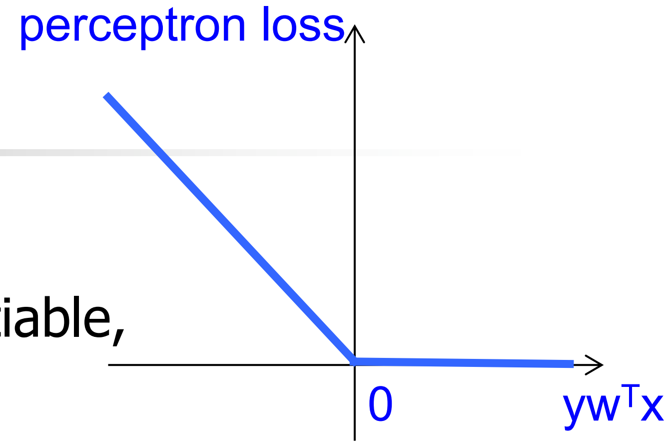
CMSC 510 – L06
Regularization Methods for Machine Learning

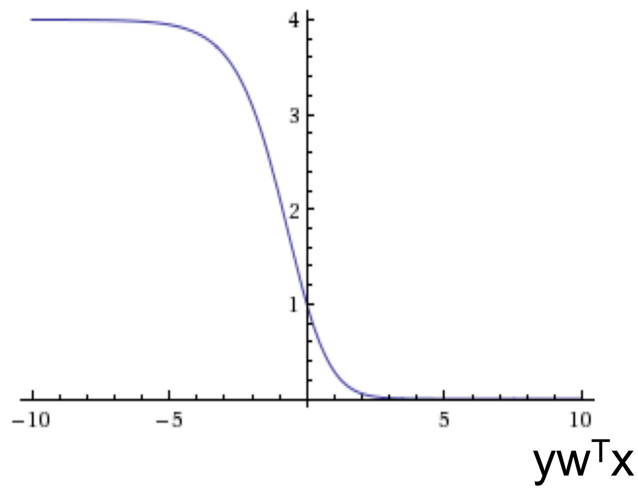
Instructor:
Dr. Tom Arodz

Recap



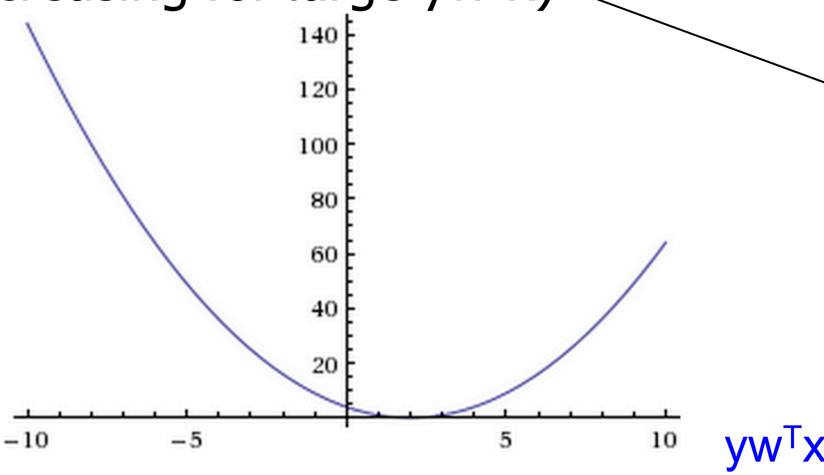
- Perceptron loss
  - convex, non-increasing, but non-differentiable, and w=0 is always a global minimum
- Quadratic loss on sigmoid activation
  - differentiable, non-increasing but non-convex => local minima
- LDA: quadratic loss directly on yw^Tx
  - convex, differentiable, but non-monotonic (increasing for large yw^Tx)

