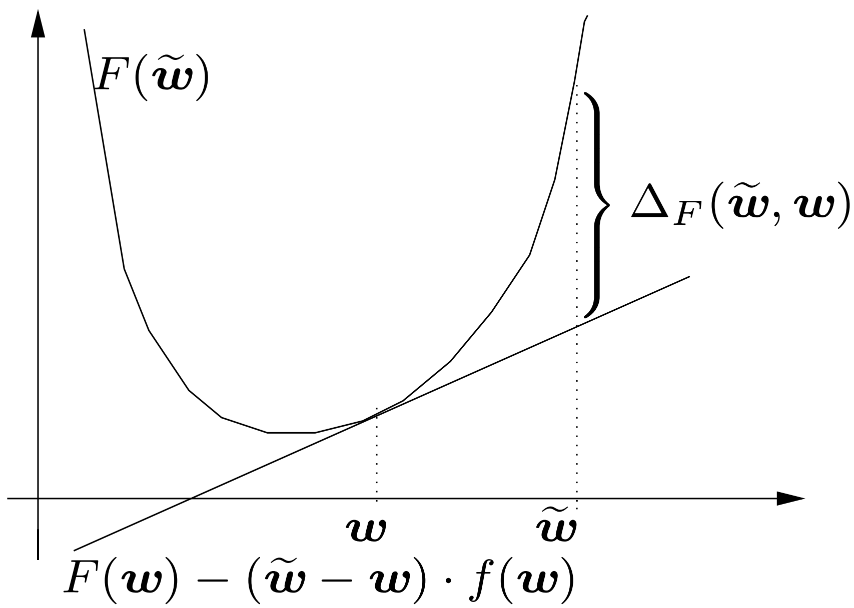
Bregman Divergences [Br,CL,Cs]

For any differentiable convex function F

$$\Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) = F(\widetilde{\boldsymbol{w}}) - F(\boldsymbol{w}) - (\widetilde{\boldsymbol{w}} - \boldsymbol{w}) \cdot \underbrace{\nabla_{\boldsymbol{w}} F(\boldsymbol{w})}_{f(\boldsymbol{w})}$$

$$= F(\widetilde{\boldsymbol{w}}) - \begin{array}{c}\text{supporting hyperplane}\\ \text{through } (\boldsymbol{w}, F(\boldsymbol{w}))\end{array}$$



Bregman Divergences: Simple Properties

1. $\Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w})$ is convex in $\widetilde{\boldsymbol{w}}$
2. $\Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) \geq 0$
   If F convex equality holds iff $\widetilde{\boldsymbol{w}} = \boldsymbol{w}$
3. Usually not symmetric: $\Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) \neq \Delta_F(\boldsymbol{w}, \widetilde{\boldsymbol{w}})$
4. Linearity (for $a \geq 0$):
   $\Delta_{F+a\,H}(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) = \Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) + a\,\Delta_H(\widetilde{\boldsymbol{w}}, \boldsymbol{w})$
5. Unaffected by linear terms ($a \in \mathbf{R}$, $\boldsymbol{b} \in \mathbf{R}^n$):
   $\Delta_{H+a\widetilde{\boldsymbol{w}}+\boldsymbol{b}}(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) = \Delta_H(\widetilde{\boldsymbol{w}}, \boldsymbol{w})$

Bregman Divergences: more properties

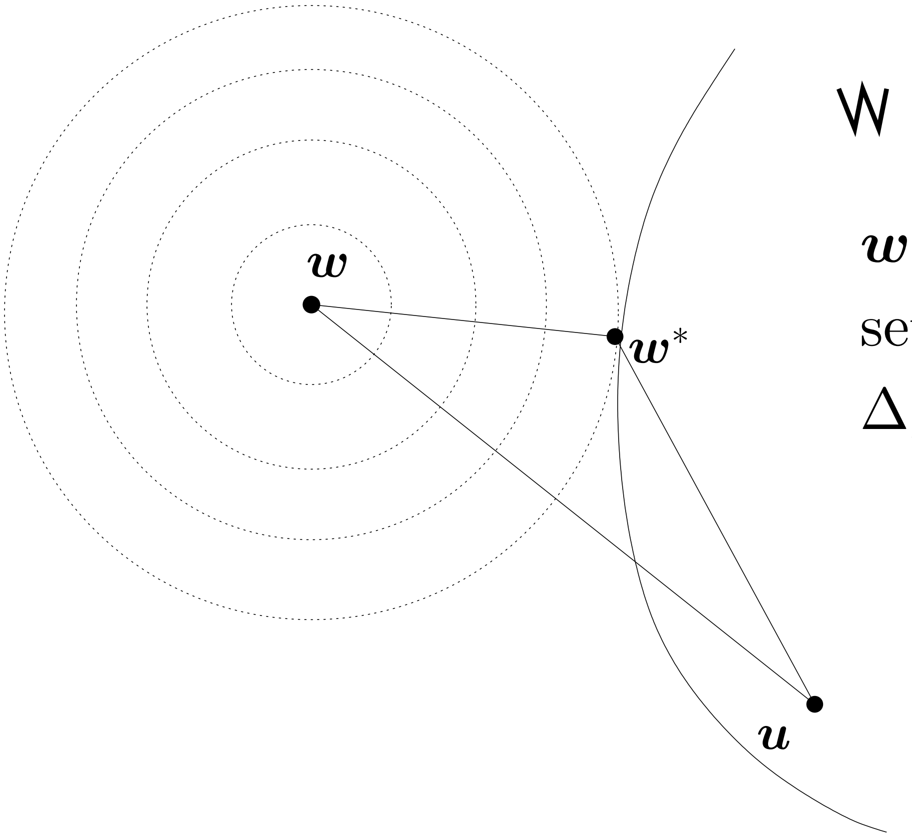
6. $\nabla_{\widetilde{\boldsymbol{w}}} \Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w})$

$$
\begin{aligned}
&= \nabla F(\widetilde{\boldsymbol{w}}) - \nabla_{\widetilde{\boldsymbol{w}}}(\widetilde{\boldsymbol{w}} \nabla_{\boldsymbol{w}} F(\boldsymbol{w})) \\
&= f(\widetilde{\boldsymbol{w}}) - f(\boldsymbol{w})
\end{aligned}
$$

