For any differentiable convex function $F$

$$
\Delta_F(\tilde{w}, w) = F(\tilde{w}) - F(w) - (\tilde{w} - w) \cdot \nabla_w F(w)
$$

$$
= F(\tilde{w}) - \text{supporting hyperplane through } (w, F(w))
$$
Bregman Divergences: Simple Properties

1. $\Delta_F(\tilde{w}, w)$ is convex in $\tilde{w}$

2. $\Delta_F(\tilde{w}, w) \geq 0$
   
   If $F$ convex equality holds iff $\tilde{w} = w$

3. Usually not symmetric: $\Delta_F(\tilde{w}, w) \neq \Delta_F(w, \tilde{w})$

4. Linearity (for $a \geq 0$):
   
   $\Delta_{F+aH}(\tilde{w}, w) = \Delta_F(\tilde{w}, w) + a \Delta_H(\tilde{w}, w)$

5. Unaffected by linear terms ($a \in \mathbb{R}, b \in \mathbb{R}^n$):
   
   $\Delta_{H+a\tilde{w}+b}(\tilde{w}, w) = \Delta_H(\tilde{w}, w)$
Bregman Divergences: more properties

6. $\nabla \tilde{w} \Delta_F (\tilde{w}, w)$

$$= \nabla F(\tilde{w}) - \nabla \tilde{w}(\tilde{w} \nabla w F(w))$$

$$= f(\tilde{w}) - f(w)$$

7. $\Delta_F (w_1, w_2) + \Delta_F (w_2, w_3)$

$$= F(w_1) - F(w_2) - (w_1 - w_2)f(w_2)$$

$$= F(w_2) - F(w_3) - (w_2 - w_3)f(w_3)$$

$$= \Delta_F (w_1, w_3) + (w_1 - w_2) \cdot (f(w_3) - f(w_2))$$
A Pythagorean Theorem \([\text{Br,Cs,A,HW}]\)

\(w^*\) is projection of \(w\) onto convex set \(\mathcal{W}\) w.r.t. Bregman divergence \(\Delta_F\):

\[
\begin{align*}
    w^* &= \text{argmin} \Delta_F(u, w) \\
    &= \arg\min_{u \in \mathcal{W}} \Delta_F(u, w)
\end{align*}
\]

Theorem:

\[
\Delta_F(u, w) \geq \Delta_F(u, w^*) + \Delta_F(w^*, w)
\]
Examples

Squared Euclidean Distance

\[ F(w) = \|w\|^2_2 / 2 \]
\[ f(w) = w \]
\[ \Delta_F(\tilde{w}, w) = \|\tilde{w}\|^2_2 / 2 - \|w\|^2_2 / 2 - (\tilde{w} - w) \cdot w \]
\[ = \|\tilde{w} - w\|^2_2 / 2 \]

(Unnormalized) Relative Entropy

\[ F(w) = \sum_i (w_i \ln w_i - w_i) \]
\[ f(w) = \ln w \]
\[ \Delta_F(\tilde{w}, w) = \sum_i \left( \tilde{w}_i \ln \frac{\tilde{w}_i}{w_i} + w_i - \tilde{w}_i \right) \]
$p$-norm Algs ($q$ is dual to $p$: $\frac{1}{p} + \frac{1}{q} = 1$)

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

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

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

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

\[ F(w) = \sum_i -\ln w_i \]

\[ f(w) = -\frac{1}{w} \]

\[ \Delta_F(\tilde{w}, w) = \sum_i \left( -\ln \frac{\tilde{w}_i}{w_i} + \frac{\tilde{w}_i}{w_i} \right) - n \]
Umegaki Divergence

\[ F(W) = \text{tr} (W \ln W - W) \]

\[ f(W) = \ln W \]

\[ \Delta_F(\tilde{W}, W) = \text{tr} \left( \tilde{W}(\ln \tilde{W} - \ln W) + W - \tilde{W} \right) \]

LogDet Divergence

\[ F(W) = -\ln |W| \]

\[ f(W) = W^{-1} \]

\[ \Delta_F(\tilde{W}, W) = -\ln \frac{\tilde{W}}{|W|} + \text{tr}(\tilde{W}W^{-1}) - n \]
General Motivation of Updates [KW]

Trade-off between two terms:

\[ w_{t+1} = \arg \min_w \left( \Delta_F(w, w_t) + \eta_t L_t(w) \right) \]

\( \Delta_F(w, w_t) \) is “regularization term” and serves as measure of progress in the analysis.

