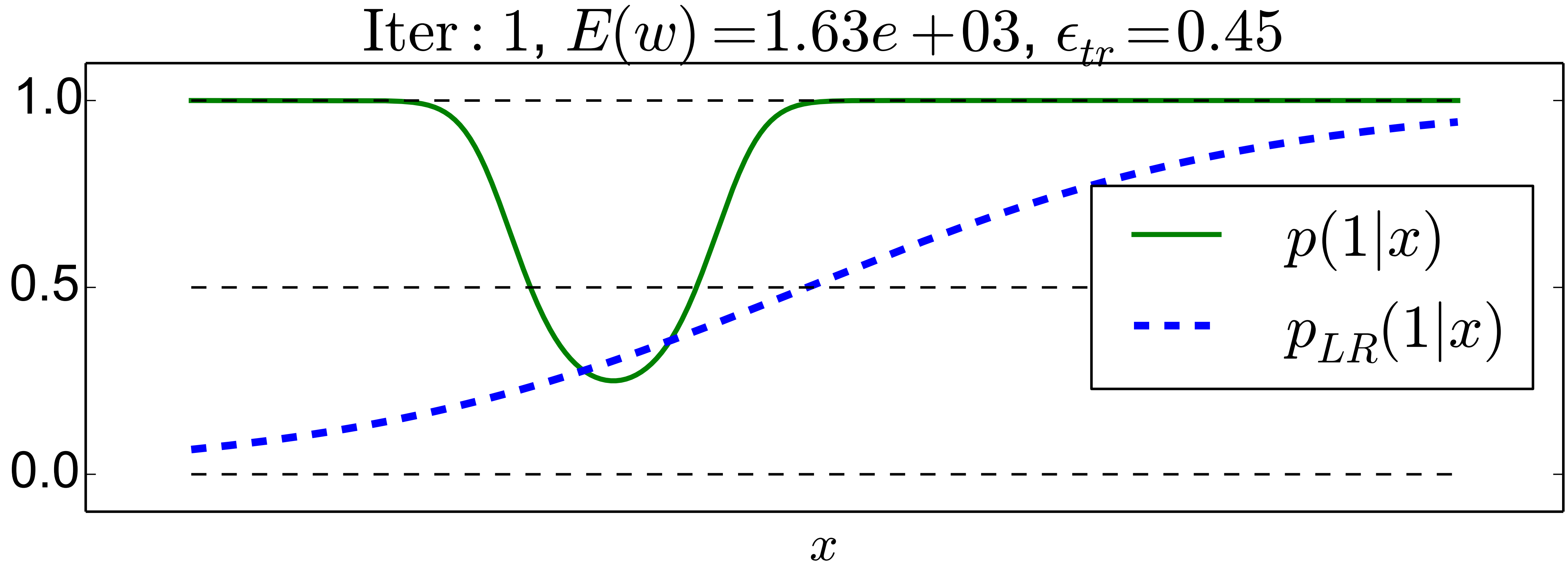
$\mathrm{Iter}: 6,\ E(w) = 3.80e+02,\ \epsilon_{tr} = 0.05$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x

