Motivation Machine Learning Updates
Lecture 6, part 2, CMPS 272, W12

Manfred K. Warmuth

University of California - Santa Cruz

February 4, 2012
Outline

1. Motivations of updates via divergences?
2. Starting with differential equations
3. Relative entropies?
4. Entropies as barriers
Motivations of updates via divergences?

Starting with differential equations

Relative entropies?

Entropies as barriers
Motivate additive and multiplicative updates

- Use linear regression as example problem
- Motivate all updates as minimizing

\[ \text{regularization} + \eta \text{ loss/objective} \]

- Sometimes additional linear constraints
- Sometimes no regularization
Online linear regression

For $t = 1, 2, \ldots$

- Get instance $x_t \in \mathbb{R}_n$
- Predict $\hat{y}_t = s_t \cdot x_t$
- Get label $y_t \in \mathbb{R}$
- Incur square loss $(y_t - \hat{y}_t)^2$
- Update $s_t \rightarrow s_{t+1}$
Two main update families - linear regression

- **Additive**

\[ s_{t+1} = s_t - \eta (s_t \cdot x_t - y_t)x_t \]

Motivated by squared Euclidean distance
- Weights can go negative
- Gradient Descent (GD) - SKIING

- **Multiplicative**

\[ s_{t+1,i} = \frac{s_{t,i} e^{-\eta (s_t \cdot x_t - y_t)x_t,i}}{Z_t} \]

Motivated by relative entropy
- Updated weight vector stays on probability simplex
- Exponentiated Gradient (EG) - LIFE !

[KW97]
Additive Updates

**Goal**

Minimize tradeoff between closeness to last share vector and loss on last example

\[ s_{t+1} = \arg\min_s U(s) \]

\[ U(s) = \|s - s_t\|_2^2 + \eta (s \cdot x_t - y_t)^2 \]

\( \eta > 0 \) is the learning rate/speed
Additive updates

\[
\frac{\partial U(s)}{\partial s_i} \bigg|_{s_i = s_{t+1}, i} = 2(s_{t+1,i} - s_{t,i}) + 2\eta(s \cdot x_t - y_t)x_{t,i} = 0
\]

Therefore,

\[
\text{implicit:} \quad s_{t+1} = s_t - \eta(s_{t+1} \cdot x_t - y_t)x_t
\]

\[
\text{explicit:} \quad = s_t - \eta(s_t \cdot x_t - y_t)x_t
\]
Motivations of updates via divergences?

Multiplicative updates

\[ s_{t+1} = \arg\min_{\sum_i s_i = 1} U(s) \]

where \( U(s) = \sum_i s_i \ln \frac{s_i}{s_{t,i}} + \eta (s \cdot x_t - y_t)^2 \)

Define Lagrangian

\[ L(s) = \sum_i s_i \ln \frac{s_i}{s_{t,i}} + \eta (s \cdot x_t - y_t)^2 + \lambda (\sum_i s_i - 1) \]

where \( \lambda \) Lagrange coeff.
Multiplicative updates

\[
\frac{\partial L(s)}{\partial s_i} = \ln \frac{s_i}{s_{t,i}} + 1 + \eta(s \cdot x_t - y_t)x_{t,i} + \lambda = 0
\]

\[
\ln \frac{s_{t+1,i}}{s_{t,i}} = -\eta(s_{t+1} \cdot x_t - y_t)x_{t,i} - \lambda - 1
\]

\[
s_{t+1,i} = s_{t,i} e^{-\eta(s_{t+1} \cdot x_t - y_t)x_{t,i}} e^{-\lambda-1}
\]

Enforce normalization constraint by setting \(e^{-\lambda-1}\) to \(1/Z_t\)

**implicit:** \(s_{t+1,i} = \frac{s_{t,i} e^{-\eta(s_{t+1} \cdot x_t - y_t)x_{t,i}}}{Z_t}\)

**explicit:** \(= \frac{s_{t,i} e^{-\eta s_t \cdot x_t - y_t)x_{t,i}}{Z'_t}}{Z_t}\)
Connection to Biology

- One species for each of the $n$ dimension
- Fitness rates of species $i$ in time interval $(t, t+1]$ clamped to
  
  $$w_i = -\eta(s_{t+1} \cdot x_t - y_t)x_{t,i}$$

- Fitness
  
  $$W_i = e^{w_i}$$

Algorithm can be seen as running a population

- Can’t go to continuous time, since data points arrive at discrete times 1, 2, 3, ...
Outline