When loss \( L \) is convex (in \( w \))

\[ \nabla_w (\Delta_F(w, w_t) + \eta_t L_t(w)) = 0 \]

iff

\[ f(w) - f(w_t) + \eta_t \nabla L_t(w) = 0 \]

\[ \approx \nabla L_t(w_t) \]

\[ \Rightarrow w_{t+1} = f^{-1}(f(w_t) - \eta_t \nabla L_t(w_t)) \]
Quadratic Loss

\[ L_t(w) = \frac{1}{2}(y_t - w \cdot x_t)^2 \]

\[ w_t = (-3/2, 1) \]
\[ x_t = (1, -0.5) \]
\[ y_t = 1 \]
Divergence: Euclidean Distance Squared

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

$$w_t = (-3/2, 1)$$

$$x_t = (1, -0.5)$$

$$y_t = 1$$
Divergence + $\eta$ Loss

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

$$\mathbf{w}_t = \left(-\frac{3}{2}, 1\right)$$
$$\mathbf{x}_t = (1, -0.5)$$
$$y_t = 1$$
$$\eta = 0.2$$
Divergence: 10-norm algorithm divergence

\[ \Delta_F(w, w_t) \text{ where } F(w) = \frac{1}{2} \| w \|_{10}^2 \]

\[ w_t = (-3/2, 1) \]
\[ x_t = (1, -0.5) \]
\[ y_t = 1 \]
Loss + $\eta$ Divergence

$$\Delta F(w, w_t) + \eta \frac{1}{2}(y_t - w \cdot x_t)^2, \text{ where } F(w) = \frac{1}{2}||w||_1^2$$

$w_t = (-3/2, 1)$
$x_t = (1, -0.5)$
$y_t = 1$
$\eta = 0.2$
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 $w$
Bregman Div. Lead to Good Loss Function

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

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

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

Called **matching loss** for $h$  

Matching loss is **convex** in $w$

<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>$\frac{1}{2}(w \cdot x - y)^2$ square loss</td>
</tr>
<tr>
<td>$z$</td>
<td>$\frac{1}{2}z^2$</td>
<td>$\frac{1}{2}(w \cdot x - y)^2$ square loss</td>
</tr>
<tr>
<td>$\frac{e^z}{1+e^z}$</td>
<td>$\ln(1 + e^z)$</td>
<td>$\ln(1 + e^{w \cdot x}) - yw \cdot x$ logistic loss</td>
</tr>
<tr>
<td>$\text{sign}(z)$</td>
<td>$</td>
<td>z</td>
</tr>
</tbody>
</table>
Idea behind the matching loss

If transfer function and loss match, then

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

Then update has simple form:

$$f(w_{t+1}) = f(w_t) - \eta_t (h(w_t \cdot x) - y_t) x_t$$

This can be exploited in proofs

But not absolutely necessary
One only needs convexity of $L(h(w \cdot x), y)$ in $w$ [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
  Hinge Loss

Exponent. gradient
  with
  Hinge Loss

Perceptron

Normalized

Winnow

Known linear threshold algorithms for ±1-classification case are gradient-based algorithms with hinge loss
\[ w_{t+1} = \arg \min_w \left( ||w - w_t||^2 / 2 + \eta H L(w \cdot x_t, g^{-1}(y_t)) \right) \]

\[ = w_t - \eta (\text{sign}(w_{t+1} \cdot x_t)) - y_t \right) x_t \]

\[ \approx w_t - \eta (\text{sign}(w_t \cdot x_t) - y_t) \right) x_t \]

\[ \hat{y}_t \]
Normalized Winnow

\[ w_{t+1} \]

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

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

\[ \approx w_{t,i} e^{\hat{y}_t y_t - y_t x_{t,i}} / \text{normalization} \]
Figure 1: The matching loss between estimate \( \hat{a} \) and measurement \( a \) is the area from \( \hat{a} \) to \( a \) underneath the transfer function. (If \( a \) is less than \( \hat{a} \), then the loss is the area above the transfer function.)

\[
\text{loss}(\hat{a}, a) = \int_{a}^{\hat{a}} (h(z) - h(a)) \, dz
\]
Figure 2: On the top row we plot three choices of the transfer function $h(a)$ and their derivatives: The first is the sigmoid function, i.e. $h(a) = s(a) := \frac{e^a}{1 + e^a}$, the others two are shifted versions of the sigmoid. In the bottom row we plot the loss($\hat{a}, a$) as a function of the estimate $\hat{a}$ for fixed activities $a = -3, 0, 3$ (The transfer function is always specified in the plot above). Note that locally the losses are quadratic and the steepness of the bowl is determined by $h'(a)$. 
Trade-off between two divergences [KW]

\[ w_{t+1} = \text{argmin}_w \left( \Delta_F(w, w_t) + \eta_t \Delta_H(w \cdot x_t, h^{-1}(y_t)) \right) \]

- 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
Bregman divergences and exponential families?