delta rule loss



This wasn't a problem for quadratic over sigmoid

Square loss

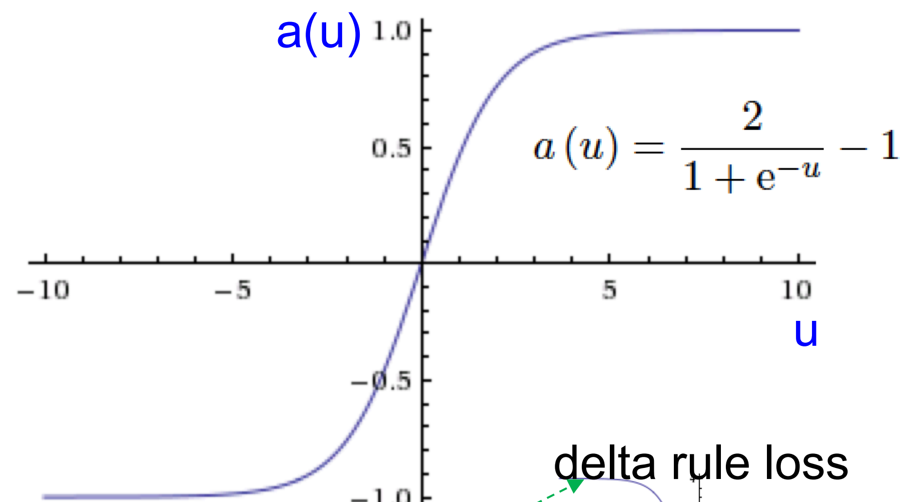


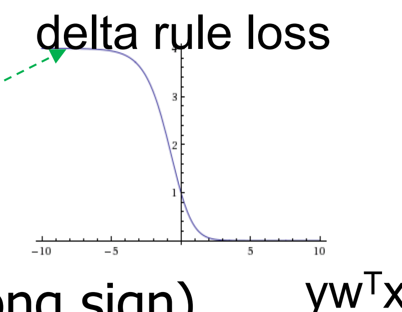
Differentiable perceptron $h(x)=a(w^Tx)$

- Another problem with delta rule:

$$\mathbf{w}^{\dagger}{}_{t+1} = \mathbf{w}^{\dagger}{}_{t} - \frac{c}{2}\,\frac{\partial \ell\left(h,\mathbf{z}\right)}{\partial \mathbf{w}^{\dagger}}\bigg|_{\mathbf{w}^{\dagger}=\mathbf{w}^{\dagger}{}_{t}} = \mathbf{w}^{\dagger}{}_{t} + c\left[y - a\left(\mathbf{w}^{\dagger}{}_{t}^{T}\mathbf{x}^{\dagger}\right)\right] a'\left(\mathbf{w}^{\dagger}{}_{t}^{T}\mathbf{x}^{\dagger}\right)\mathbf{x}^{\dagger}$$

- The update depends on:
  - (y-a(u))
    - Smaller as we approach correct prediction
  - a’(u)
    - Much smaller as we approach correct prediction (*u* v.large)
    - And also very small when we have really incorrect prediction (*u* high magnitude, wrong sign)





- Very slow learning for large |u|

Yet another loss

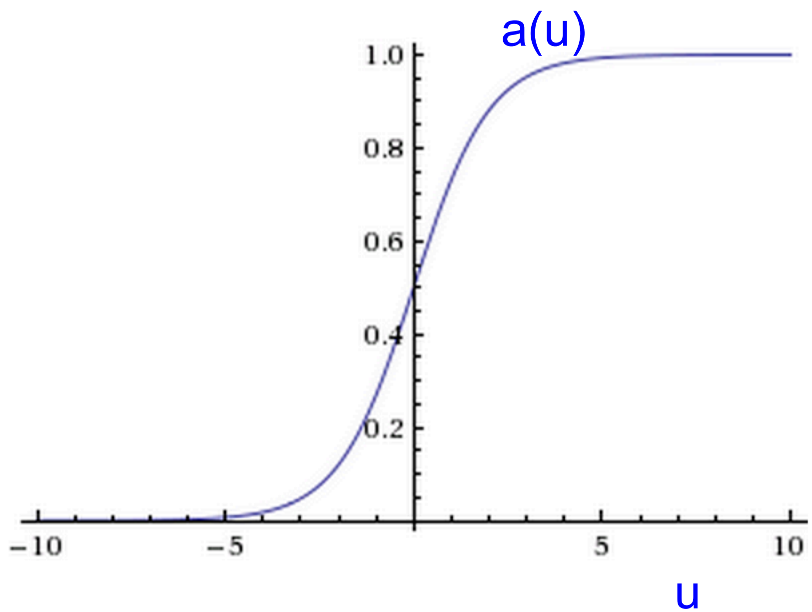
- Let's try with a unipolar sigmoid activation function
  - Same as bipolar sigmoid, except return [0,1] not [-1,1]

$$a(u) = \frac{1}{1+\mathrm{e}^{-u}}$$

$$a(-u) = 1 - a(u)$$

$$a'(u) = a(u)(1 - a(u))$$

Sigmoid activation function:
 h(x) = +1 if $a(w^T x) > 0.5$
 h(x) = –1 if $a(w^T x) < 0.5$



Differentiable perceptron $h(x)=a(w^T x)$

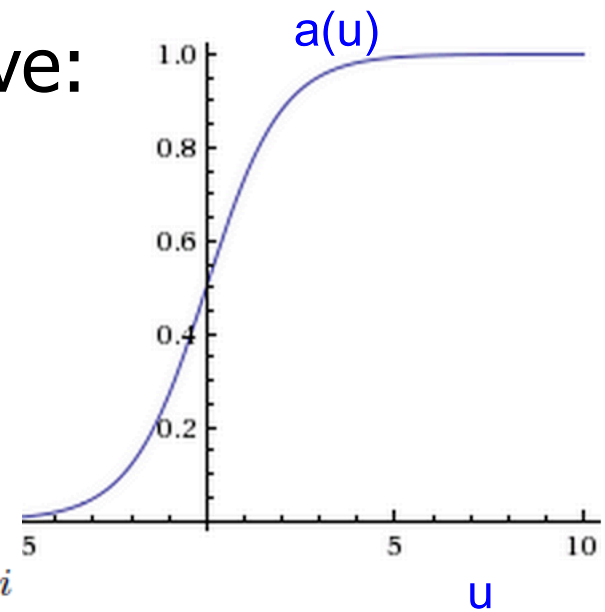
$$a'(u) = a(u)(1-a(u))$$

- Another problem with delta rule:

$$\mathbf{w}^{\dagger}{}_{t+1} = \mathbf{w}^{\dagger}{}_{t} - \frac{c}{2}\frac{\partial \ell\left(h,\mathbf{z}\right)}{\partial \mathbf{w}^{\dagger}}\Big|_{\mathbf{w}^{\dagger}=\mathbf{w}^{\dagger}{}_{t}} = \mathbf{w}^{\dagger}{}_{t} + c\left[y - a\left(\mathbf{w}^{\dagger}{}_{t}^{T}\mathbf{x}^{\dagger}\right)\right]\cancel{a'\left(\mathbf{w}^{\dagger}{}_{t}^{T}\mathbf{x}^{\dagger}\right)}\,\mathbf{x}^{\dagger}$$

- Let’s get rid of the a’ term and have:

$$\boxed{-\frac{\partial \ell(a,y)}{\partial w_i} = (y-a)x_i}$$



- Our loss would have to be:

$$\frac{\partial \ell}{\partial w_i} = \frac{\partial \ell(a,y)}{\partial a}\frac{\partial a(u)}{\partial u}\frac{\partial w^T x}{\partial w_i}$$

$$\frac{\partial \ell(a,y)}{\partial w_i} = \frac{\partial \ell(a,y)}{\partial a}\frac{\partial a(u)}{\partial u}\frac{\partial w^T x}{\partial w_i} = \frac{\partial \ell(a,y)}{\partial a}a(u)(1-a(u))x_i$$

$$\boxed{\frac{\partial \ell(a,y)}{\partial a} = -\frac{(y-a)}{a(1-a)}}$$

- That is: $\frac{\partial \ell(a,1)}{\partial a} = -\frac{(1-a)}{a(1-a)} = -\frac{1}{a}$ $\qquad \frac{\partial \ell(a,0)}{\partial a} = -\frac{(0-a)}{a(1-a)} = -\frac{-1}{1-a}$

Differentiable perceptron $h(x)=a(w^T x)$

- Let's try to find a loss without a' term in gradient

$$\boxed{-\frac{\partial \ell(a,y)}{\partial w_i} = (y-a)x_i}$$

- Our loss would have to be: $\frac{\partial \ell(a,y)}{\partial a} = -\frac{(y-a)}{a(1-a)}$

$$\frac{\partial \ell(a,1)}{\partial a} = -\frac{(1-a)}{a(1-a)} = -\frac{1}{a} \qquad \frac{\partial \ell(a,0)}{\partial a} = -\frac{(0-a)}{a(1-a)} = -\frac{-1}{1-a}$$

$$\frac{\partial -\log(a)}{\partial a} = -\frac{1}{a}$$

$$\frac{\partial -\log(1-a)}{\partial a} = \frac{\partial -\log(1-a)}{\partial 1-a}\frac{\partial(1-a)}{\partial a} = -\frac{-1}{1-a}$$

- In a form of a single equation:

$$\ell(a,y) = -[y\log(a) + (1-y)\log(1-a)]$$

y=1 y=0

Differentiable perceptron $h(x)=a(w^T x)$

- Let's try to find a loss without a' term in gradient

$$\boxed{-\frac{\partial \ell(a,y)}{\partial w_i} = (y-a)x_i}$$

- Our loss would have to be:

$$\ell(a,y) = -[y \log(a) + (1-y)\log(1-a)]$$

- Or in another form:

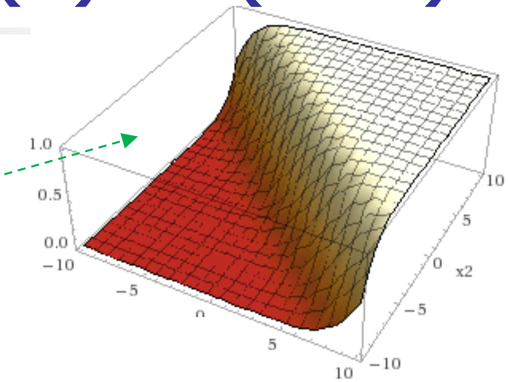
$$\ell(a,1) = -\log\left(\frac{1}{1+e^{-w^T x}}\right) = \log(1+e^{-w^T x})$$

$$\ell(a,0) = -\log\left(1-\frac{1}{1+e^{-w^T x}}\right) = -\log\left(\frac{e^{-w^T x}}{1+e^{-w^T x}}\right) = \log(1+e^{w^T x})$$