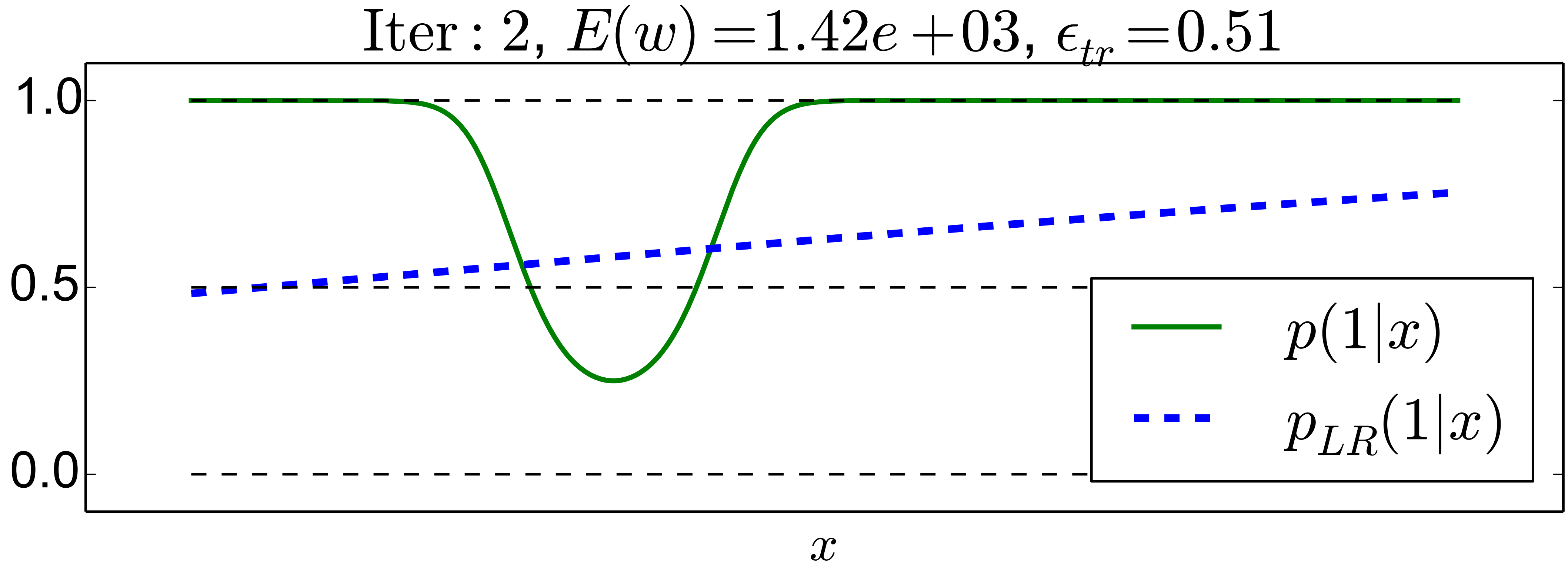


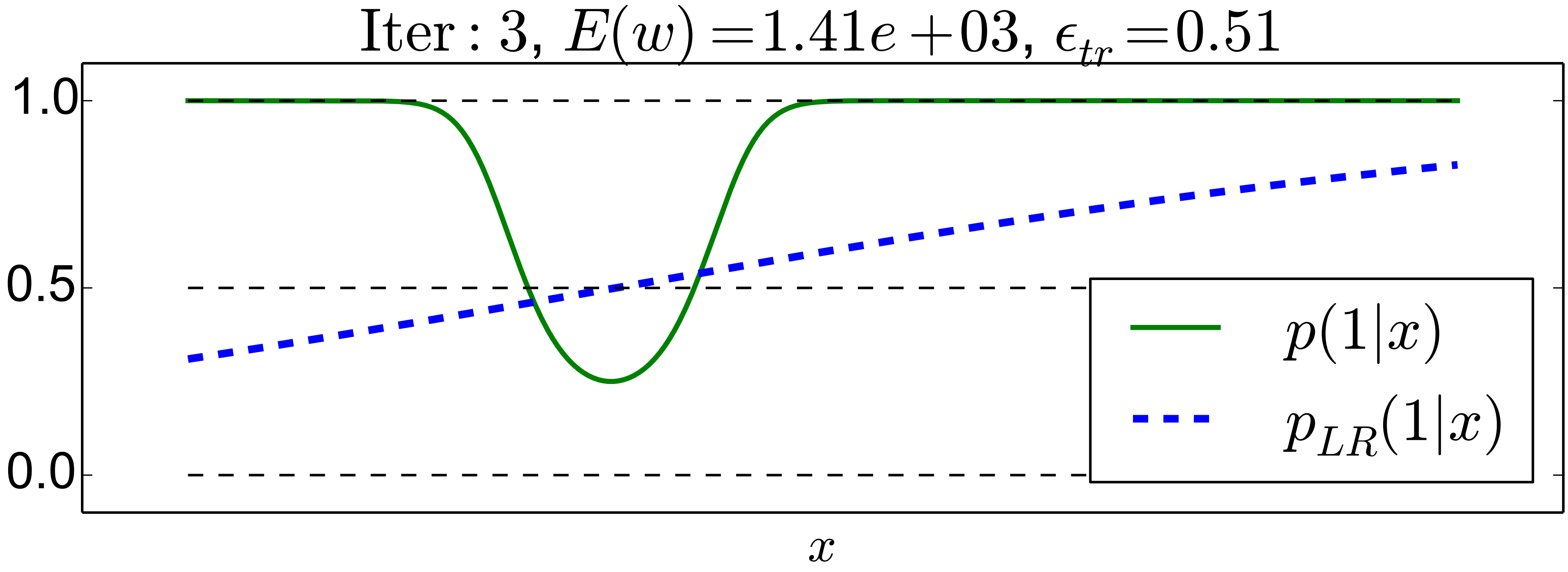
$\mathrm{Iter}: 53,\ E(w) = 3.75e+02,\ \epsilon_{tr} = 0.05$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x

