Algorithm $\text{EG}_\pm(U,(s^+, s^-), \eta)$

**Parameters:**
- $L$: a loss function from $\mathbb{R} \times \mathbb{R}$ to $[0, \infty)$,
- $U$: the total weight of the weight vectors,
- $s^+$ and $s^-$: a pair of start vectors in $[0, 1]^S$, with $\sum_{s^+} (s^+_i + s^-_i) = 1$, and
- $\eta$: a learning rate in $[0, \infty)$.

**Initialization:** Before the first trial, set $w_1^+ = Us^+$ and $w_1^- = Us^-$. 

**Prediction:** Upon receiving the $t$th instance $x_t$, give the prediction

$$\hat{y}_t = (w_t^+ - w_t^-) \cdot x_t,$$

**Update:** Upon receiving the $t$th outcome $y_t$, update the weights according to the rules

$$w_{t+1}^+ = U \left( \frac{w_t^+ r_t^+}{\sum_{i=1}^U (w_t^+_i r_t^+_i + w_t^-_i r_t^-_i)} \right),$$

$$w_{t+1}^- = U \left( \frac{w_t^- r_t^-}{\sum_{i=1}^U (w_t^+_i r_t^+_i + w_t^-_i r_t^-_i)} \right),$$

where

$$r_t^+ = \exp(-\eta L^+(\hat{y}_t, y_t) \psi_{x_t}),$$

$$r_t^- = \exp(\eta L^-(\hat{y}_t, y_t) \psi_{x_t}),$$

No $U$ in exponent


**FIG. 3.** Exponential gradient algorithm with positive and negative weights $\text{EG}_\pm(U,(s^+, s^-), \eta)$. 

Therefore, $w_t^+ \cdot x_t = (w_t^+ - w_t^-) \cdot x_t$. Hence, the predictions of $\text{EG}_\pm$ on $S$ and $\text{EG}_\pm$ on $S'$ are identical, so $\text{EG}_\pm$ is a result of applying a simple transformation to $\text{EG}_L$.

This transformation leads to an algorithm that in effect uses a weight vector $w_t^+ - w_t^-$, which can contain negative components. Further, by using the scaling factor $U$, we can make the weight vector $w_t^+ - w_t^-$ range over all vectors $w \in \mathbb{R}$ for which $||w||_1 \leq U$. Although $||w_t^+||_1 + ||w_t^-||_1$ is always exactly $U$, vectors $w_t^+ - w_t^-$ with $||w_t^+ - w_t^-||_1 < U$ result simply from having both $w_t^+, w_t^- > 0$ and $w_t^+, w_t^- > 0$ for some $i$. For other examples of reductions of this type, see Littlestone et al. (1995).

The parameters of $\text{EG}_\pm$ are a loss function $L$, a scaling factor $U$, a pair $(s^+, s^-)$ of start vectors in $[0, 1]^S$ with $\sum_{s^+} (s^+_i + s^-_i) = 1$, and a learning rate $\eta$. We simply write $\text{EG}_\pm$ for $\text{EG}_\pm$ where $L$ is the square loss function. As the start vectors for $\text{EG}_\pm$, one would typically use $s^+ = s^- = (1/(2N), ..., 1/(2N))$. This gives $w_1^+ - w_1^- = 0$. A typical learning rate function could be $\eta = 1/(3U^2X^2)$ where $X$ is an estimated upper bound for the maximum $L_{\infty}$ norm $\max_i ||x_i||_\infty$ of the instances. More detailed theoretical results are given in Theorem 5.11.

Again, we introduce one particular variable learning rate version of $\text{EG}_\pm$. We use the name $\text{EGV}_\pm$ for the algorithm that is as $\text{EG}_\pm$ except that (3.10) and (3.11) are replaced by