Two Types of Setup

**Regression**
- Examples are tuples \((x_t, a_t)\)
  - \(x_t \in \mathbb{R}^n\): data point (example)
  - \(a_t \in \mathbb{R}\): true concentration (activity)
- Linear activation (estimate): \(\hat{a}_t = w \cdot x_t\)

**Classification**
- Examples are tuples \((x_t, y_t)\)
  - \(x_t \in \mathbb{R}^n\): data point (example)
  - \(y_t \in [0, 1]\): true probability (label)
- Probability (label) estimate: \(\hat{y}_t = h(w \cdot x_t)\)
Why are we doing this?

- Want to design good loss functions for a given problem
- Loss functions should be steep in important areas and flat in unimportant areas
- Asymmetric loss functions
- Flexible method
Clarke Grid Type Loss
Designing Loss for Clarke Grid

**Goal:** Accurately predict glucose levels of people with diabetes

- Non-symmetric loss: Assymetry is needed ...
- Low concentrations more important than high ones
- Loss defined by the Clarke Grid
- How do you optimize such loss?
Single Neuron

\[
\hat{y} = h(\hat{a}) \\
y = h(a)
\]

Post: \( \text{Loss}(y, \hat{y}) \)

\[
\hat{a} = w \cdot x \\
a = h^{-1}(y)
\]

Pre: \( \text{Loss}(a, \hat{a}) \)
Matching Loss in Pre Domain

\[ \Delta_h(\hat{a}, a) = \int_a^{\hat{a}} (h(z) - h(a)) \, dz \]
Matching Loss as Bregman Divergence

\[ \Delta_h(\hat{a}, a) = \int_a^{\hat{a}} (h(z) - h(a)) \, \partial z \]

- **Square Loss:** \( h(z) = z \)
  \[ \Delta_h(\hat{a}, a) = \frac{1}{2} (\hat{a} - a)^2 \]

- **Logistic Loss:** \( h(z) = \frac{e^z}{1 + e^z} \)
  \[ \Delta_h(\hat{a}, h^{-1}(y)) = \ln(1 + e^{\hat{a}}) + y \ln y + (1 - y) \ln(1 - y) - y \hat{a} \]
Dual View of Matching Loss

$$\Delta_h(\hat{a}, a) = \int_a^{\hat{a}} (h(z) - h(a)) \, dz$$  \hspace{1cm} \text{Pre}$$

$$h(z) = p \quad h^{-1}(p) = z$$
$$dz = (h^{-1}(p))' \, dp$$

Integ. by parts

$$= \int_{h(a)}^{h(\hat{a})} (p - h(a)) (h^{-1}(p))' \, dp$$

$$= (h(\hat{a}) - h(a)) h^{-1}(h(\hat{a})) - \int_{h(a)}^{h(\hat{a})} (h^{-1}(p)) dp$$

$$= \int_{h(\hat{a})}^{h(a)} (h^{-1}(p) - h^{-1}(h(\hat{a}))) \, dp$$

$$= \Delta_{h^{-1}}(h(a), h(\hat{a}))$$  \hspace{1cm} \text{Post}$$
Two domains

- **Pre**: linear activations $\to (a, \hat{a})$

  \[
  \Delta_h(\hat{a}, a) = \int_a^{\hat{a}} (h(z) - h(a)) \, dz
  \]

  **Regression**: labels are $a$.

- **Post**: estimation of probability $\to (y, h(\hat{a}))$

  \[
  \Delta_{h^{-1}}(y, h(\hat{a})) = \int_{h(\hat{a})}^{y} \left( h^{-1}(p) - h^{-1}(h(\hat{a})) \right) \, dp
  \]

  **Classification**: labels are $y = h(a)$.
Why are we doing this?

- Want to design good matching losses given a problem
- Post loss:
  - Insensitive to shifting
  - Loss is not fixed
- Pre loss:
  - Very sensitive to shifting
  - Loss is fixed by the transfer function
  - Allows for design of “fancy” losses
Define the transfer function $h(\hat{a})$ as

$$h(\hat{a}) = \frac{e^{\alpha(w \cdot x + \beta)}}{1 + e^{\alpha(w \cdot x + \beta)}},$$

where $\alpha$ scales the sigmoid and $\beta$ shifts it.

