In certain application domains, such as weather prediction, one typically prefers to output a probability (e.g., the chance of rain) instead of a binary prediction (e.g., it will rain). This task corresponds to the problem of learning the function \( \eta(x) = \mathbb{P}(Y = 1 \mid X = x) \) in a binary classification problem. A popular approach to do that is known as \textbf{logistic regression}: we train a predictor \( g : \mathcal{X} \to \mathbb{R} \) and then use \( \sigma(g(x)) \) to predict \( \eta(x) \). The function \( \sigma : \mathbb{R} \to \mathbb{R} \), called logistic, is defined by

\[
\sigma(z) = \frac{1}{1 + e^{-z}} \in (0, 1)
\]

Because we estimate a probability, an appropriate loss function is the logarithmic loss (here we use logarithms in base 2 for convenience),

\[
\ell(y, \hat{y}) = \mathbb{I}\{y = +1\} \log_2 \frac{1}{\hat{y}} + \mathbb{I}\{y = -1\} \log_2 \frac{1}{1 - \hat{y}}
\]

Noting that \( 1 - \sigma(z) = \sigma(-z) \), we can write the identity

\[
\mathbb{I}\{y = +1\} \log_2 \frac{1}{\hat{y}} + \mathbb{I}\{y = -1\} \log_2 \frac{1}{1 - \hat{y}} = \log_2 \left( 1 + e^{-yg(x)} \right)
\]

where \( \hat{y} = \sigma(g(x)) \). The right-hand side of the above identity is a function known as \textbf{logistic loss}, and is typically defined using \( \hat{y} = g(x) \) as follows,

\[
\ell(y, \hat{y}) = \log_2 \left( 1 + e^{-y\hat{y}} \right)
\]

We now describe the important case of logistic regression when \( g(x) \) is a linear model \( w^\top x \). Given a training set \( S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \), let \( \ell_t(w) = \log_2 \left( 1 + e^{-y_t w^\top x_t} \right) \), we show how to compute \( \nabla \ell_t(w) \). Let \( s_t = w^\top x_t \). First, observe that

\[
\frac{d}{ds_t} \log_2 \left( 1 + e^{-y s_t} \right) = \frac{1}{\ln 2} \frac{-y_t e^{-y_t s_t}}{1 + e^{-y_t s_t}} = \frac{1}{\ln 2} \frac{-y_t}{1 + e^{y_t s_t}} = \frac{-y_t \sigma(-y_t s_t)}{\ln 2}
\]

Therefore,

\[
\nabla \ell_t(w) = \left( \frac{d}{ds_t} \log_2 \left( 1 + e^{-y s_t} \right) \bigg|_{s_t = w^\top x_t} \right) x_t = \frac{-\sigma(-y_t w^\top x_t)}{\ln 2} y_t x_t.
\]

The gradient descent update can then be written as

\[
w_{t+1} = w_t + \eta_t \sigma(-y_t w^\top x_t) y_t x_t
\]

where we hid the \( \ln 2 \) factor in the learning rate \( \eta_t \).
To avoid overfitting, logistic regression is often used with a regularization term that enforces stability,
\[
\ell_t(w) = \log_2 \left(1 + e^{-y_t w^\top x_t}\right) + \frac{\lambda}{2} \|w\|^2.
\]

If we run stochastic gradient descent using regularized logistic regression we get an algorithm similar to Pegasos for regularized hinge loss.

**Surrogate losses** \( \ell : \{-1, 1\} \times \mathbb{R} \rightarrow \mathbb{R} \) are convex upper bounds on the zero-one loss function for binary classification. We already encountered three of them:

- Hinge loss \( \ell(y, \tilde{y}) = [1 - y \tilde{y}]_+ \)
- Boosting loss \( \ell(y, \tilde{y}) = e^{-y \tilde{y}} \)
- Logistic loss \( \ell(y, \tilde{y}) = \log_2 \left(1 + e^{-y \tilde{y}}\right) \)

where \( y \in \{-1, 1\} \) and \( \tilde{y} \in \mathbb{R} \).

As many surrogate losses exist, we may wonder whether some of them should be preferred over the others. We now define an important criterion, called consistency, that a surrogate loss may satisfy with respect to the function \( \eta(x) = \mathbb{P}(Y = 1 \mid X = x) \) which defines the Bayes optimal predictor \( f^* \).