1. Motivations of updates via divergences?
2. Starting with differential equations
3. Relative entropies?
4. Entropies as barriers
GD via differential equations

\[ \dot{s}_t = -\eta \nabla L(s_t) \]

Here \( f(x_t) = \frac{\partial f(x_t)}{\partial t} \) Two ways to discretize:

**Forward Euler**

\[
\frac{s_{t+h} - s_t}{h} = -\eta \nabla L(s_t) \\
\]

\[
s_{t+h} = s_t - \eta h \nabla L(s_t) \\
\]

explicit: \( s_{t+1} = s_t - \eta h \nabla L(s_t) \)

**Backward Euler**

\[
\frac{s_{t+h} - h - s_{t+h}}{-h} = -\eta \nabla L(s_{t+h}) \\
\]

\[
s_{t+h} = s_t - \eta h \nabla L(s_{t+h}) \\
\]

implicit: \( s_{t+1} = s_t - \eta h \nabla L(s_{t+1}) \)
Starting with differential equations

EGU via differential equation

\[ \log s_t = -\eta \nabla L(s_t) \]

Two ways to discretize:

**Forward Euler**

\[
\frac{\log s_{t+h,i} - \log s_{t,i}}{h} = -\eta \nabla L(s_t)_i \\
\Rightarrow s_{t+h,i} = s_{t,i} e^{-\eta h \nabla L(s_t)_i} \\
\text{explicit: } s_{t+1,i} \equiv h=1 s_{t,i} e^{-\eta \nabla L(s_t)_i}
\]

**Backward Euler**

\[
\frac{\log s_{t+h} - h,i - \log s_{t+h,i}}{-h} = -\eta \nabla L(s_{t+h})_i \\
\Rightarrow s_{t+h,i} = s_{t,i} e^{-\eta h \nabla L(s_{t+h})_i} \\
\text{implicit: } s_{t+1,i} \equiv h=1 s_{t,i} e^{-\eta \nabla L(s_{t+1})_i}
\]

Derivation of normalized update more involved
Handle normalization by going one dimension lower

\[
\ln \frac{s_{t,i}}{1 - \sum_{j=1}^{n-1} s_{t,j}} = -\eta (\nabla L(s_t)_i - \nabla L(s_t)_n)
\]

**Backward Euler**

\[
\ln \frac{s_{t+1,i}}{1 - \sum_{j=1}^{n-1} s_{t+1,j}} - \ln \frac{s_{t,i}}{1 - \sum_{j=1}^{n-1} s_{t,j}} = -\eta (\nabla L(s_{t+1})_i - \nabla L(s_{t+1})_n)
\]

*implicit:*

\[
s_{t+1,i} = \frac{s_{t,i} e^{-\eta \nabla L(s_{t+1})_i}}{\sum_{j=1}^{n} s_{t,j} e^{-\eta \nabla L(s_{t+1})_j}}
\]
Big picture

- Batch: tradeoff between regularization and total loss
- Bregman divergence based on increasing link function $f$

$$s_{t+1} = \min \Delta_f(s, s_t) + \eta L(s)$$

$$f(s_t) = -\eta \nabla L(s_t)$$

- GD: $f(s) = s$  
  EGU: $f(s) = \ln(s)$

- Implicity update: $s_{t+1} = f^{-1}(f(s_t) - \eta \nabla L(s_{t+1}))$
  Go to rate domain, subtract gradient, come back
  - called mirror descent

- Regularization determines how parameter space is searched
## Two main families of updates

<table>
<thead>
<tr>
<th>$| \cdot |^2$</th>
<th>relative entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Widrow Hoff update</td>
<td>EG update</td>
</tr>
<tr>
<td>Perceptron</td>
<td>Winnow</td>
</tr>
<tr>
<td>Backprop</td>
<td>Weighted Majority algorithm</td>
</tr>
<tr>
<td>SVMs</td>
<td>Baysian update</td>
</tr>
<tr>
<td>Newton’s method</td>
<td>Boosting</td>
</tr>
</tbody>
</table>

### Different properties

<table>
<thead>
<tr>
<th>ignores small weights</th>
<th>smoothed 1-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>log of a Gaussian</td>
<td>information theoretic</td>
</tr>
<tr>
<td>rotation invariant</td>
<td>not rotation invariant</td>
</tr>
</tbody>
</table>
Outline