For the Clarke Grid type loss use piece of sigmoid that puts more weight on the smaller activations.
In the bottom row we plot the $\Delta(\hat{a}, a)$ as a function of the estimate $\hat{a}$ for fixed activities $a = -3, 0, 3$. Note that locally the losses are quadratic and the steepness of the bowl is determined by $h'(a)$. 
3D View of the Loss

Regular Sigmoid, Left Piece of Sigmoid, Right Piece of Sigmoid
General Shift Formulas

\[ h_{\alpha,\beta}(a) = h(\alpha(a + \beta)) \]  \hspace{1cm} (1)
\[ h_{\alpha,\beta}\left(\frac{1}{\alpha}a - \beta\right) = h(a) \]  \hspace{1cm} (2)
\[ h_{\alpha,\beta}^{-1}(y) = \frac{1}{\alpha}h^{-1}(y) - \beta \]  \hspace{1cm} (3)
\[ h_{\alpha,\beta}^{-1}\left(\frac{1}{\alpha}a - \beta\right) = h(a) \]  \hspace{1cm} (4)
\[ h_{\alpha,\beta}(h_{\alpha,\beta}^{-1}(y)) \overset{(5)}{=} h_{\alpha,\beta}\left(\frac{1}{\alpha}h^{-1}(y) - \beta\right) \overset{(6)}{=} h(h^{-1}(y)) = y \]
\[ h_{\alpha,\beta}^{-1}(h_{\alpha,\beta}(a)) \overset{(3)}{=} h_{\alpha,\beta}^{-1}(h(\alpha(a + \beta))) \overset{(5)}{=} \frac{1}{\alpha}(h^{-1}(h(\alpha(a + \beta)))) - \beta = a \]
Logistic Regression ($y \in [0, 1]$)

$$\Delta_{h_{\alpha, \beta}^{-1}} (y, h_{\alpha, \beta}(\frac{1}{\hat{\alpha}} - \beta)) = \int_{h(\hat{\alpha})}^{y} \left( h_{\alpha, \beta}^{-1}(p) - h_{\alpha, \beta}^{-1}(h_{\alpha, \beta}(\frac{1}{\hat{\alpha}} - \beta)) \right) dp$$

$$= \int_{h(\hat{\alpha})}^{y} \left( \frac{1}{\alpha} h^{-1}(p) - \frac{1}{\alpha} \hat{\alpha} - \beta \right) dp$$

$$= \frac{1}{\alpha} \int_{h(\hat{\alpha})}^{y} \left( h^{-1}(p) - h^{-1}(h(\hat{\alpha})) \right) dp$$

$$= \frac{1}{\alpha} \Delta_{h^{-1}} (y, h(\hat{\alpha}))$$
Another view

\[
\Delta_{h_{\alpha,\beta}} \left( \frac{1}{\alpha} \hat{a} - \beta, \ h_{\alpha,\beta}^{-1}(y) \right) = \int_{h_{\alpha,\beta}(y)}^{\frac{1}{\alpha} \hat{a} - \beta} (h_{\alpha,\beta}(z) - h_{\alpha,\beta}(h_{\alpha,\beta}^{-1}(y))) \, dz
\]

\[
z = \frac{1}{\alpha} v - \beta \quad \Rightarrow \\
= \frac{1}{\alpha} \int_{h^{-1}(y)}^{\hat{a}} (h_{\alpha,\beta}(\frac{1}{\alpha} v - \beta) - y) \, dv
\]

\[
= \frac{1}{\alpha} \int_{h^{-1}(y)}^{\hat{a}} (h(v) - y) \, dv
\]

\[
= \frac{1}{\alpha} \Delta_h(\hat{a}, h^{-1}(y))
\]
Why are we doing this?

- Want to design good matching losses given a problem
- Post loss:
  - Insensitive to shifting
  - Loss is not fixed
- Pre loss:
  - Very sensitive to shifting
  - Loss is fixed by the transfer function
  - Allows for design of “fancy” losses