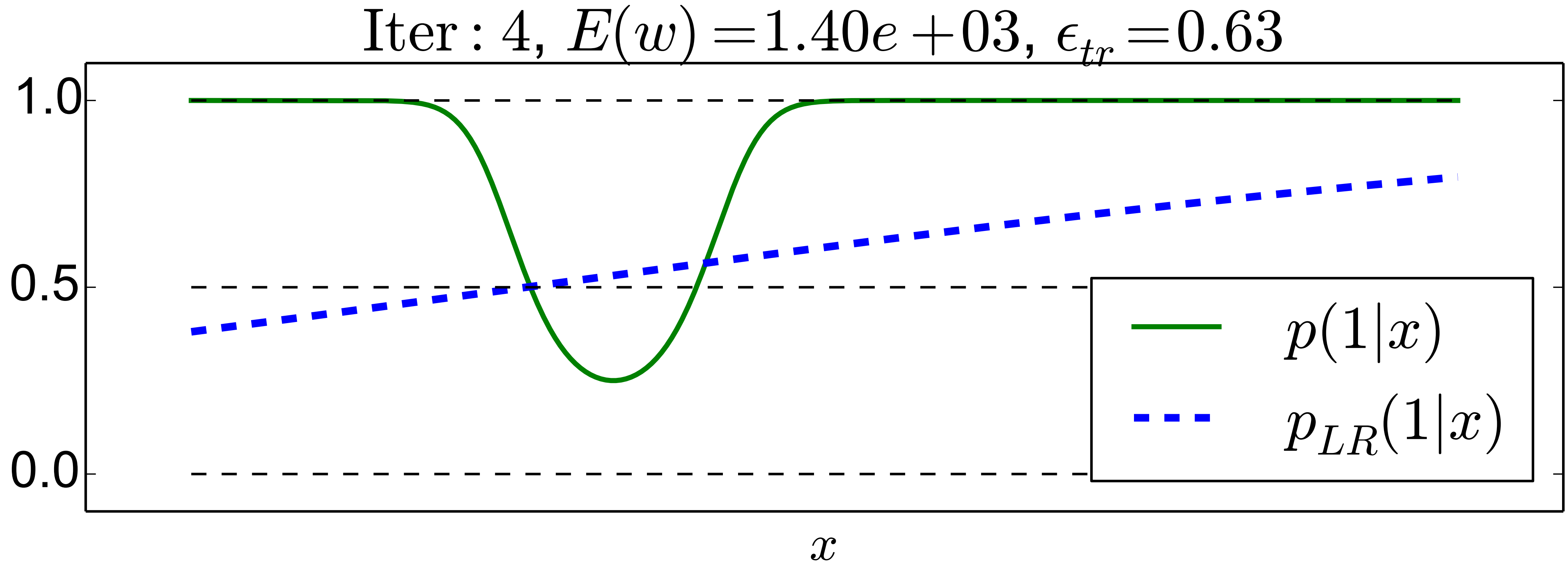

$p(x|1)$
$p(x|2)$
x
1.0
0.5
0.0
$p(1|x)$
$p(2|x)$
x

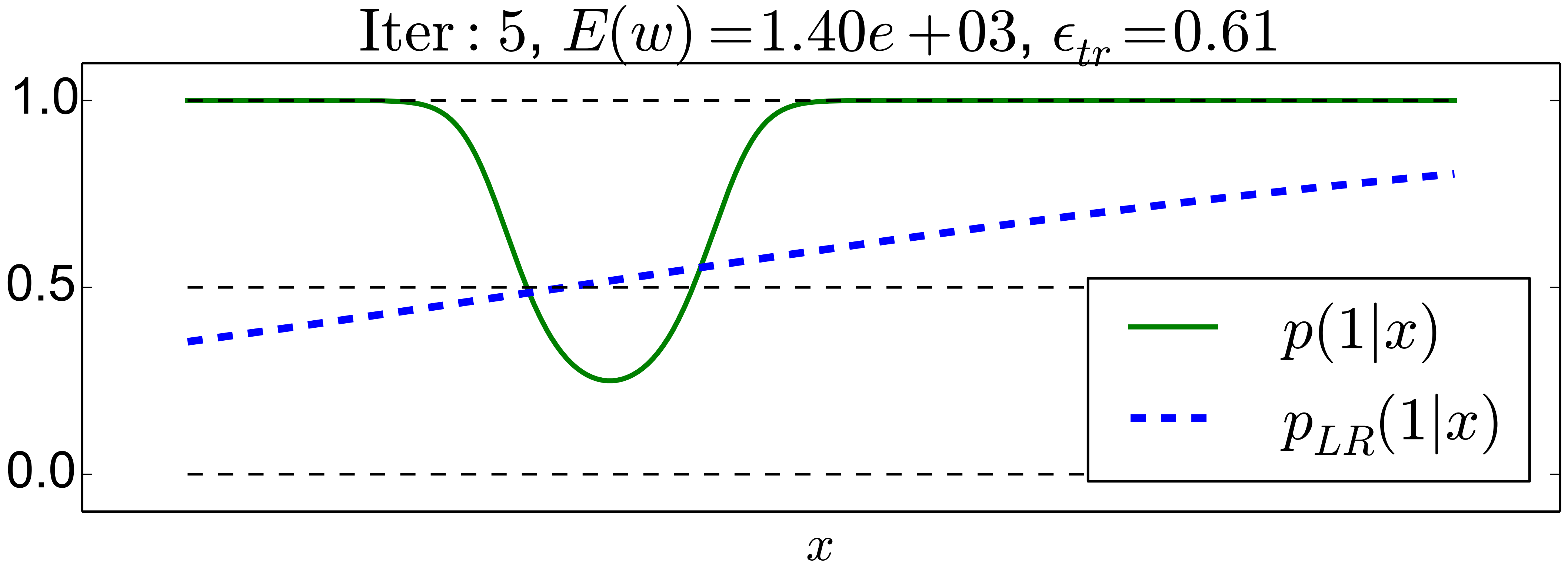

$\mathrm{Iter}: 0,\ E(w) = 4.08e+03,\ \epsilon_{tr} = 