- Exponential family of distributions
- Inherent duality

\[ w_{t+1} = f^{-1} (f(w_t) - \eta \nabla L_t(w_t)) \]

<table>
<thead>
<tr>
<th>primal param.</th>
<th>dual param.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( w_t )</td>
<td>( f )</td>
</tr>
<tr>
<td>( w_{t+1} )</td>
<td>( f^{-1} )</td>
</tr>
</tbody>
</table>
Exponential Family of Distributions

- Parametric density functions
  
  \[ P_G(x|\theta) = e^{\theta \cdot x - G(\theta)} P_0(x) \]

- \( \theta \) and \( x \) vectors in \( \mathbb{R}^d \)

- Cumulant function \( G(\theta) \) assures normalization
  
  \[ G(\theta) = \ln \int e^{\theta \cdot x} P_0(x) \, dx \]

- \( G(\theta) \) is convex function on convex set \( \Theta \subseteq \mathbb{R}^d \)

- \( G \) characterizes members of the family

- \( \theta \) is natural parameter
• Expectation parameter

\[ \mu = \int x P_G(x|\theta) dx = E_{\theta}(x) = g(\theta) \]

where \( g(\theta) = \nabla_{\theta} G(\theta) \)

• Second convex function \( F(\mu) \) on space \( g(\Theta) \)

\[ F(\mu) = \theta \cdot \mu - G(\theta) \]

• \( G(\theta) \) and \( F(\mu) \) are convex conjugate functions

• Let \( f(\mu) = \nabla_{\mu} F(\mu) \)

• \( f(\mu) = g^{-1}(\mu) \)
Primal & Dual Parameters

- \( \theta \) is the natural parameter
- \( \mu \) is the expectation parameter

Parameter transformations:
- \( g(\theta) = \mu \)
- \( f(\mu) = \theta \)

- \( \theta \) and \( \mu \) are dual parameters
Gaussian (unit variance)

\[ P(x|\theta) \sim e^{-\frac{1}{2}(\theta-x)^2} = e^{\theta \cdot x - \frac{1}{2} \theta^2} e^{\frac{1}{2}x^2} \]

Cumulant function: \( G(\theta) = \frac{1}{2} \theta^2 \)

Parameter transformations:
\[ g(\theta) = \theta = \mu \quad \text{and} \quad f(\mu) = \mu = \theta \]

Dual convex function: \( F(\mu) = \theta \cdot \mu - G(\theta) = \frac{1}{2} \mu^2 \)

Square loss: \( L_t(\theta) = \frac{1}{2}(\theta_t - x_t)^2 \)
Examples $x_t$ are coin flips in $\{0, 1\}$

$$P(x|\mu) = \mu^x (1 - \mu)^{1-x}$$

$\mu$ is the probability (expectation) of 1

Natural parameter: $\theta = \ln \frac{\mu}{1-\mu}$

$$P(x|\theta) = \exp \left( \theta x - \ln(1 + e^\theta) \right)$$

Cumulant function: $G(\theta) = \ln(1 + e^\theta)$

Parameter transformations:

$$\mu = g(\theta) = \frac{e^\theta}{1 + e^\theta} \text{ and } \theta = f(\mu) = \ln \frac{\mu}{1 - \mu}$$

Dual function: $F(\mu) = \mu \ln \mu + (1 - \mu) \ln(1 - \mu)$

Log loss: $L_t(\theta) = -x_t \theta + \ln(1 + e^\theta)$

$$= -x_t \ln \mu - (1 - x_t) \ln(1 - \mu)$$
Examples $x_t$ are natural numbers in $\{0, 1, \ldots\}$

$$P(x|\mu) = \frac{e^{-\mu} \mu^x}{x!}$$

$\mu$ is expectation of $x$

Natural parameter: $\theta = \ln \mu$

$$P(x|\theta) = \exp \left( \theta x - e^{\theta} \right) \frac{1}{x!}$$

Cumulant function: $G(\theta) = e^\theta$

Parameter transformations:

$$\mu = g(\theta) = e^\theta \text{ and } \theta = f(\mu) = \ln \mu$$

Dual function: $F(\mu) = \mu \ln \mu - \mu$

Loss: $L_t(\theta) = -x_t \theta + e^\theta + \ln x_t!$

$$= -x_t \ln \mu + \mu + \ln x_t!$$
Let \( P(x|\theta) \) and \( P(x|\tilde{\theta}) \) denote two distributions with cumulant function \( G \)