7. $\Delta_F(\boldsymbol{w}_1, \boldsymbol{w}_2) + \Delta_F(\boldsymbol{w}_2, \boldsymbol{w}_3)$

$$
\begin{aligned}
= \quad & F(\boldsymbol{w}_1) - F(\boldsymbol{w}_2) - (\boldsymbol{w}_1 - \boldsymbol{w}_2) f(\boldsymbol{w}_2) \\
& F(\boldsymbol{w}_2) - F(\boldsymbol{w}_3) - (\boldsymbol{w}_2 - \boldsymbol{w}_3) f(\boldsymbol{w}_3) \\
= \quad & \Delta_F(\boldsymbol{w}_1, \boldsymbol{w}_3) + (\boldsymbol{w}_1 - \boldsymbol{w}_2) \cdot (f(\boldsymbol{w}_3) - f(\boldsymbol{w}_2))
\end{aligned}
$$

A Pythagorean Theorem [Br,Cs,A,HW]



$\boldsymbol{w}^*$ is projection of $\boldsymbol{w}$ onto convex set $\mathcal{W}$ w.r.t. Bregman divergence Δ_F:

$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{u}\in\mathcal{W}} \Delta_F(\boldsymbol{u}, \boldsymbol{w})$$

Theorem:

$$\Delta_F(\boldsymbol{u}, w) \geq \Delta_F(\boldsymbol{u}, \boldsymbol{w}^*) + \Delta_F(\boldsymbol{w}^*, \boldsymbol{w})$$

Examples

Squared Euclidean Distance

$$
\begin{aligned}
F(\boldsymbol{w}) &= ||\boldsymbol{w}||_2^2/2 \\
f(\boldsymbol{w}) &= \boldsymbol{w} \\
\Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) &= ||\widetilde{\boldsymbol{w}}||_2^2/2 - ||\boldsymbol{w}||_2^2/2 - (\widetilde{\boldsymbol{w}} - \boldsymbol{w}) \cdot \boldsymbol{w} \\
&= ||\widetilde{\boldsymbol{w}} - \boldsymbol{w}||_2^2/2
\end{aligned}
$$

(Unnormalized) Relative Entropy

$$
\begin{aligned}
F(\boldsymbol{w}) &= \sum_i (w_i \ln w_i - w_i) \\
f(\boldsymbol{w}) &= \ln \boldsymbol{w} \\
\Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) &= \sum_i \left(\widetilde{w_i} \ln \frac{\widetilde{w_i}}{w_i} + w_i - \widetilde{w_i} \right)
\end{aligned}
$$

Examples-2 [GLS,GL]

p-norm Algs (q is dual to p: $\frac{1}{p} + \frac{1}{q} = 1$)

$$F(\boldsymbol{w}) = \frac{1}{2}||\boldsymbol{w}||_q^2$$

$$f(\boldsymbol{w}) = \nabla \frac{1}{2}||\boldsymbol{w}||_q^2$$

$$\Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) = \frac{1}{2}||\widetilde{\boldsymbol{w}}||_q^2 + \frac{1}{2}||\boldsymbol{w}||_q^2 - (\widetilde{\boldsymbol{w}} - \boldsymbol{w}) \cdot f(\boldsymbol{w})$$

When $p = q = 2$ this reduces to squared Euclidean distance (Widrow-Hoff).

Examples-3

Burg entropy

$$
\begin{aligned}
F(\boldsymbol{w}) &= \sum_i -\ln w_i \\
f(\boldsymbol{w}) &= -\frac{1}{\boldsymbol{w}} \\
\Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) &= \sum_i \left(-\ln \frac{\widetilde{w_i}}{w_i} + \frac{\widetilde{w_i}}{w_i} \right) - n
\end{aligned}
$$

General Motivation of Updates [KW]

Trade-off between two term:

$$\boldsymbol{w}_{t+1} = \underset{\boldsymbol{w}}{\operatorname{argmin}}\Big(\underbrace{\Delta_F(\boldsymbol{w}, \boldsymbol{w}_t)}_{weight\ domain} + \eta_t \underbrace{L_t(\boldsymbol{w})}_{label\ domain} \Big)$$

$\Delta_F(\boldsymbol{w}, \boldsymbol{w}_t)$ is "regularization term" and serves as measure of progress in the analysis.

When loss L is convex (in $\boldsymbol{w}$)

$$\nabla_{\boldsymbol{w}}(\Delta_F(\boldsymbol{w}, \boldsymbol{w}_t) + \eta_t L_t(\boldsymbol{w})) = 0$$

iff

$$f(\boldsymbol{w}) - f(\boldsymbol{w}_t) + \eta_t \underbrace{\nabla L_t(\boldsymbol{w})}_{\approx \nabla L_t(\boldsymbol{w}_t)} = 0$$

$$\Rightarrow \quad \boldsymbol{w}_{t+1} = f^{-1}\left(f(\boldsymbol{w}_t) - \eta_t \nabla L_t(\boldsymbol{w}_t)\right)$$

