Tracking the Best Expert

Mark Herbster & Manfred Warmuth (Machine Learning, 1998)

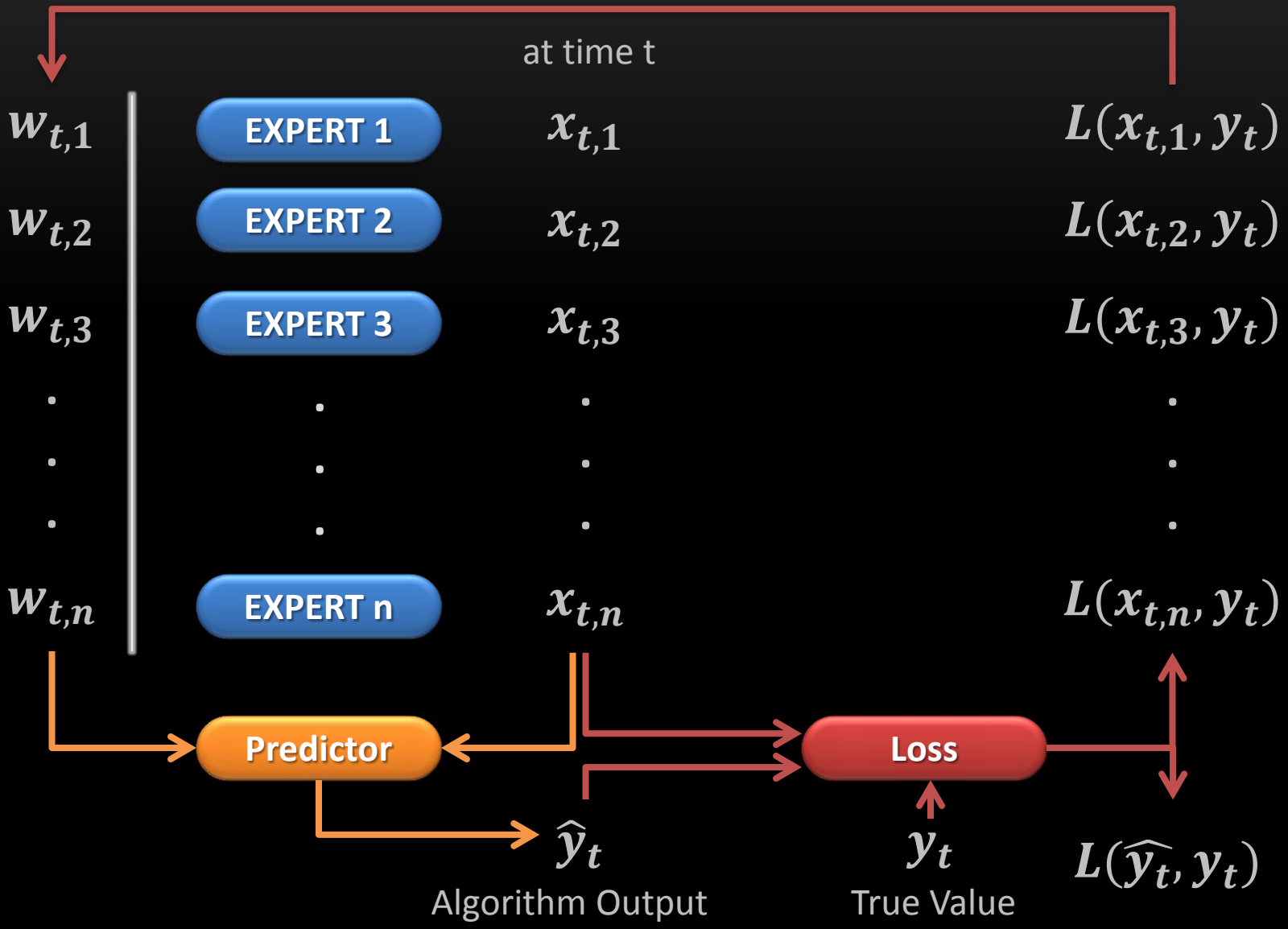
Ohil K Manyam

March 2nd, 2011

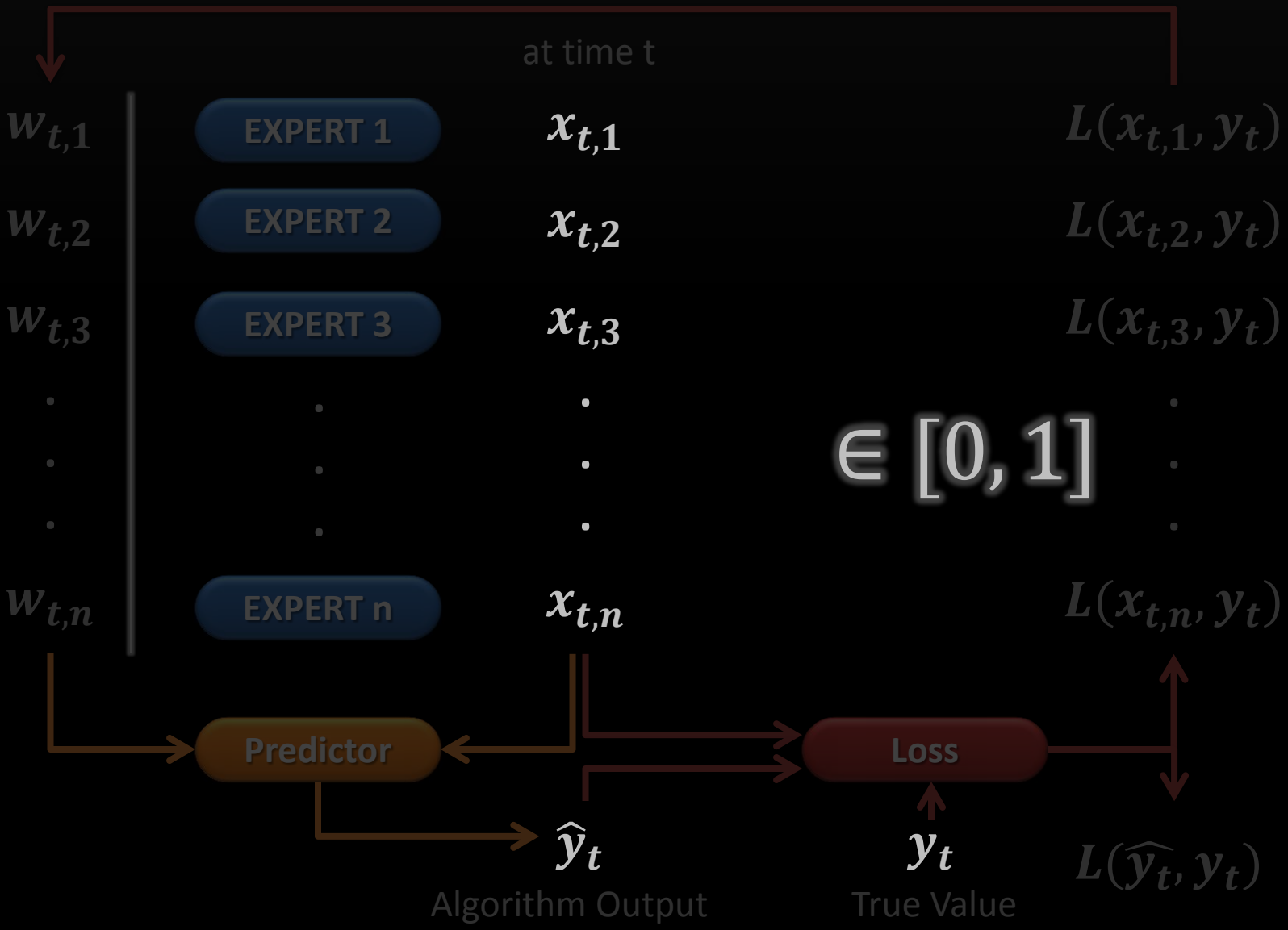
Outline

- Prediction with expert advice – recap
- New setting - segment experts
- Existing solutions
- New solutions
  - Fixed Share
  - Variable Share
  - Proximity Variable Share
- Experiments & results