1. Motivations of updates via divergences?
2. Starting with differential equations
3. Relative entropies?
4. Entropies as barriers
Purpose of relative entropies

- Regularizer that leads to multiplicative updates.
  Ratios of logs in derivatives
  Undoing the logs give multiplicative updates
- Measure of progress in analysis
- Value of optimization problem
  - characterized the population
  - as score in evaluating markets
Send symbol $X$ on channel

$$P(X = x_i)$$

<table>
<thead>
<tr>
<th>$X$</th>
<th>$P(X = x_i)$</th>
<th>$-\log P(x_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$\frac{1}{2}$</td>
<td>1</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$\frac{1}{4}$</td>
<td>2</td>
</tr>
<tr>
<td>$x_3$</td>
<td>$\frac{1}{8}$</td>
<td>3</td>
</tr>
<tr>
<td>$x_4$</td>
<td>$\frac{1}{8}$</td>
<td>3</td>
</tr>
</tbody>
</table>

$-\log P(x_i)$ is measure of surprise in bits

$-\log 1 = 0$ no surprise

$-\log 0 = \infty$ infinite surprise

$-\log \frac{1}{2^i} = i$ $i$ bits of surprise
Entropy equals expected surprise

\[ H(X) = \sum_i P(x_i) \log \frac{1}{p(x_i)} \]

\[ = \frac{1}{2} 1 + \frac{1}{4} 2 + \frac{1}{8} 3 + \frac{1}{8} 3 \]

\[ = 1\frac{3}{4} \text{ bits} \]

\[ -p \log p - (1 - p) \log(1 - p) \]

3-dim entropy unif. in dim. \( n \)

\[ \log(n) \]
Assigns symbols a bitstring (codeword)
- any sequence of codewords must be uniquely decodable

Expected codelength

\[ L(C) = \sum_i p(x_i) \ell_C(x_i) \]

Optimal code \( C^* \)
- \( L(C^*) \) is minimum
  - Thm: \( H(X) \leq L(C^*) \leq H(X) + 1 \)
  - Thm: Huffman codes are optimal
  - More info: first five chapters of Cover & Thomas
Relative entropy between distributions \( p \) and \( q \)

\[
\Delta(p, q) = \sum_i p_i \log \frac{p_i}{q_i}
\]

\[
= \sum_i p_i \log \frac{1}{q_i} - \sum_i p_i \log \frac{1}{p_i}
\]

expected codelength of best codebook for \( q \)  
expected codelength of best codebook for \( p \)

where all expectations are wrt \( p \)
Relative entropy to the uniform distribution

\[ \Delta(p, \frac{1}{n}) = \sum p_i \log \frac{p_i}{1/n} \]
\[ = \sum p_i \log p_i + p_i \log n \]
\[ = \log n - H(p) \]
\[ \geq 0 \]

0 at center of simplex

In general

\[ \Delta(p, q) \geq 0, \]

where equality holds iff \( p = q \)
Relative entropies?

Which argument should be the variable?

not too steep at boundary
motivates EG and Boosting

steep at boundary
Two relative entropies in 3D

Both are barriers for simplex
Use of relative entropy (w. first argument as var.)

- As regularizer in motivation of update
- As measure of progress in analysis

Squared Euclidean distance “ignores” the simplex
Outline

1 Motivations of updates via divergences?
2 Starting with differential equations
3 Relative entropies?
4 Entropies as barriers
Each inequality constraint gives a term

- **Inequality** constraint $s_i \geq 0$ is handled by term $s_i \ln \frac{s_i}{s_{t,i}}$
- Entire simplex by $\sum_i s_i \ln \frac{s_i}{s_{t,i}}$
- Normalization **equality** constraint $\sum_i s_i = 1$ kept as constraint

The method

- Form one relative entropy term per inequality constraint
- Keep equality constraints on the outside
- Minimized relative entropy plus $\eta$ times loss subject to linear constraint
Keeping variable with fitness rate $w$ in interval