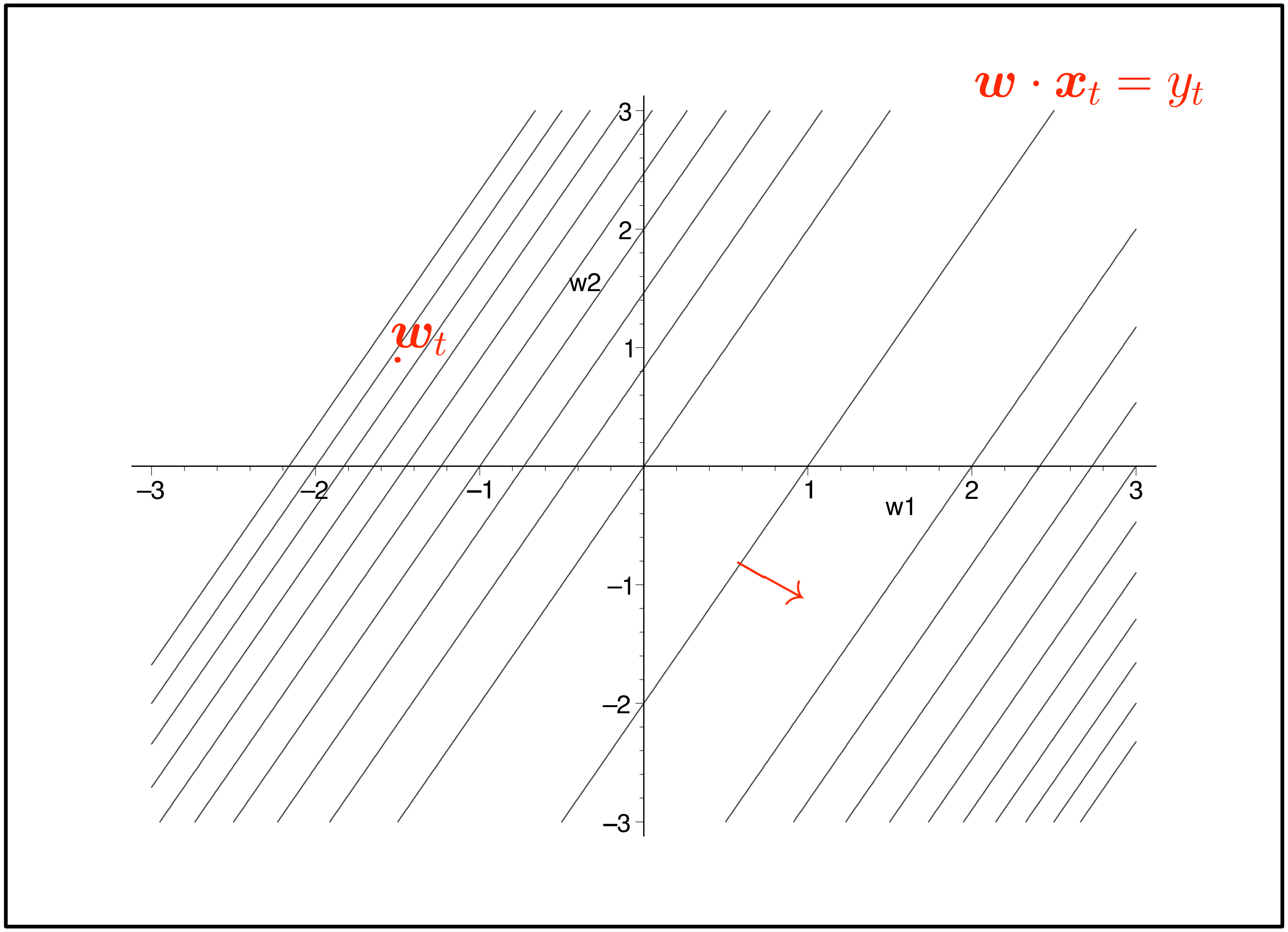
Quadratic Loss

$$L_t(\boldsymbol{w}) = \frac{1}{2}(y_t - \boldsymbol{w} \cdot \boldsymbol{x}_t)^2$$