Prediction with expert advice



Prediction with expert advice



Prediction with expert advice

Minimize additional loss of algorithm over loss of the best expert

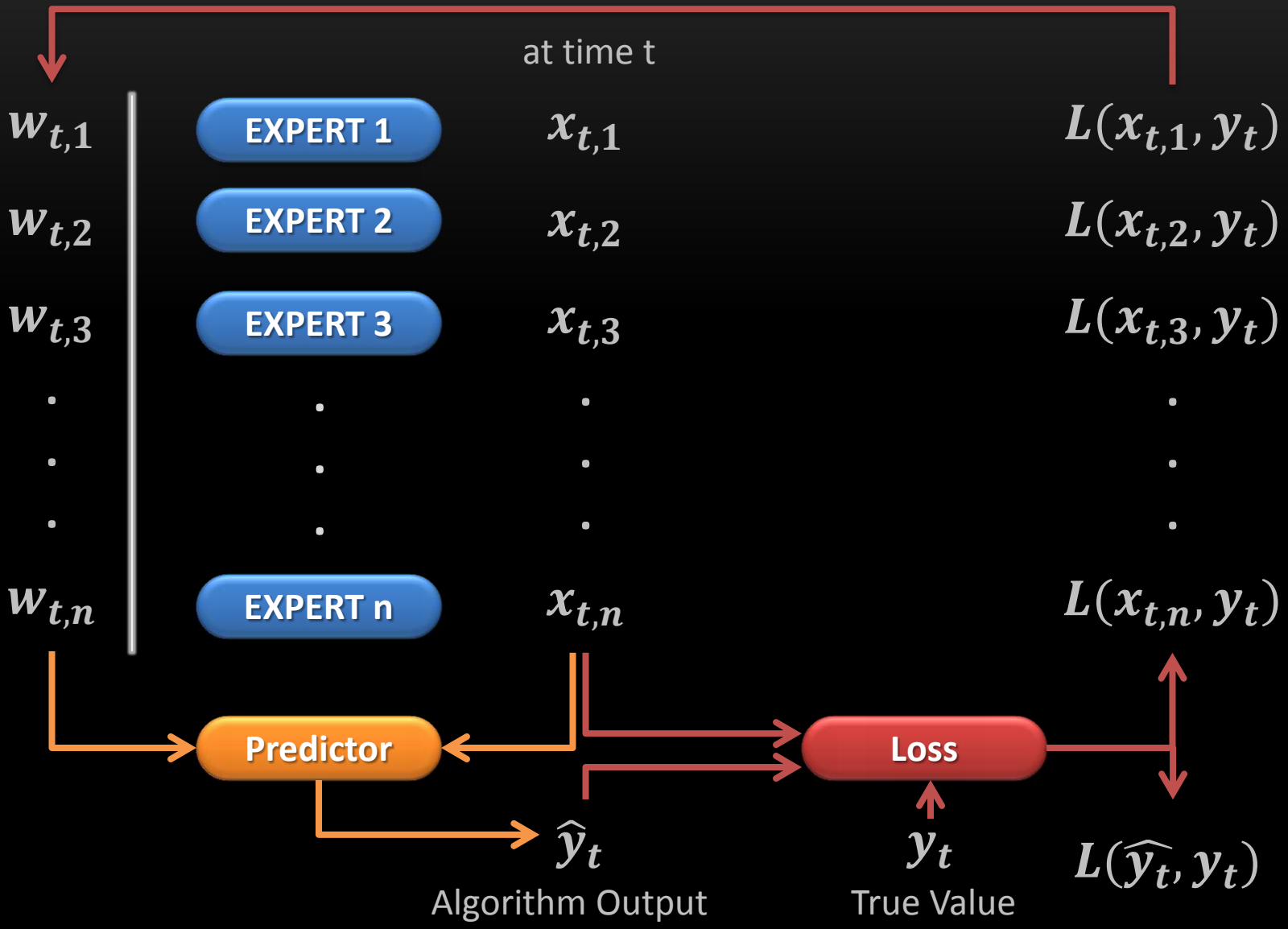
$$L' = \sum_{t=1}^{T} L(\hat{y}_t, y_t) - \min_i \left(\sum_{t=1}^{T} L(x_{t,i}, y_t) \right)$$