- If we go back to y being +1 or -1:

$$\ell(a,y) = \log(1+e^{-yw^T x})$$

Differentiable perceptron $h(x)=a(w^Tx)$



- Our loss on result of $a(w^Tx)$:

$$\ell(a,y) = -[y\log(a) + (1-y)\log(1-a)]$$

- We can rewrite it as:

$$P_{\text{data}}(y=1|x) = y \qquad P_{\text{data}}(y=0|x) = 1-y$$

$$P_{\text{model}}(y=1|x) = a(w^Tx) \qquad P_{\text{model}}(y=0|x) = 1-a(w^Tx)$$

$$\ell(a,y) = -\sum_{i=\{0,1\}} P_{\text{data}}(y=i|x)\log P_{\text{model}}(y=i|x)$$

- P_{data} is a distribution over y={0,1}, and takes values {0,1}
  - a sample has 100% of class 1, or 100% of class 0
- For P_{model}, we see a term $-\log_2 p_i = \log_2 1/p_i$
  - What is the meaning of this term?

Information theory

- Surprise $I(P_A)$ associated with seeing an event A that is supposed to happen with probability P_A:
  - Certain = no surprise: $P_A=1 \Rightarrow I(P_A)=0$
  - Lower probability => higher surprise: $P_A < P_B \quad \Rightarrow \quad I(P_A) > I(P_B)$
- Surprise $I(P_{A+B})$ associated with seeing events A and B that are supposed to happen with probabilities P_A, P_B
  - $I(P_{A+B}) \Rightarrow I(P_A)+I(P_B)$ if events A, B are independent
    - $P_{A+B}=P_AP_B \quad \Rightarrow \quad I(P_AP_B)=I(P_A)+I(P_B)$
  - What form can the function $I(P)$ take?
    - Is 0 for argument of 1
    - Is decreasing
    - Turns multiplication into addition

Information theory

- Surprise $I(P_A)$ associated with seeing an event A that is supposed to happen with probability P_A:
  - $P_A=1 \Rightarrow I(P_A)=0$
- Lower probability => higher surprise
  - $P_A < P_B \Rightarrow I(P_A) > I(P_B)$
- Surprise $I(P_{A+B})$ associated with seeing events A and B that are supposed to happen with probabilities P_A, P_B
  - $I(P_{A+B}) \Rightarrow I(P_A)+I(P_B)$ if events A, B are independent
    - $P_{A+B}=P_AP_B \Rightarrow I(P_AP_B)=I(P_A)+I(P_B)$
  - What form can the function I(P) take?
    - $I(p)=-\log(p)$
      - $-\log(1)=0$
      - $-\log(ab)=-\log(a) + -\log(b)$
      - -log is decreasing, since log is increasing
    - By convention, in CS we take $\log_2$

Information theory

- Interpretation of $-\log_2 p_i = \log_2 1/p_i$
- Measure of surprise: $I(p)=-\log_2(p)$
- We see event with probability p_i – how surprised are we?
  - Sun rising in the morning: $p_i = 1$, surprise = 0
  - Heads after a fair coin toss: $p_i = 1/2$, surprise = 1 bit
  - Two heads from two fair coins: $p_i = 1/2*1/2$, surprise = 2 bits
  - "3" on a 4-sided dice: $p_i = 1/4$, surprise = 2 bits
  - Once-in-a-hundred-years heat-wave: $p_i = 1/100$, surprise = 6.64 bits
  - Win on a "1:1,048,576 chance" lottery, $p_i = 2^{-20}$, surprise = 20 bits
  - Sun NOT rising in the morning: $p_i = 0$, surprise = ∞ bits

Information theory

0 log0 =0
log is log2

- **Entropy** of distribution *p*:

$$H(p) = E_p[\log \frac{1}{p}] = E_p[-\log p] = -\sum_i p_i \log p_i$$

  - **Expected surprise** from samples from distr. p
    - E.g. $p_0=p_1=1/2$ high surprise H=1, $p_0=1$ no surprise H=0
- Example: hotter than median summer
  - heat wave $p_1=1/2$ (surprise=1bit) or not $p_0=1/2$ (surprise=1bit)
  - We observe 128,000 summers, we get 64,000 heat waves
    - H(p)=H(p,p)=-1/2 log2(1/2) - 1/2 log2(1/2) = 1
  - Highest possible entropy = expected surprise for "this-or-that"
- Example: once-in-a-(CS)-century heat wave
  - heat wave $q_1=1/128$ (surprise=7bits) or not $q_0=127/128$ (surprise=0.011 bits)
  - We observe 128,000 summers, we get 1,000 heat waves
    - H(q)=H(q,q)=-1/128 log2(1/128) - 127/128 log2(127/128) = 0.0659
  - Much lower entropy = expected surprise

Information theory

0 log0 =0
log is log2

- Entropy of distribution *p*:

$$H(p) = E_p[\log \tfrac{1}{p}] = E_p[-\log p] = -\sum_i p_i \log p_i$$

  - Amount of surprise from samples from distr. P

- Cross-entropy of *p* and *q*:

$$H(p,q) = E_p[-\log q] = -\sum_i p_i \log q_i$$

  - Surprise when assuming samples came from distrib. q (use q to calc. surprise of each even)
    but they actually come from distrib. p (use p to calc. the mean/expect.)