0.50$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x

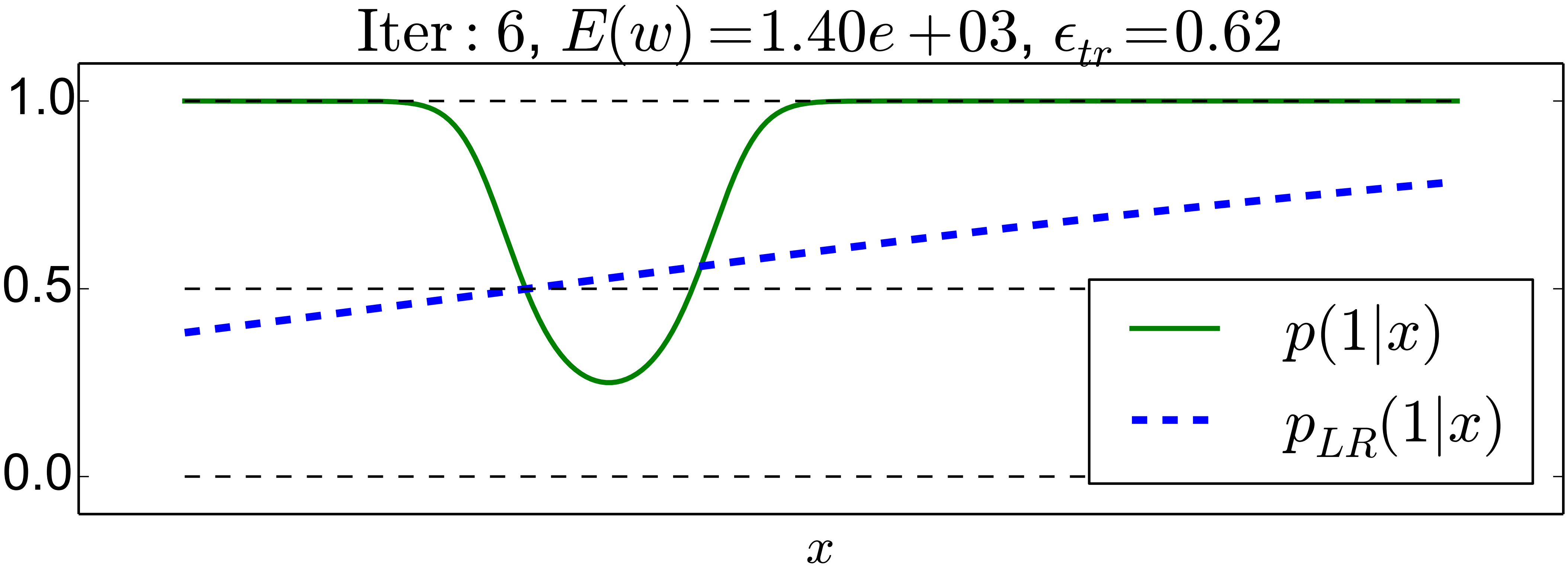

$\mathrm{Iter:1},\ E(w)=1.63e+03,\ \epsilon_{tr}=0.45$
1.0
0.5
0.0
x
$p(1|x)$
$p_{LR}(1|x)$