$\boldsymbol{w}_t = (-3/2, 1)$

$\boldsymbol{x}_t = (1, -0.5)$

$y_t = 1$



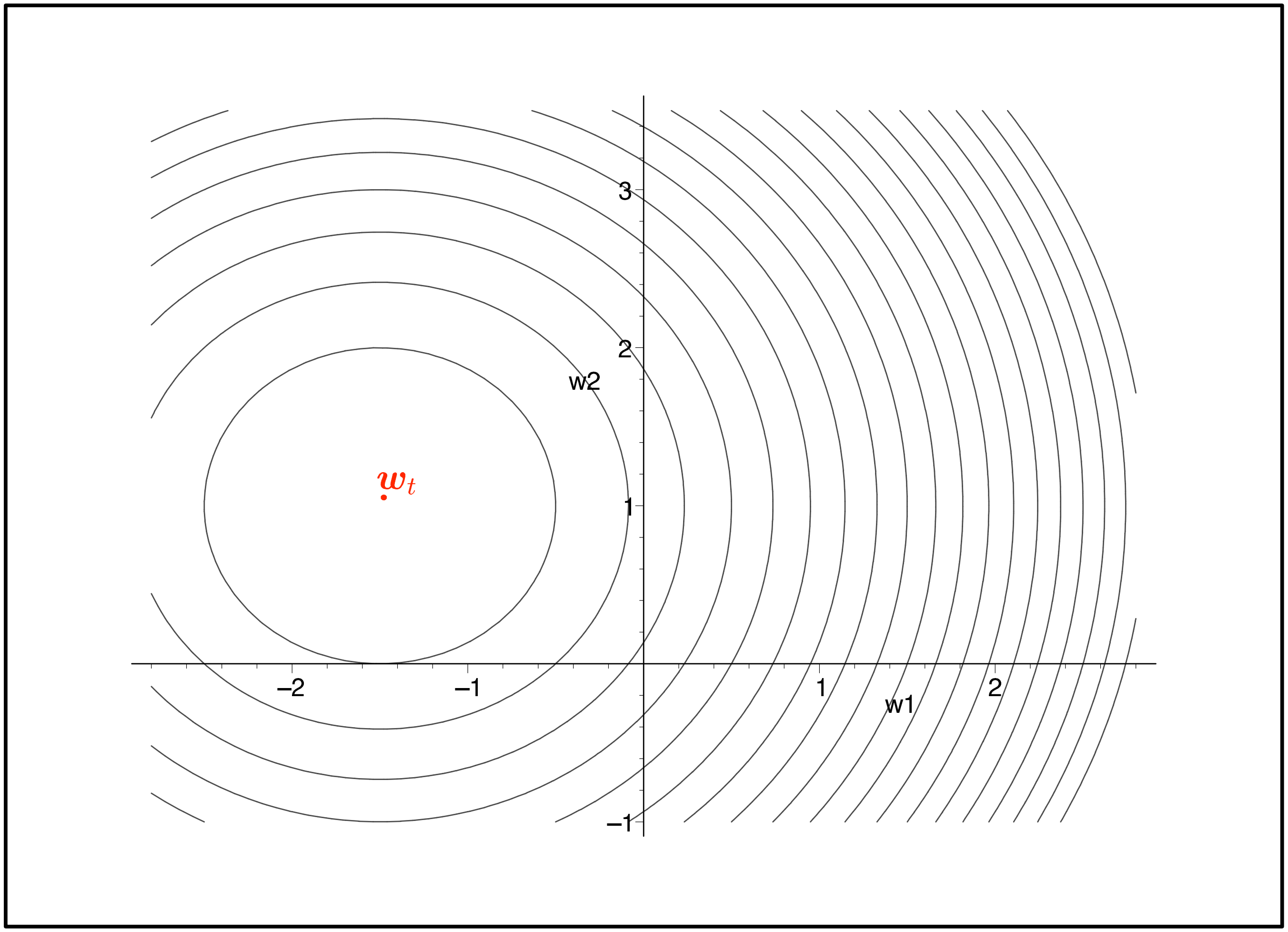
Divergence: Euclidean Distance Squared

$$\Delta_F(\boldsymbol{w}, \boldsymbol{w}_t) = \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_t\|_2^2$$

$\boldsymbol{w}_t = (-3/2, 1)$

$\boldsymbol{x}_t = (1, -0.5)$

$y_t = 1$



Divergence + η Loss

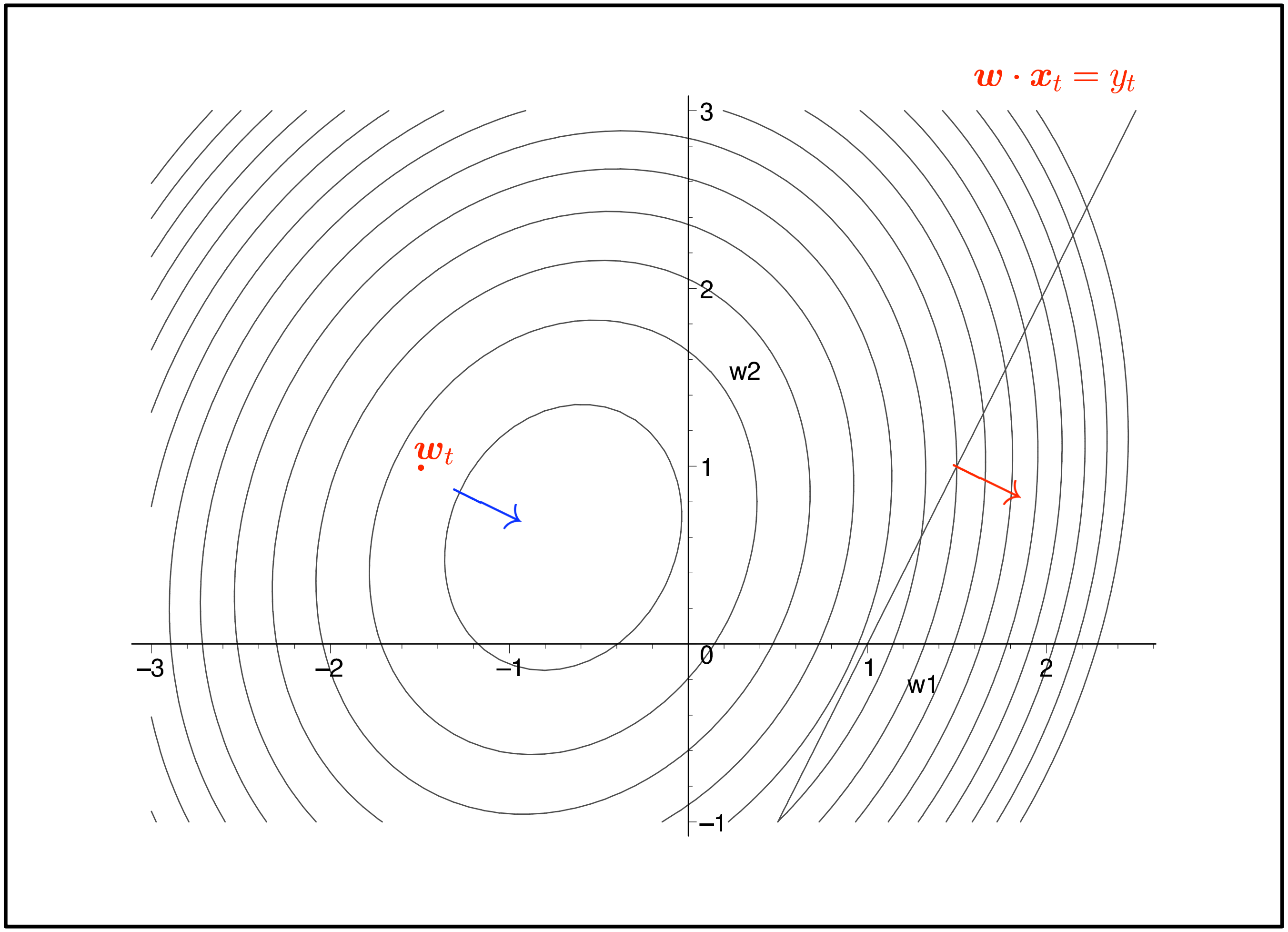
$$\frac{1}{2}\|\boldsymbol{w}-\boldsymbol{w}_t\|_2^2+\eta\frac{1}{2}(y_t-\boldsymbol{w}\cdot\boldsymbol{x}_t)^2$$

