Nonlinear Regression

\[ \hat{y} = h(w \cdot x) \]

- Sigmoid function \( h(z) = \frac{1}{1+e^{-z}} \)
- For a set of examples \((x_1, y_1), \ldots, (x_T, y_T)\)
  total loss \( \sum_{t=1}^{T} (h(w \cdot x) - y_t)^2 / 2 \)
  can have exponentially many minima
  in weight space

[Bu, AHW]
Want loss that is convex in $\mathbf{w}$
Bregman Div. Lead to Good Loss Function

\[ \int_{h^{-1}(y)} w \cdot x (h(z) - y) \, dz = H(w \cdot x) - H(h^{-1}(y)) - (w \cdot x - h^{-1}(y)) y \]

\[ = \Delta_H(w \cdot x, h^{-1}(y)) \quad (h = \nabla H) \]
Use $\Delta_H(\mathbf{w} \cdot \mathbf{x}, h^{-1}(y))$ as loss of $\mathbf{w}$ on $(\mathbf{x}, y)$

Called matching loss for $h$ [AHW, HKW]

Matching loss is convex in $\mathbf{w}$

<table>
<thead>
<tr>
<th>transfer f.</th>
<th>$H(z)$</th>
<th>match. loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h(z)$</td>
<td>$\frac{1}{2}z^2$</td>
<td>$\frac{1}{2}(\mathbf{w} \cdot \mathbf{x} - y)^2$</td>
</tr>
<tr>
<td>$z$</td>
<td>$\frac{1}{2}z^2$</td>
<td>square loss</td>
</tr>
<tr>
<td>$\frac{e^z}{1+e^z}$</td>
<td>$\ln(1 + e^z)$</td>
<td>$\ln(1 + e^{\mathbf{w} \cdot \mathbf{x}}) - y\mathbf{w} \cdot \mathbf{x}$</td>
</tr>
<tr>
<td>$\text{sign}(z)$</td>
<td>$</td>
<td>z</td>
</tr>
</tbody>
</table>

hinge loss
Idea behind the matching loss

If transfer function and loss match, then

$$\nabla_{\mathbf{w}} \Delta_H (\mathbf{w} \cdot \mathbf{x}, h^{-1}(y)) = h(\mathbf{w} \cdot \mathbf{x}) - y$$

Then update has simple form:

$$f(\mathbf{w}_{t+1}) = f(\mathbf{w}_t) - \eta_t (h(\mathbf{w}_t \cdot \mathbf{x}) - y_t) \mathbf{x}_t$$

This can be exploited in proofs

But not absolutely necessary
One only needs convexity of $L(h(\mathbf{w} \cdot \mathbf{x}), y)$ in $\mathbf{w}$

[Ce]
For transfer function $h(z) = \text{sign}(z)$

$H(z) = |z|$

Matching loss is **hinge loss**

$$HL(w \cdot x, h^{-1}(y)) = \max\{0, -y w \cdot x\}$$

Convex in $w$ but not differentiable
Motivation of linear threshold algs

Gradient descent
  with Perceptron
  Hinge Loss

Expon. gradient
  with Normalized
  Hinge Loss

Known linear threshold algorithms for $\pm 1$-classification case are gradient-based algorithms with hinge loss
Perceptron

\[ w_{t+1} \]

\[ = \arg\min_{\mathbf{w}} \left( \| \mathbf{w} - \mathbf{w}_t \|^2 / 2 + \eta HL(\mathbf{w} \cdot \mathbf{x}_t, g^{-1}(y_t)) \right) \]

\[ = \mathbf{w}_t - \eta \left( \text{sign}(\mathbf{w}_{t+1} \cdot \mathbf{x}_t) - y_t \right) \mathbf{x}_t \]

\[ \approx \mathbf{w}_t - \eta \left( \text{sign}(\mathbf{w}_t \cdot \mathbf{x}_t) - y_t \right) \mathbf{x}_t \]
Normalized Winnow

\[ w_{t+1} = \arg\min_w \left( \sum_{i=1}^{n} w_i \ln \frac{w_i}{w_{t,i}} + \eta HL(w \cdot x_t, g^{-1}(y_t)) \right) \]

\[ = w_{t,i} e^{-\eta (\text{sign}(w \cdot x_t) - y_t) x_{t,i}} / \text{normalization} \]

\[ \approx w_{t,i} e^\hat{y}_t / \text{normalization} \]
Trade-off between two divergences [KW]

\[ w_{t+1} = \arg\min_w (\Delta F(w, w_t) + \eta_t \Delta_H(w \cdot x_t, h^{-1}(y_t))) \]

parameter divergence + matching loss divergence

Both divergences are convex in \( w \)

\[ w_{t+1} = f^{-1}(f(w_t) - \eta_t(h(w_t \cdot x_t) - y_t)x_t) \]

Generalization of the “delta”-rule