- Solution
  - Halving Algorithm
  - Hedge Algorithm

- Additional Loss $\sim O(c \ln n)$
  (c depends only on L for large class of functions)

Segment experts

Segment experts



Segment experts

- l trials
- k expert shifts
- n experts

Partition

"Sequence" S – divided into $(k+1)$ "segments"

Segment experts

Segment experts

- $\vec{e} = (e_0, e_1, \dots, e_k)$
- $1 \leq e_i \leq n$
- $e_i \neq e_{i+1}$

Segment boundary

- $\vec{t} = (t_1, t_2, \dots, t_k)$
- $1 \leq t_i \leq l$
- $[t_i .. t_{i+1})$ - i^{th} segment

Segment experts

- $P_{l,n,k,\vec{t},\vec{e}}(S)$ – describes the partitioning
- Algorithm's loss

$$L(S,A) = \sum_{t=1}^{l} L(y_t, \widehat{y_t})$$

- Loss of partition

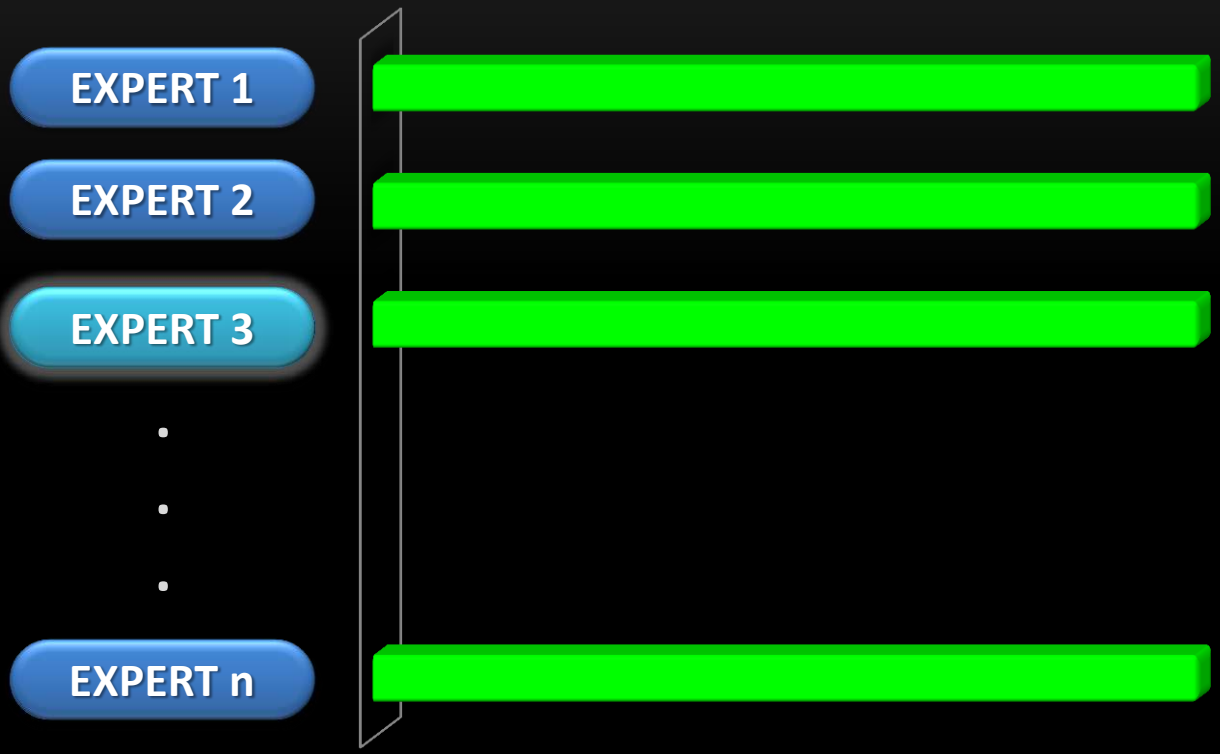
$$L\left(P_{l,n,k,\vec{t},\vec{e}}(S)\right) = \sum_{i=0}^{k} L([t_i..t_{i+1}), e_i)$$