- Example: climate change
  - heat wave q_1=1/128 (surprise=7bits) or not q_0=127/128 (surprise=0.011 bits)
  - We expect that if we see 128,000 summers, we get 1,000 heat waves
    - H(q)=H(q,q)=-1/128 log2(1/128) - 127/128 log2(127/128) = 0.0659
  - We observe 128,000 summers, we get 64,000 heat waves
    - So we have observed p_1=1/2 p_0=1/2
    - Our expectation was that it came from q_1=1/128 q_0=127/128
    - H(p,q) = -1/2 log2(1/128) – 1/2 log2(127/128) = 3.5 >> 0.0659 =H(q)

Information theory

0 log0 =0
log is log2

- Entropy of distribution *p*:

$$H(p) = E_p[\log \frac{1}{p}] = E_p[-\log p] = -\sum_i p_i \log p_i$$

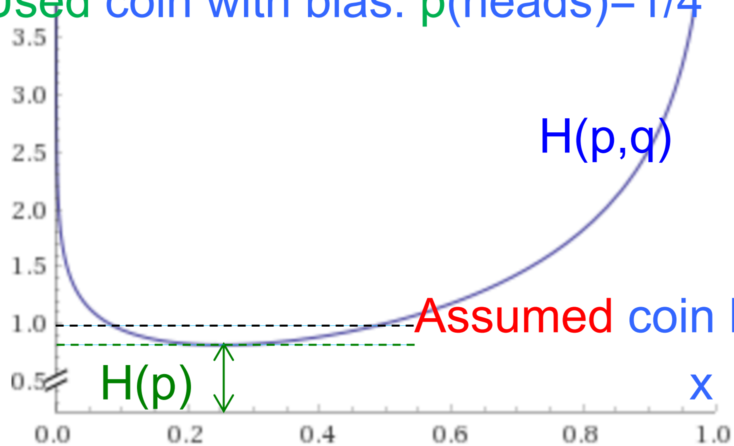
  - Amount of surprise from samples from distr. p
    - E.g. $p_0=p_1=1/2$ high surprise H=1, $p_0=1$ no surprise H=0

- Cross-entropy of *p* and *q*:

$$H(p,q) = E_p[-\log q] = -\sum_i p_i \log q_i$$

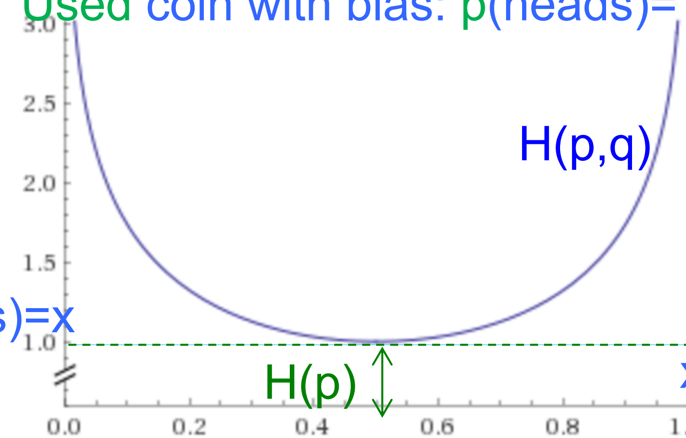
  - Surprise when assuming samples came from distrib. q but they actually come from distrib. p

Used coin with bias: p(heads)=1/4



Assumed coin bias with q(heads)=x

Used coin with bias: p(heads)=1/2



Information theory

0 log0 =0
log is log2

- **Entropy** of distribution *p*:

$$H(p) = E_p[\log \frac{1}{p}] = E_p[-\log p] = -\sum_i p_i \log p_i$$

  - Amount of surprise from samples from distr. p

- **Cross-entropy** of *p* and *q*:

$$H(p,q) = E_p[-\log q] = -\sum_i p_i \log q_i$$

  - Surprise when assuming samples came from distrib. q but they actually come from distrib. p

- **Relative entropy** of *q* w.r.t. to *p* (from *p* to *q*):

$$D_{KL}(p||q) = -E_p[\log \frac{q}{p}] = E_p[\log p] + E_p[-\log q] = H(p,q) - H(p)$$

a.k.a **Kullback-Leibler divergence** of *q* from *p*

  - **Additional** surprise when assuming samples came from distrib. q but they actually come from distrib. p

Cross-entropy loss

0 log0 =0
log is log2

- Cross-entropy of *p* and *q*:

$$H(p,q) = E_p[-\log q] = -\sum_i p_i \log q_i$$

- Kullback-Leibler divergence of *q* from *p*

$$D_{KL}(p||q) = -E_p[\log \frac{q}{p}] = E_p[\log p] + E_p[-\log q] = H(p,q) - H(p)$$

- If p_i has 1 for one i, and 0 for all other i, H(p)=0
  - Cross-entropy = KL-divergence
    (additional surprise is all surprise there is)
- Loss = $H(P_{data}, P_{model}) = KL_{div}(P_{data} || P_{model})$
  - We observe data where class is certain
    (e.g. *P(class 1) for given x* = 1 or 0)
  - Model says data came from distrib.
    $a(w^Tx)$ = *P(class 1) for given x* is
- How surprised we are by data (y) if model y=h(x) was true?

