Correction of the EG+- update

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

Parameters:

L: a loss function from $\mathbf{R} \times \mathbf{R}$ to $[0, \infty)$,

U: the total weight of the weight vectors,

$\mathbf{s}^+$ and $\mathbf{s}^-$: a pair of start vectors in $[0, 1]^N$, with $\sum_{i=1}^{N} (s_i^+ + s_i^-) = 1$, and

η: a learning rate in $[0, \infty)$.

Initialization: Before the first trial, set $\mathbf{w}_1^+ = U\mathbf{s}^+$ and $\mathbf{w}_1^- = U\mathbf{s}^-$.

Prediction: Upon receiving the tth instance $\mathbf{x}_t$, give the prediction

$$\hat{y}_t = (\mathbf{w}_t^+ - \mathbf{w}_t^-) \cdot \mathbf{x}_t.$$

Update: Upon receiving the tth outcome y_t, update the weights according to the rules

$$w_{t+1,i}^+ = U \cdot \frac{w_{t,i}^+ r_{t,i}^+}{\sum_{j=1}^{N} (w_{t,j}^+ r_{t,j}^+ + w_{t,j}^- r_{t,j}^-)} \tag{3.8}$$

$$w_{t+1,i}^- = U \cdot \frac{w_{t,i}^- r_{t,i}^-}{\sum_{j=1}^{N} (w_{t,j}^+ r_{t,j}^+ + w_{t,j}^- r_{t,j}^-)}, \tag{3.9}$$

where

$$r_{t,i}^+ = \exp(-\eta L'_{y_t}(\hat{y}_t)\, \cancel{U} x_{t,i}) \tag{3.10}$$

No U in exponent

$$r_{t,i}^- = \exp(\eta L'_{y_t}(\hat{y}_t)\, \cancel{U} x_{t,i}) = \frac{1}{r_{t,i}^+} \tag{3.11}$$

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

therefore, $\mathbf{w}'_t \cdot \mathbf{x}'_t = (\mathbf{w}_t^+ - \mathbf{w}_t^-) \cdot \mathbf{x}_t$. Hence, the predictions of $\mathrm{EG}^{\pm}_{L}$ on S and EG_L on S' are identical, so $\mathrm{EG}^{\pm}_{L}$ is a result of applying a simple transformation to EG_L. This transformation leads to an algorithm that in effect uses a weight vector $\mathbf{w}_t^+ - \mathbf{w}_t^-$, which can contain negative components. Further, by using the scaling factor U, we can make the weight vector $\mathbf{w}_t^+ - \mathbf{w}_t^-$ range over all vectors $\mathbf{w} \in \mathbf{R}$ for which $\|\mathbf{w}\|_1 \leqslant U$. Although $\|\mathbf{w}_t^+\|_1 + \|\mathbf{w}_t^-\|_1$ is always exactly U, vectors $\mathbf{w}_t^+ - \mathbf{w}_t^-$ with $\|\mathbf{w}_t^+ - \mathbf{w}_t^-\|_1 < U$ result simply from having both $w_{t,i}^+ > 0$ and $w_{t,i}^- > 0$ for some i. For other examples of reductions of this type, see Littlestone *et al.* (1995).

The parameters of $\mathrm{EG}^{\pm}_{L}$ are a loss function L, a scaling factor U, a pair $(\mathbf{s}^+, \mathbf{s}^-)$ of start vectors in $[0, 1]^N$ with $\sum_{i=1}^{N} (s_i^+ + s_i^-) = 1$, and a learning rate η. We simply write $\mathrm{EG}^{\pm}$ for $\mathrm{EG}^{\pm}_{L}$ where L is the square loss function. As the start vectors for $\mathrm{EG}^{\pm}$, one would typically use $\mathbf{s}^+ = \mathbf{s}^- = (1/(2N), ..., 1/(2N))$. This gives $\mathbf{w}_1^+ - \mathbf{w}_1^- = \mathbf{0}$. A typical learning rate function could be $\eta = 1/(3U^2X^2)$ where X is an estimated upper bound for the maximum L_∞ norm $\max_t \|\mathbf{x}_t\|_\infty$ of the instances. More detailed theoretical results are given in Theorem 5.11.

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