$\boldsymbol{w}_t = (-3/2, 1)$

$\boldsymbol{x}_t = (1, -0.5)$

$y_t = 1$

$\eta = 0.2$



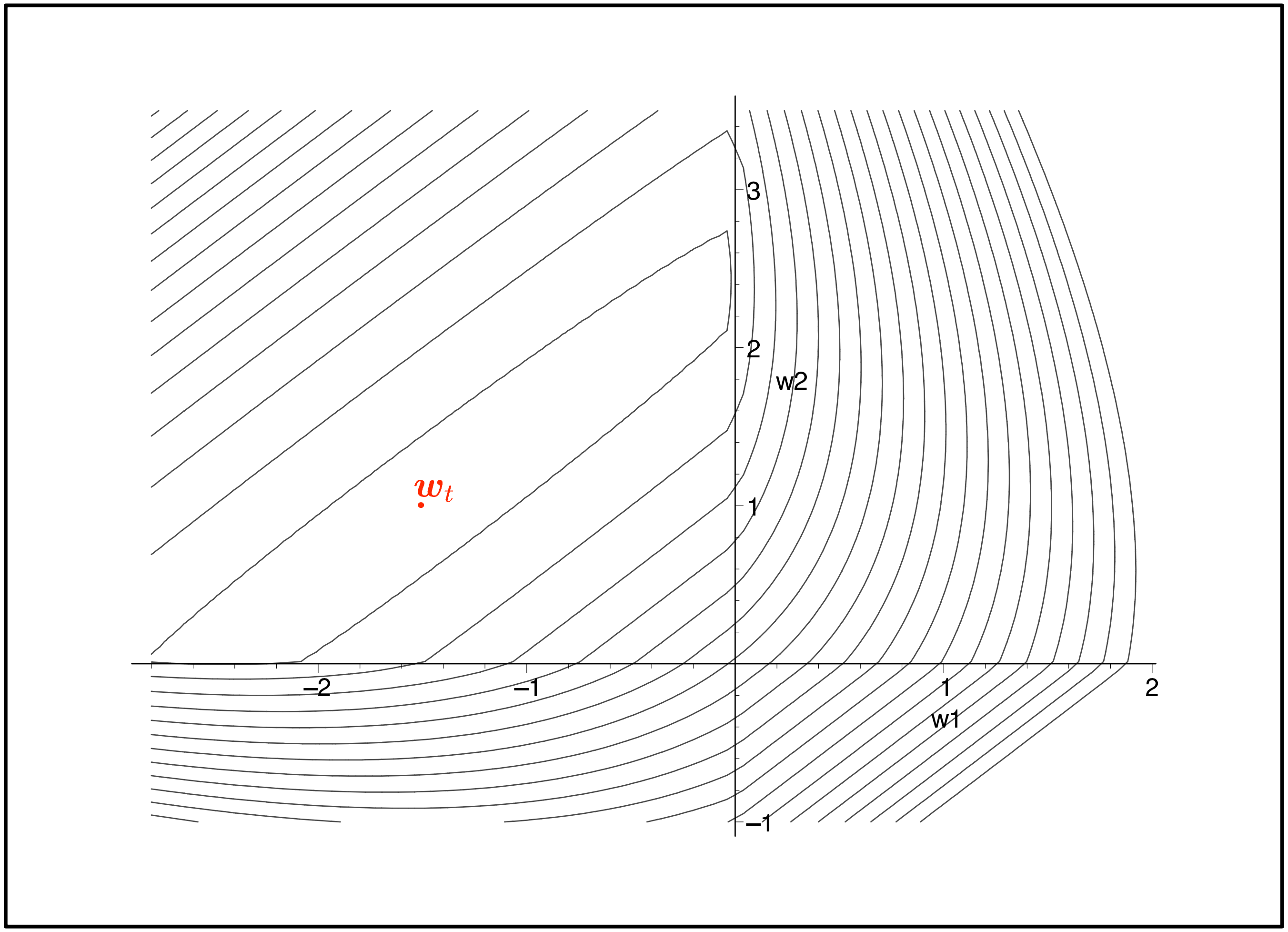
Divergence: 10-norm algorithm divergence

$\Delta_F(\boldsymbol{w}, \boldsymbol{w}_t)$ where $F(\boldsymbol{w}) = \frac{1}{2}||\boldsymbol{w}||_{10}^2$