Cross-entropy loss

- Can we extend it to more than 2 classes?
- Easy:

$$\ell(a, y) = -\sum_{i=\{0,1\}} P_{\text{data}}(y = i|x) \log P_{\text{model}}(y = i|x$$

i=1 to num_classes

$a(w_1^T x) =$ *P(class 1 | x)*
$a(w_2^T x) =$ *P(class 2 | x)*
$a(w_3^T x) =$ *P(class 3 | x)*
$a(w_4^T x) =$ *P(class 4 | x)*
…

- But: $a(w_1^T x)$, $a(w_2^T x)$, $a(w_3^T x)$, … must be a probability distribution over classes
  - >=0
  - Add up to 1

Soft-max

$$\ell(a, y) = -\sum_{i=\{0,1\}} P_{\text{data}}(y = i|x) \log P_{\text{model}}(y = i|x)$$

i=1 to num_classes

$a(w_1^T x) =$ *P(class 1 | x)*
$a(w_2^T x) =$ *P(class 2 | x)*
$a(w_3^T x) =$ *P(class 3 | x)*
$a(w_4^T x) =$ *P(class 4 | x)*
...

- But: $a(w_1^T x)$, $a(w_2^T x)$, $a(w_3^T x)$, …
  must be a probability distribution over classes
- Soft-max:
  - $a(w_1^T x) = \exp(w_1^T x) / \Sigma_k \exp(w_k^T x)$
    - exp(…) makes values > 0
    - Then we normalize to add up to 1

Cross-entropy loss

- For y in {0,1} - it's called: cross-entropy loss

$$\ell(a,y) = \log(1 + e^{-yw^T x})$$

$$P_{\text{data}}(y=1|x) = y \qquad P_{\text{data}}(y=1|x) = 1-y$$

$$P_{\text{model}}(y=1|x) = a(w^T x) \qquad P_{\text{model}}(y=0|x) = 1 - a(w^T x)$$

- Can we extend it to more than 2 classes?

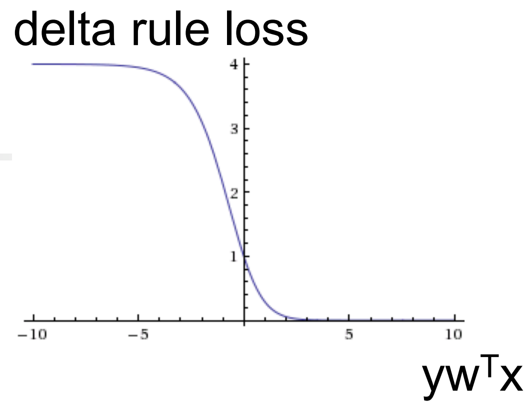
$$\ell(a,y) = -\sum_{i=\{0,1\}} P_{\text{data}}(y=i|x) \log P_{\text{model}}(y=i|x)$$

- Easy:

i=1 to num_classes

$a(w_1^T x) =$ *P(class 1 | x)*
$a(w_2^T x) =$ *P(class 2 | x)*
$a(w_3^T x) =$ *P(class 3 | x)*
$a(w_4^T x) =$ *P(class 4 | x)*
…