A surrogate loss function \( \ell \) is **consistent** if, for all \( x \in \mathcal{X} \),

\[
\text{sgn}(g^*(x)) = f^* \quad \text{for} \quad g^*(x) = \arg\min_{\tilde{y} \in \mathbb{R}} \mathbb{E}[\ell(Y, \tilde{y}) \mid X = x]
\]

In other words, the sign of the Bayes optimal predictor for the surrogate loss must be the Bayes optimal classifier for the zero-one loss.

We now verify the consistency of the logistic loss. By taking derivatives, it is easy to check that

\[
g^*(x) = \arg\min_{\tilde{y} \in \mathbb{R}} \left( \eta(x) \log_2 \left(1 + e^{-\tilde{y}}\right) + (1 - \eta(x)) \log_2 \left(1 + e^{\tilde{y}}\right) \right) = \ln \frac{\eta(x)}{1 - \eta(x)}
\]

which implies

\[
\text{sgn}(g^*(x)) = \text{sgn} \left( \ln \frac{\eta(x)}{1 - \eta(x)} \right) = \text{sgn}(\eta(x) - \frac{1}{2}) = f^*(x)
\]

The Bayes optimal prediction \( g^*(x) = \ln \frac{\eta(x)}{1 - \eta(x)} \) for the logistic loss is known as log-odds ratio. If we compute the conditional Bayes risk of \( g^* \) with respect to the logistic loss we get

\[
\mathbb{E} \left[ \log_2 \left(1 + e^{-Y g^*(x)}\right) \mid X = x \right] = -\eta(x) \log_2 \eta(x) - (1 - \eta(x)) \log_2 (1 - \eta(x))
\]

The quantity on the right-hand side is the entropy \( H(Y \mid X = x) \) of \( Y \) for \( X = x \). This corresponds to the expected number of bits that we receive by observing \( Y \) when \( X \) is already known. From the conditional Bayes risk, we can easily obtain the Bayes risk,

\[
\ell_D(g^*) = \mathbb{E} \left[ \log_2 \left(1 + e^{-Y g^*(x)}\right) \right] = H(Y \mid X)
\]

The quantity on the right-hand side is now the conditional entropy \( H(Y \mid X) \) of the label \( Y \) given \( X \), which corresponds the Bayes risk for the logistic loss.
Next, we verify the consistency of the hinge loss. We have

\[ g^*(x) = \arg\min_{\widehat{y} \in \mathbb{R}} \left( \eta(x) \left[1 - \widehat{y}\right]_+ + (1 - \eta(x)) \left[1 + \widehat{y}\right]_+ \right) \]

\[ = \arg\min_{\widehat{y} \in [-1, +1]} \left( \eta(x) \left[1 - \widehat{y}\right]_+ + (1 - \eta(x)) \left[1 + \widehat{y}\right]_+ \right) \]

\[ = \arg\min_{\widehat{y} \in [-1, +1]} \left( 1 + (1 - 2\eta(x))\widehat{y} \right) \]

\[ = \begin{cases} 
-1 & \text{if } \eta(x) \leq 1/2, \\
+1 & \text{otherwise} 
\end{cases} \]

\[ = f^*(x) \]

In the second inequality, we could replace \( \widehat{y} \in \mathbb{R} \) with \( \widehat{y} \in [-1, +1] \) because both functions \([1 - \widehat{y}]_+\) and \([1 + \widehat{y}]_+\) increase or remain constant outside of the interval \([-1, +1]\).

More generally, the following result holds.

**Theorem 1 (Sufficient condition for consistency of a surrogate loss).** If a surrogate loss \( \ell : \{-1, 1\} \times \mathbb{R} \to \mathbb{R} \) is such that for all \( y \in \{-1, 1\} \) the derivative \( \ell'(y, 0) \) exists and satisfies \( \ell'(y, 0) < 0 \), then \( \ell \) is consistent.

Besides the hinge loss and the logistic loss, also the boosting loss, the square loss \( \ell(y, b_y) = (1 - y b_y)^2 \) and the quadratic hinge loss \( \ell(y, \widehat{y}) = \left([1 - y \widehat{y}]_+\right)^2 \) are all consistent.