$\boldsymbol{w}_t = (-3/2, 1)$

$\boldsymbol{x}_t = (1, -0.5)$

$y_t = 1$



Loss + η Divergence

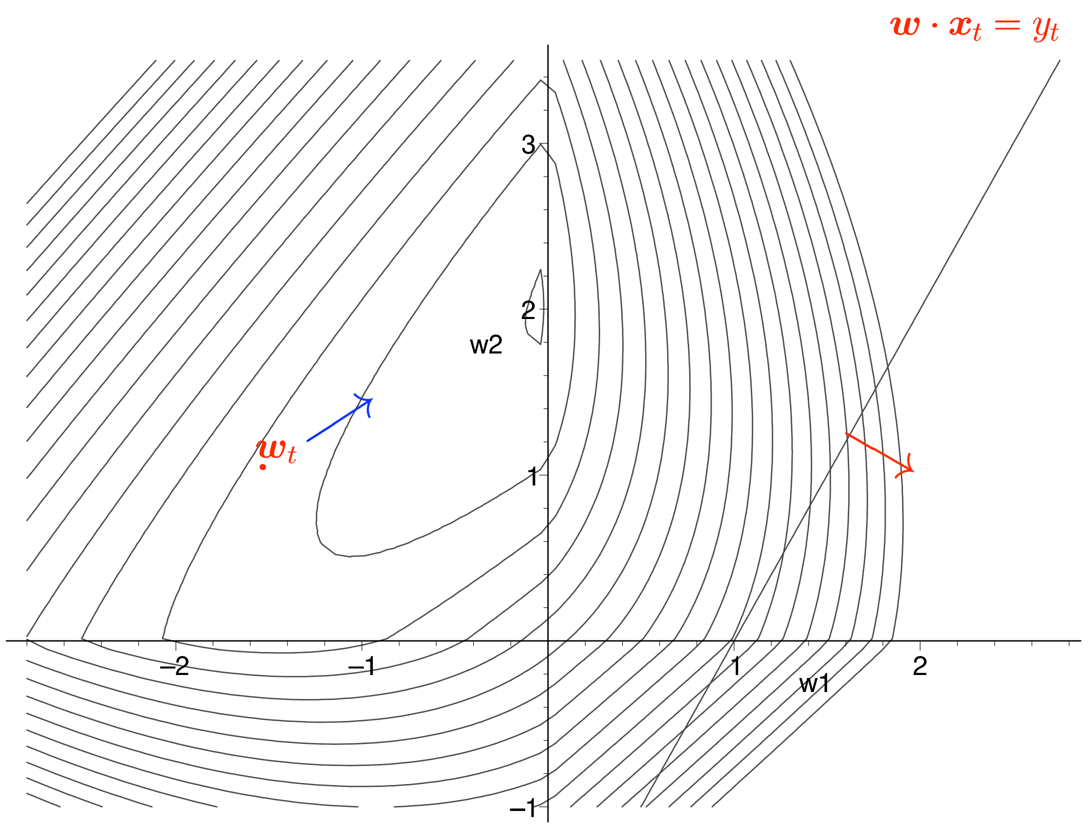
$$\Delta_F(\boldsymbol{w}, \boldsymbol{w}_t) + \eta \frac{1}{2}(y_t - \boldsymbol{w} \cdot \boldsymbol{x}_t)^2, \text{ where } F(\boldsymbol{w}) = \frac{1}{2}||\boldsymbol{w}||_{10}^2$$

