7.1.2 Relation to logistic regression

As with the separable case, we can re-cast the SVM for non separable distributions in terms of the minimization of a regularized error function.

This will also allow us to highlight similarities, and differences, compared to the logistic regression model.

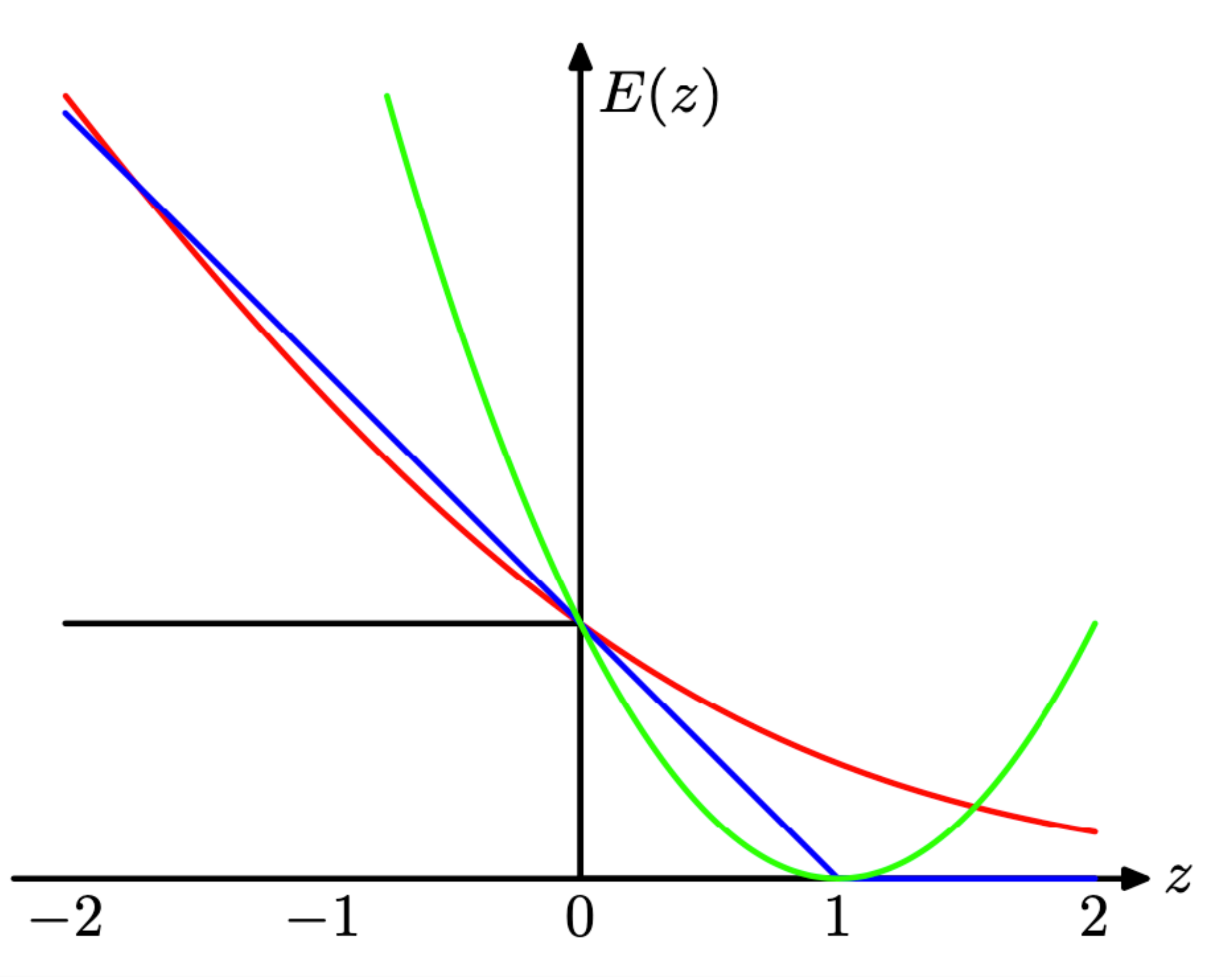
-비슷함과 다름을 logistic regression 과 비교하여 더욱 강조해준다.

Data points that are on the correct side of the margin boundary satisfying $y_n t_n \geq 1$, we have $\xi_n = 0$, and for the remaining points we have $\xi_n = 1 - t_n t_n$.

Thus the objective function can be written up to an overall multiplicative constant

$C \sum_{n=1}^{N} \xi_n + \frac{1}{2}||w||^2 \rightarrow \sum_{n=1}^{N} E_{SV}(y_n t_n) + \lambda ||w||^2$ where $\lambda = (2C)^{-1}$,

$E_{SV}(\cdot)$: hinge error function defined by $E_{SV}(y_n t_n) = [1 - y_n t_n]_+$ ($[\cdot]_+$: the positive part.)



Hinge error function : An approximation to the misclassification error,

(i.e., the error function that ideally we would like to minimize, which is also shown in figure.)

In the logistic regression model, it is easy to handle target variable when $t \in \{0, 1\}$.

For comparison with the support vector machine,

we reformulate maximum likelihood logistic regression using the target variable $t \in \{-1, 1\}$.

To do this,

we note that $p(t = 1|y) = \sigma(y)$ where $y(x)$ is given by (7.1), and $\sigma(y) = \frac{1}{1+e^y}$

$\Rightarrow \ p(t = -1|y) = 1 - \sigma(y) = \sigma(-y)$, and so $p(t|y) = \sigma(yt)$.

From this we can construct an error function by taking the negative logarithm of the likelihood function that, with a quadratic regularizer, takes the form $\sum_{n=1}^{N} E_{LR}(y_n t_n) + \lambda||w||^2$. Where, $E_{LR}(yt) = \ln(1 + \exp(-yt))$.

For comparison with other functions, divide by $\ln(2)$ so that the error function passes through the point (0,1).

This rescaled error function is also plotted in Figure and we see that it has a similar to SVM error function.

The key difference is that the flat region in $E_{SV}(yt)$ leads to sparse solutions.

Both the logistic error and the hinge loss can be viewed as continuous approx to the misclassification error.

Another continuous error function that has sometimes been used to solve classification problems is the squared error, which is again plotted in Figure.

It has the property, however, of placing increasing emphasis on data points that are correctly classified but that are a long way from the decision boundary on the correct side.

Such points will be strongly weighted at the expense of misclassified points, and so if the objective is to minimize the misclassification rate, then a monotonically decreasing error function would be a better choice.