- Minimize

$$L(S,A) - \min_{\vec{t},\vec{e}} L\left(P_{l,n,k,\vec{t},\vec{e}}(S)\right)$$

Using the old algorithm

Using the old algorithm



$t = 0$

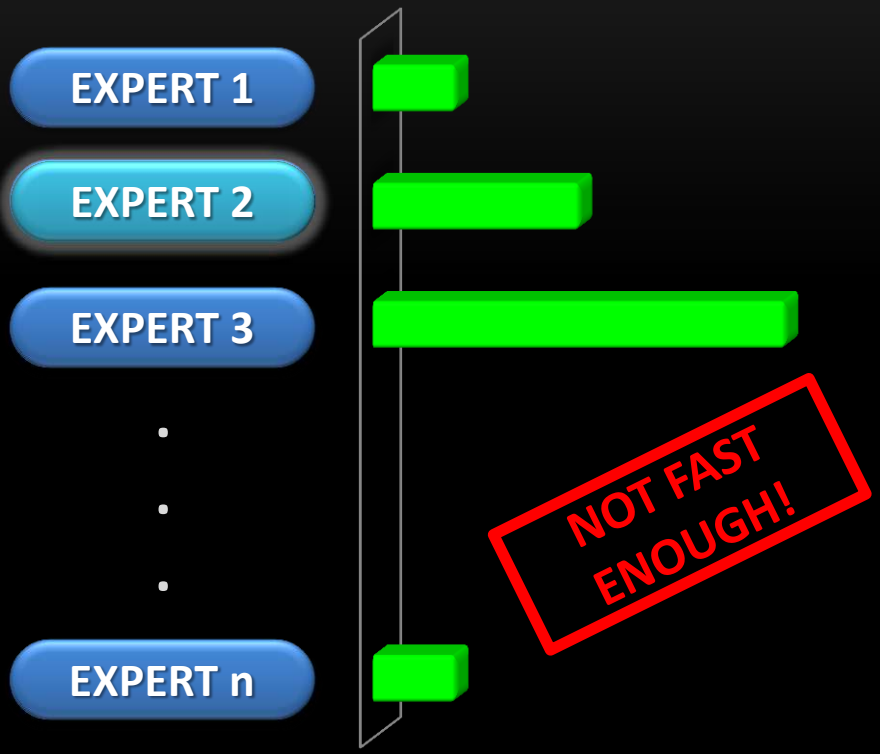
$$w_{t+1,i} = w_{t,i} e^{-\eta L(y_t, x_{t,i})}$$

Using the old algorithm

$t = t_1$

$$w_{t+1,i} = w_{t,i} e^{-\eta L(y_t, x_{t,i})}$$

Using the old algorithm



$t = t_2$

$$w_{t+1,i} = w_{t,i} e^{-\eta L(y_t, x_{t,i})}$$

Using the old algorithm

again…

Casting into single expert setting

- Total “partition experts”

$$\binom{l-1}{k} n(n-1)^k = O\left(n^{k+1}\left(\frac{el}{k}\right)^k\right)$$

Casting into single expert setting

- Total “partition experts”

$$\binom{l-1}{k} n(n-1)^k = O\left(n^{k+1}\left(\frac{el}{k}\right)^k\right)$$

- Regret bound $\sim O(c \ln n)$

$$R \leq c\left[(k+1)\ln n + k \ln\left(\frac{l}{k}\right) + k\right]$$

- Problem?
  - ✘ Inefficient : too many partition-expert weights to update
  - ✘ Regret grows with sequence length

Prior work

Prior work

- Littlestone & Warmuth (1994)
  - "The weighted majority algorithm"
  - Fixed multiplicative update of weights
  - W_{ML} : Update weight only when > threshold
- Auer & Warmuth (1998)
  - Modification of Winnow
  - Learns shifting disjunctions