$\boldsymbol{w}_t = (-3/2, 1)$

$\boldsymbol{x}_t = (1, -0.5)$

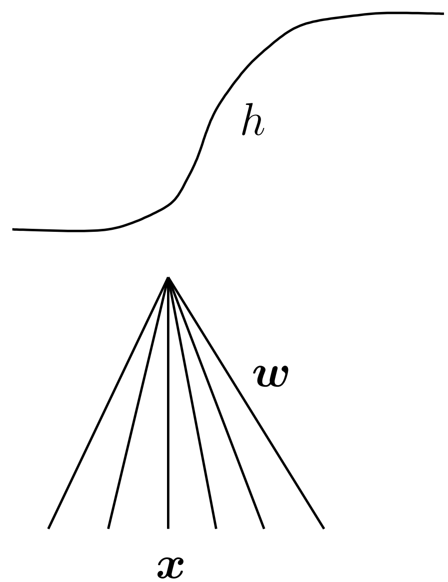
$y_t = 1$

$\eta = 0.2$

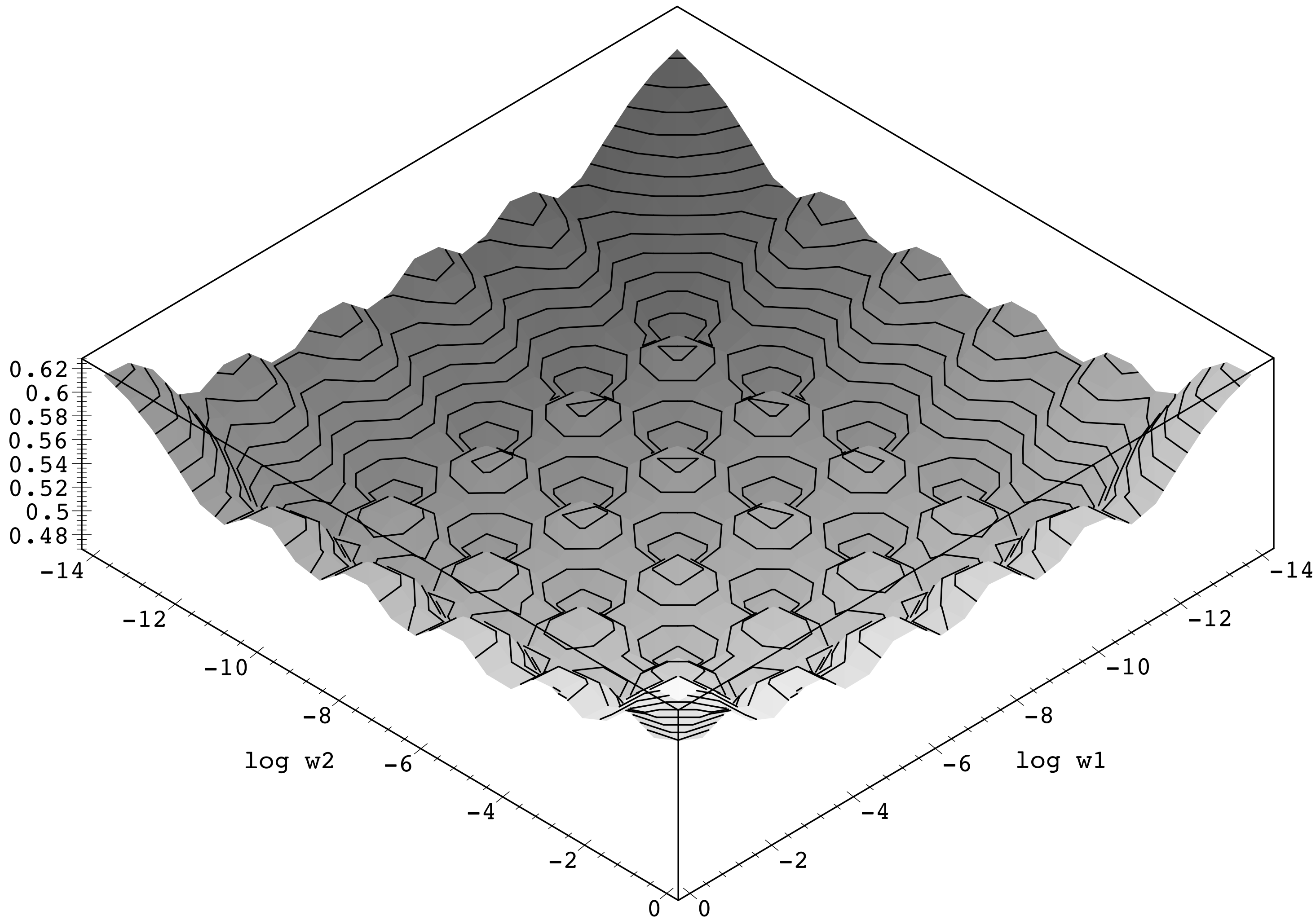


Nonlinear Regression

$\hat{y} = h(\boldsymbol{w} \cdot \boldsymbol{x})$

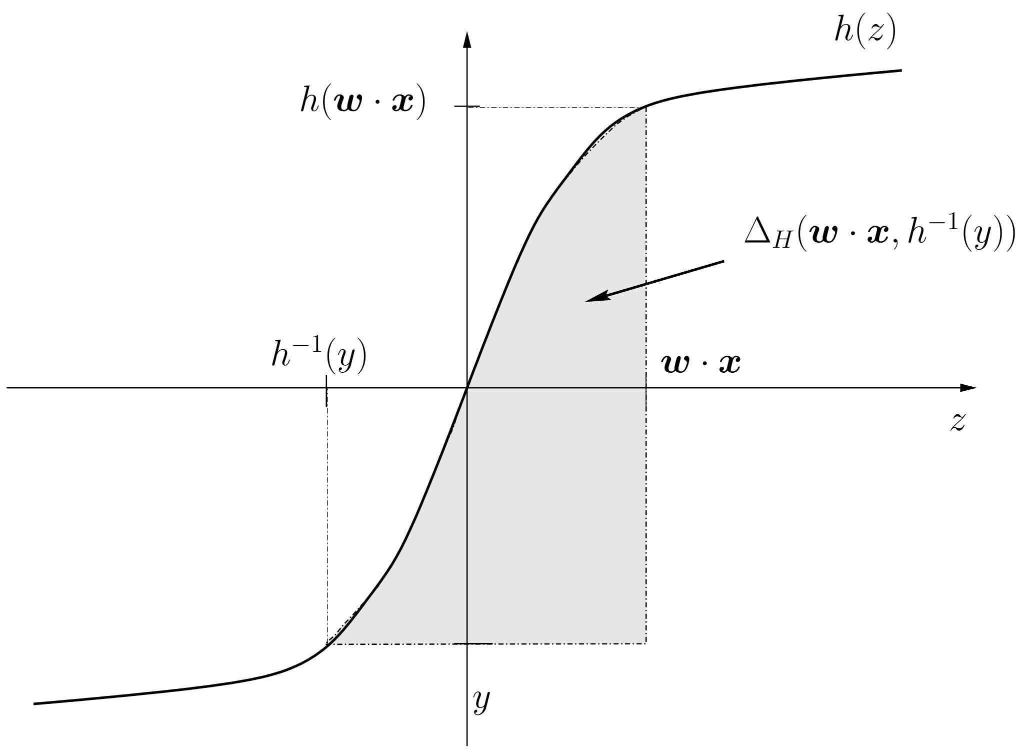


- Sigmoid function $h(z) = \frac{1}{1+e^{-z}}$
- For a set of examples $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_T, y_T)$ total loss $\sum_{t=1}^{T} h(\boldsymbol{w} \cdot \boldsymbol{x}) - y_t)^2/2$ can have exponentially many minima in weight space [Bu,AHW]



Want loss that is convex in $\boldsymbol{w}$

Bregman Div. Lead to Good Loss Function



$$(h = \nabla H)$$

$$\begin{aligned}
\int_{h^{-1}(y)}^{\boldsymbol{w} \cdot \boldsymbol{x}} (h(z) - y)\, dz &= H(\boldsymbol{w} \cdot \boldsymbol{x}) - H(h^{-1}(y)) - (\boldsymbol{w} \cdot \boldsymbol{x} - h^{-1}(y))\, y \\
&= \Delta_H(\boldsymbol{w} \cdot \boldsymbol{x}, h^{-1}(y))
\end{aligned}$$