- For $s \in [0, 1] : \left( \ln \frac{s}{1-s} \right) = w_i$

  $$s(t + 1) := \argmax_s \quad s \, w - s \ln \frac{s}{s(t)} - (1-s) \ln \frac{1-s}{1-s(t)}$$

  $$= \frac{s(t) \, e^w}{s(t) \, e^w + 1 - s(t)}$$

- For $s \in [\ell, u] : \left( \ln \frac{s-\ell}{u-s} \right) = w_i$

  $$s(t + 1) := \argmax_s \quad s \, w - (s - \ell) \ln \frac{s-\ell}{s(t)-\ell} - (u-s) \ln \frac{u-s}{u-s(t)}$$

  $$= \frac{u(s(t)-\ell) \, e^w + \ell(u-s(t))}{u-s(t) + e^w(s(t)-\ell)}$$

If $w > 0$, then $s(t) \to u$.
If $w < 0$, then $s(t) \to \ell$
The derivation of the update - solution of hw3

\[ \frac{d}{ds} \left( sw - (s - \ell) \ln \frac{s - \ell}{s(t) - \ell} - (u - s) \ln \frac{u - s}{u - s(t)} \right) \bigg|_{s=s(t+1)} \]

\[ w - \ln \frac{s(t + 1) - \ell}{s(t) - \ell} - 1 + \ln \frac{u - s(t + 1)}{u - s(t)} + 1 = 0 \]

\[ \frac{s(t + 1) - \ell}{s(t) - \ell} = \frac{u - s(t + 1)}{u - s(t)} e^w \]

\[ s(t + 1)((s(t) - \ell) e^w + u - s(t)) = u(s(t) - \ell) e^w + \ell(u - s(t)) \]

\[ s(t + 1) = \frac{u(s(t) - \ell) e^w + \ell(u - s(t))}{(s(t) - \ell) e^w + u - s(t)} \]
Keeping variables in intervals and normalizing

$s_i \in [\ell_i, u_i]$ and $\sum_i s_i = 1$:

$$s(t+1) := \arg\max_{\sum_i s_i = 1} \mathbf{s} \cdot \mathbf{w} - \sum_{i=1}^{n} (s_i - \ell_i) \ln \frac{s_i - \ell_i}{s_i(t) - \ell_i} - (u_i - s_i) \ln \frac{u_i - s_i}{u_i - s_i(t)}$$

$$s_i(t+1) = \frac{u_i(s_i(t) - \ell_i) e^{w_i + \lambda} + \ell_i(u_i - s_i(t))}{(s_i(t) - \ell_i) e^{w_i + \lambda} + u_i - s(t)}$$

$\lambda$ is determine with binary search
Is this how nature normalizes?
Does nature normalize?
Negatives shares?

In replicator/Hedge shares lie on simplex

\[ s(t + 1) = \arg\max_{\sum_i s_i = 1} w \cdot s - \Delta(s, s(t)) \]

\[ s_i(t + 1) = \frac{s_i(t) e^{w_i}}{Z} \]

Two share vectors \( s^+ \geq 0 \) and \( s^- \geq 0 \)

\[
\begin{align*}
[s^+(t + 1), s^-(t + 1)] &= \arg\max_{\sum_i s_i^+ + s_i^- = 1} \\
&\left( w \cdot s^+ + (-w) \cdot s^- - \Delta(s^+, s^+(t)) - \Delta(s^-, s^-(t)) \right) \\
&\quad \left( (s^+ - s^-) \cdot w \right)
\end{align*}
\]

\[ s_i^+(t + 1) = \frac{s_i^+(t) e^{w_i}}{Z'} \quad \text{and} \quad s_i^-(t + 1) = \frac{s_i^-(t) e^{-w_i}}{Z'} \]

where \( Z' = \sum_{i=1}^{n} \left( s_i^+(t) e^{w_i} + s_i^-(t) e^{-w_i} \right) \)
Negative shares continued

- So $s_i^+$ has rate $w_i$ / fitness factor $W_i^+ = e^{w_i}$
- and $s_i^-$ has rate $-w_i$ / fitness factor $W_i^- = e^{-w_i} = \frac{1}{W_i^+}$.
- Each species split into two
  - one goes forward the other one backward in time
- $s^+ - s^-$ lies in the diamond $\{ r \in \mathbb{R}^n : \sum_{i=1}^n |r_i| \leq 1 \}$
  which is the convex hull of the $2n$ corners $\pm e_i$
- How does nature implement negative numbers?
Conclusion

- How do you characterize the family of updates used by nature?
- Is it relative entropy regularization subject to constraints?
- How does nature normalize?
- Does it employ negative numbers?
- What is the continuous form of the relative entropy?