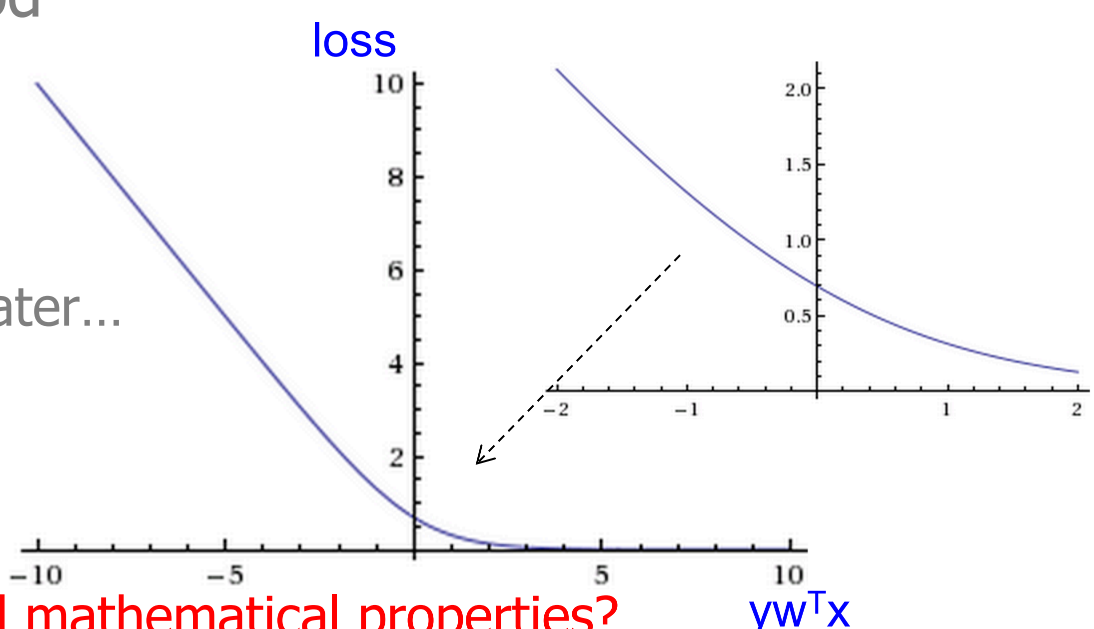
Logistic regression



- Logistic loss:

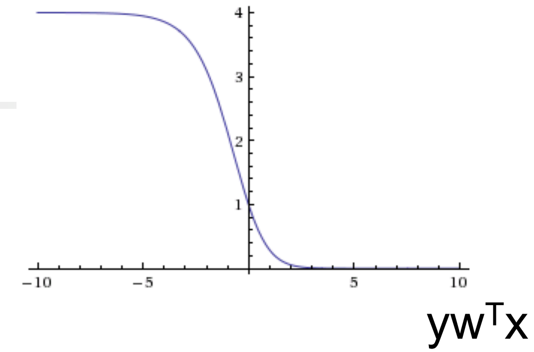
$$\ell(h, z) = \ln(1 + \mathrm{e}^{-yw^T x})$$

- Derived from: $a(u) = \frac{1}{1+\mathrm{e}^{-u}}$
  - Cross-entropy loss over $a(w^Tx)$
  - Maximum likelihood estimate for $P(y|x,w) = a(w^Tx)$
    - We will see that later…



  - Does it have good mathematical properties?

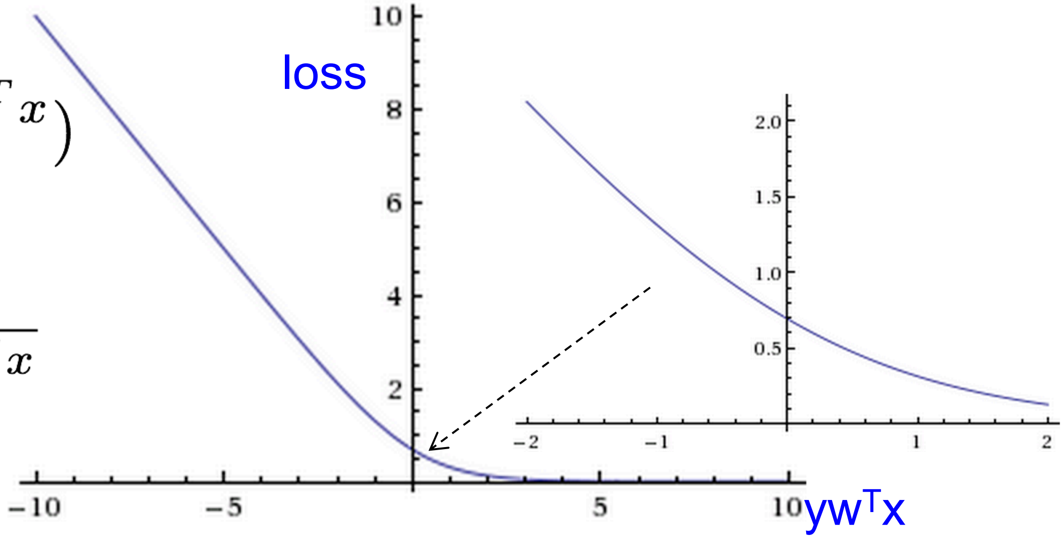
Logistic regression



- Let's try another loss function:
  - Logistic loss

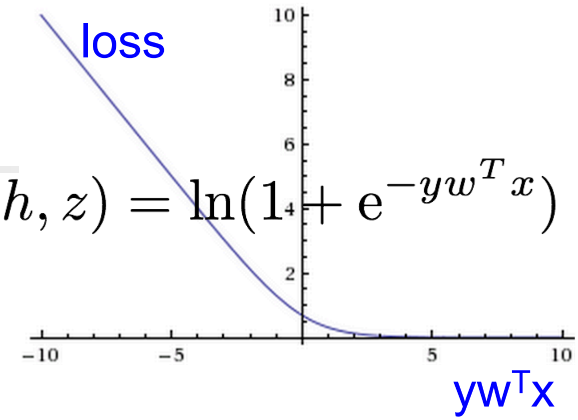
$$\ell(h,z) = \ln(1 + \mathrm{e}^{-yw^Tx})$$

$$\nabla_w \ell(h,z) = \frac{yx}{1 + \mathrm{e}^{yw^Tx}}$$



- Differentiable: gradient $\nabla_{\mathbf{w}}$ easy to calculate
- Constant derivative/step size for highly incorrect predictions
- Doesn't go to 0 too quickly for correct predictions
- Monotonic, non-increasing, no big penalty for correct predictions
- Convex: no local minima, gradient descent will converge towards global minimum of empirical risk
- w=0 ("no prediction") is rarely a global minimum