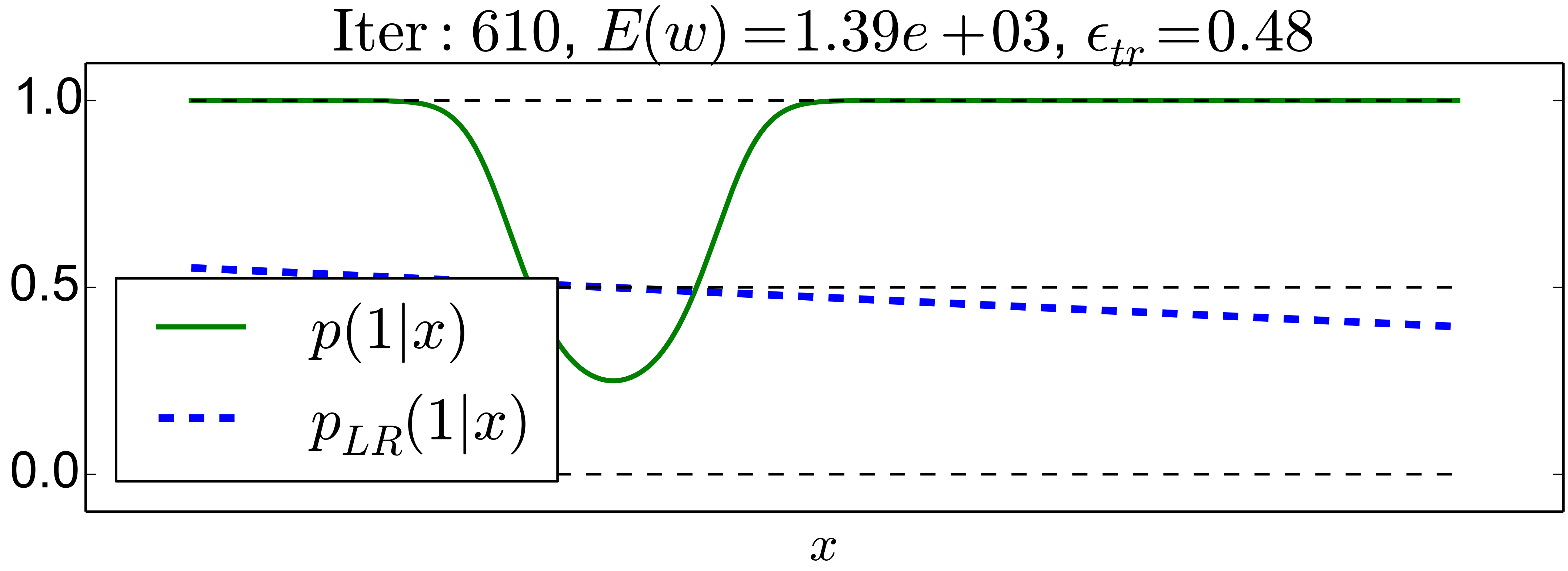

$\mathrm{Iter}: 2,\ E(w) = 1.42e+03,\ \epsilon_{tr} = 0.51$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x


$\mathrm{Iter:3},\ E(w)=1.41e+03,\ \epsilon_{tr}=0.51$
1.0
0.5
0.0
x
$p(1|x)$
$p_{LR}(1|x)$


$\mathrm{Iter}: 4,\ E(w) = 1.40e+03,\ \epsilon_{tr} = 0.63$
1.0
0.5
0.0
x
$p(1|x)$
$p_{LR}(1|x)$


$\mathrm{Iter}: 5,\ E(w) = 1.40e+03,\ \epsilon_{tr} = 0.61$
1.0
0.5
0.0
x
$p(1|x)$
$p_{LR}(1|x)$


Iter : 6, $E(w)$ =1.40e+03, ϵ_{tr} =0.62
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x


$\mathrm{Iter}: 610,\ E(w) = 1.39e+03,\ \epsilon_{tr} = 0.48$
1.0
0.5
0.0
$p(1|x)$
$p_{LR}(1|x)$
x