\[
\Delta_G(\tilde{\theta}, \theta) = \int_x P_G(x|\theta) \ln \frac{P_G(x|\theta)}{P_G(x|\tilde{\theta})} dx \\
= \int_x P_G(x|\theta)(\theta \cdot x - G(\theta) - \theta \cdot x + G(\tilde{\theta})) dx \\
= G(\tilde{\theta}) - G(\theta) - (\tilde{\theta} - \theta) \cdot (\int_x P_G(x|\theta) x dx) \\
= G(\tilde{\theta}) - G(\theta) - (\tilde{\theta} - \theta) \cdot \mu \\
F(\mu) = \theta \cdot \mu - G(\theta) \\
= F(\mu) - F(\tilde{\mu}) - (\mu - \tilde{\mu}) \cdot \tilde{\theta} \\
= \Delta_F(\mu, \tilde{\mu}) \quad [A, BN, AW]
Area unchanged When Slide Flipped

\[ \Delta_G(\theta, \tilde{\theta}) = \Delta_F(\mu, \tilde{\mu}) \]
\[
\Delta_G(\theta, \tilde{\theta}) = G(\theta) - G(\tilde{\theta}) - (\theta - \tilde{\theta}) \cdot g(\tilde{\theta})
\]

\[
\begin{align*}
= & \quad \int_{\tilde{\theta}}^{\theta} (g(\tau) - g(\tilde{\theta})) \cdot d\tau \\
\text{flip} = & \quad \int_{\mu}^{\tilde{\mu}} (f(\sigma) - f(\mu)) \cdot d\sigma \\
= & \quad F(\tilde{\mu}) - F(\mu) - (\tilde{\mu} - \mu) \cdot f(\mu) \\
= & \quad \Delta_F(\tilde{\mu}, \mu)
\end{align*}
\]
Dual divergence for Bernoulli

\[ G(\theta) = \ln(1 + e^{\theta}) \]
\[ F(\mu) = \mu \ln \mu + (1 - \mu) \ln(1 - \mu) \]

\[ g(\theta) = \frac{e^{\theta}}{1 + e^{\theta}} = \mu \]
\[ f(\mu) = \ln \frac{\mu}{1 - \mu} = \theta \]

\[ \Delta_G(\tilde{\theta}, \theta) = \ln(1 + e^{\tilde{\theta}}) - \ln(1 + e^{\theta}) - (\tilde{\theta} - \theta) \frac{e^{\theta}}{1 + e^{\theta}} \]

\[ \Delta_F(\mu, \tilde{\mu}) = \mu \ln \frac{\mu}{\tilde{\mu}} + (1 - \mu) \ln \frac{1 - \mu}{1 - \tilde{\mu}} \]

Binary relative entropy
Sum of binary relative entropies is parameter divergence for BEG
Dual divergence for Poisson

\[ G(\theta) = e^\theta \quad F(\mu) = \mu \ln \mu - \mu \]

\[ g(\theta) = e^\theta = \mu \quad f(\mu) = \ln \mu = \theta \]

\[ \Delta_G(\tilde{\theta}, \theta) = e^{\tilde{\theta}} - e^\theta - (\tilde{\theta} - \theta)e^\theta \]

\[ \Delta_F(\mu, \tilde{\mu}) = \mu \ln \frac{\mu}{\tilde{\mu}} + \tilde{\mu} - \mu \]

Unnormalized relative entropy
Sum of unnormalized relative entropies is parameter for UEG (e.g. Winnow)
Dual matching loss for sigmoid transfer func.

\[
H(z) = \ln(1 + e^z) \quad K(r) = r \ln r + (1 - r) \ln(1 - r) \\
h(z) = \frac{e^z}{1 + e^z} = r \quad k(r) = \ln \frac{r}{1-r} = z \\

K \text{ dual to } H \text{ and } k = h^{-1}
\]

\[
\Delta_H(w \cdot x, h^{-1}(y)) = \ln(1 + e^{w \cdot x}) - y w \cdot x + y \ln y + (1 - y) \ln(1 - y)
\]

By duality logistic loss is same as entropic loss

\[
\Delta_K(y, h(w \cdot x)) = y \ln \frac{y}{h(w \cdot x)} + (1 - y) \ln \frac{1 - y}{1 - h(w \cdot x)}
\]

Matching loss for logistic transfer function