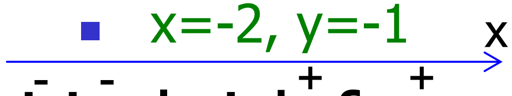
Logistic regression



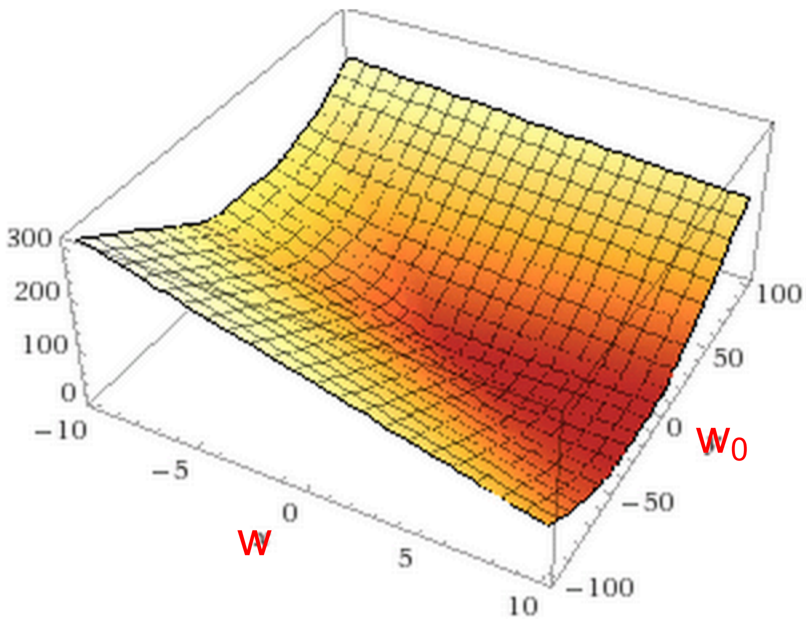
$$\ell(h,z) = \ln(1 + e^{-yw^Tx})$$

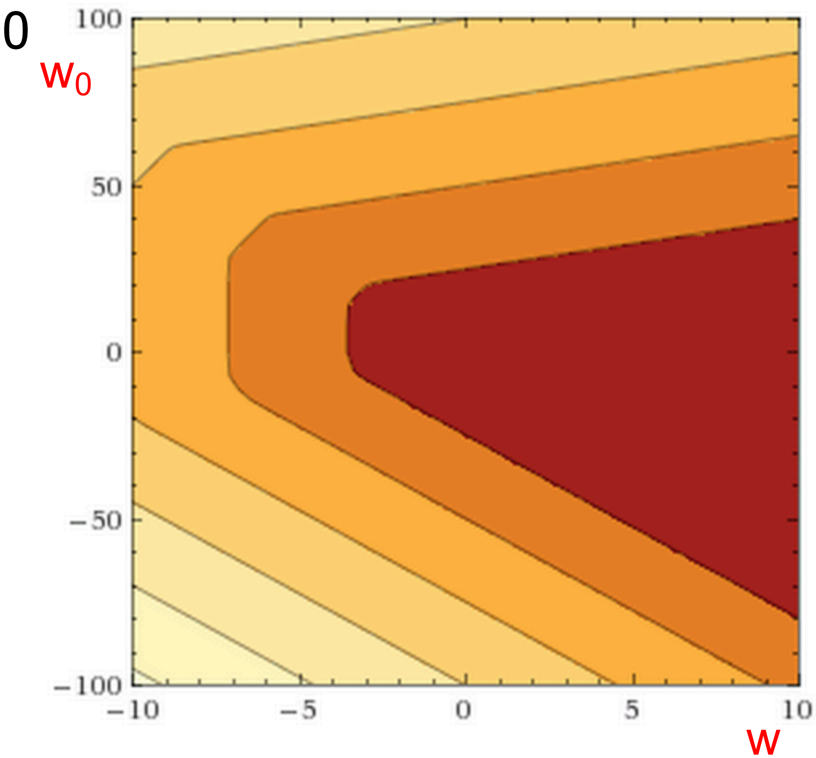
- Logistic loss - Four samples:
  - x=7, y=1
  - x=4, y=1
  - x=-1, y=-1
  - x=-2, y=-1

$$\min \frac{1}{m}\sum_{i=1}^{m} \ln(1 + e^{-y_i w^T x_i})$$



- Empirical risk for any w, w_0





- No local minima!