Use $\Delta_H(\boldsymbol{w}\cdot\boldsymbol{x}, h^{-1}(y))$ as loss of $\boldsymbol{w}$ on $(\boldsymbol{x}, y)$

Called matching loss for h [AHW,HKW]

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

transfer f. $h(z)$	$H(z)$	match. loss $d_H(\boldsymbol{w}\cdot\boldsymbol{x}, h^{-1}(y))$
z	$\frac{1}{2}z^2$	$\frac{1}{2}(\boldsymbol{w}\cdot\boldsymbol{x} - y)^2$ square loss
$\frac{e^z}{1+e^z}$	$\ln(1+e^z)$	$\ln(1+e^{\boldsymbol{w}\cdot\boldsymbol{x}}) - y\boldsymbol{w}\cdot\boldsymbol{x}$ $+y\ln y + (1-y)\ln(1-y)$ logistic loss
$\text{sign}(z)$	$\lvert z\rvert$	$\max\{0, -y\boldsymbol{w}\cdot\boldsymbol{x}\}$ hinge loss

Idea behind the matching loss

If transfer function and loss match, then

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

Then update has simple form:

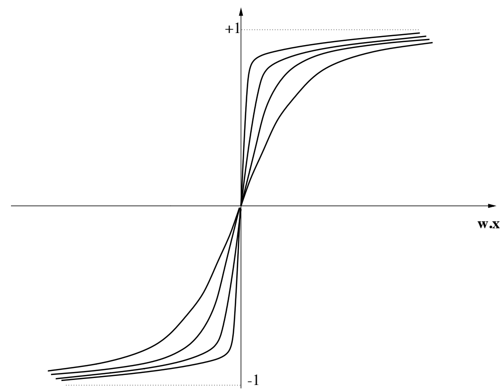
$$f(\boldsymbol{w}_{t+1}) = f(\boldsymbol{w}_t) - \eta_t (h(\boldsymbol{w}_t \cdot \boldsymbol{x}) - y_t) \boldsymbol{x}_t$$

This can be exploited in proofs

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

Sigmoid in the Limit

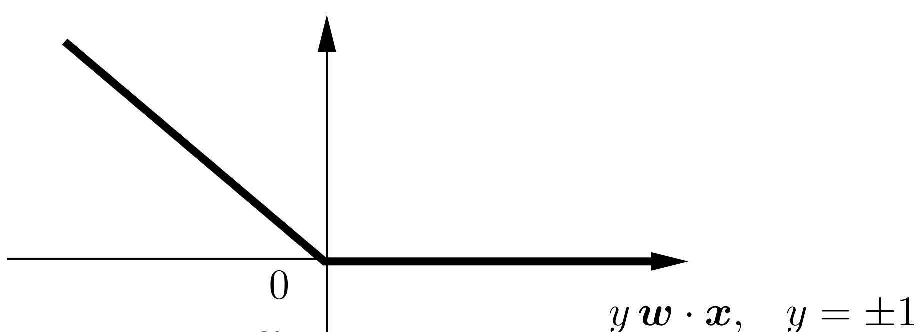
For transfer function $h(z) = \text{sign}(z)$



$H(z) = |z|$

Matching loss is hinge loss [GW]

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



Convex in $\boldsymbol{w}$ but not differentiable

Motivation of linear threshold algs

Gradient descent with Hinge Loss	Perceptron
Expon. gradient with Hinge Loss	Normalized Winnow

Known linear threshold algorithms for ± 1-classification case are gradient-based algorithms with hinge loss

Perceptron

$$
\begin{aligned}
& \boldsymbol{w}_{t+1} \\
& = \underset{\boldsymbol{w}}{\operatorname{argmin}} \left(||\boldsymbol{w} - \boldsymbol{w}_t||^2/2 + \eta \, HL(\boldsymbol{w} \cdot \boldsymbol{x}_t, g^{-1}(y_t)) \right) \\
& = \boldsymbol{w}_t - \eta \left(\operatorname{sign}(\boldsymbol{w}_{t+1} \cdot \boldsymbol{x}_t) \right) - y_t \right) \boldsymbol{x}_t \\
& \approx \boldsymbol{w}_t - \eta \left(\underbrace{\operatorname{sign}(\boldsymbol{w}_t \cdot \boldsymbol{x}_t)}_{\hat{y}_t} - y_t \right) \boldsymbol{x}_t
\end{aligned}
$$

Normalized Winnow

$$\boldsymbol{w}_{t+1}$$

$$= \operatorname*{argmin}_{\boldsymbol{w}} \left(\sum_{i=1}^{n} w_i \ln \frac{w_i}{w_{t,i}} + \eta\, HL(\boldsymbol{w} \cdot \boldsymbol{x}_t, g^{-1}(y_t)) \right)$$

$$= w_{t,i}\; e^{-\eta\, (\mathrm{sign}(\boldsymbol{w} \cdot \boldsymbol{x}_t) - y_t)\, x_{t.i}} / \text{ normalization}$$

$$\approx w_{t,i}\; e^{-\eta\, (\underbrace{\mathrm{sign}(\boldsymbol{w}_t \cdot \boldsymbol{x}_t)}_{\hat{y}_t} - y_t)\, x_{t.i}} / \text{ normalization}$$

Trade-off between two divergences [KW]

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}}\big(\underbrace{\Delta_F(\boldsymbol{w}, \boldsymbol{w}_t)}_{\substack{\text{parameter}\\ \text{divergence}}} + \eta_t \underbrace{\Delta_H(\boldsymbol{w} \cdot \boldsymbol{x}_t, h^{-1}(y_t))}_{\substack{\text{matching}\\ \text{loss divergence}}}\big)$$

Both divergences are convex in $\boldsymbol{w}$

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

Generalization of the "delta"-rule