New algorithms

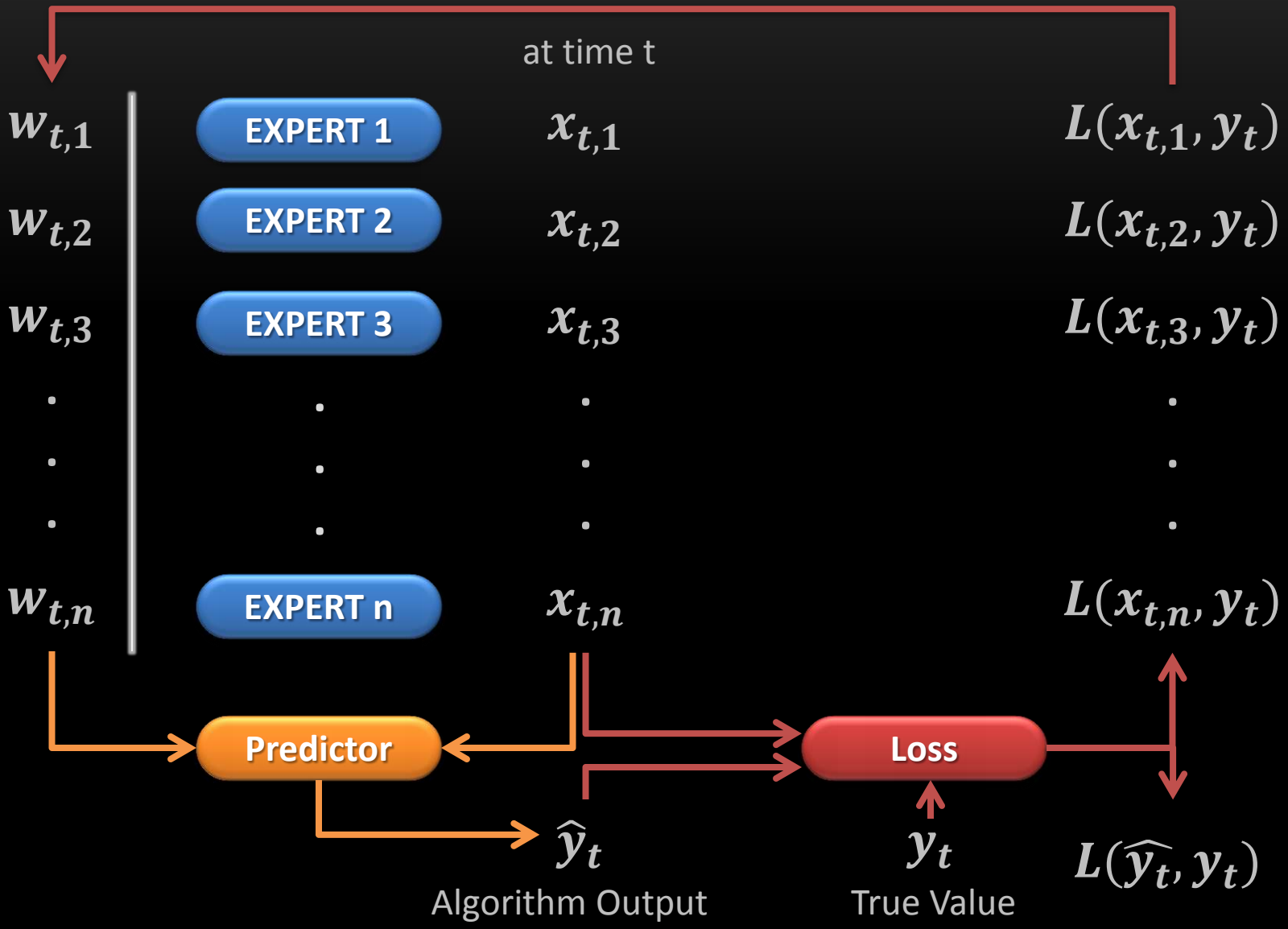
New algorithms

- Static Expert

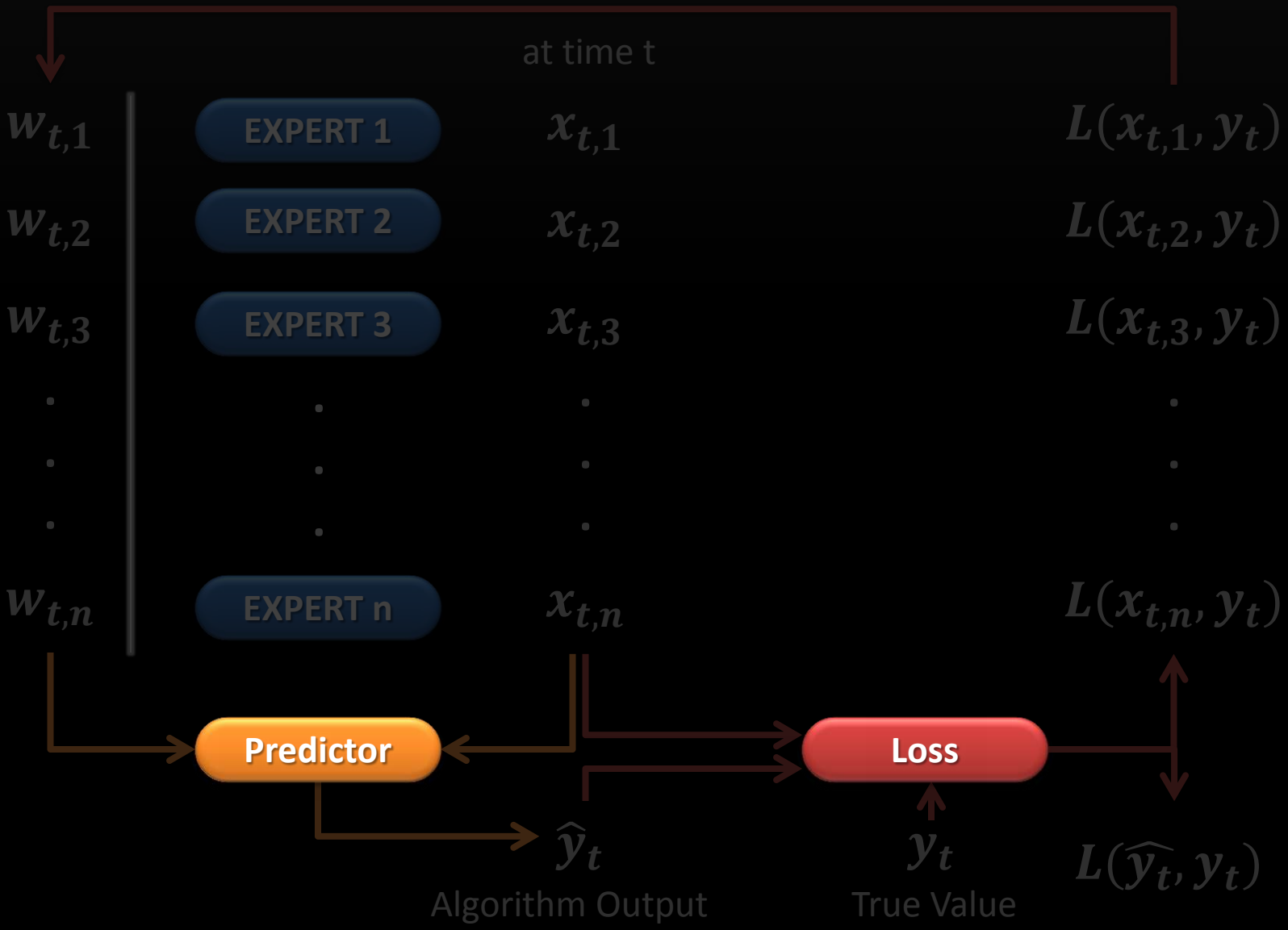
- Weight sharing
  - Fixed Share
  - Variable Share
  - Proximity Variable Share

… but before that

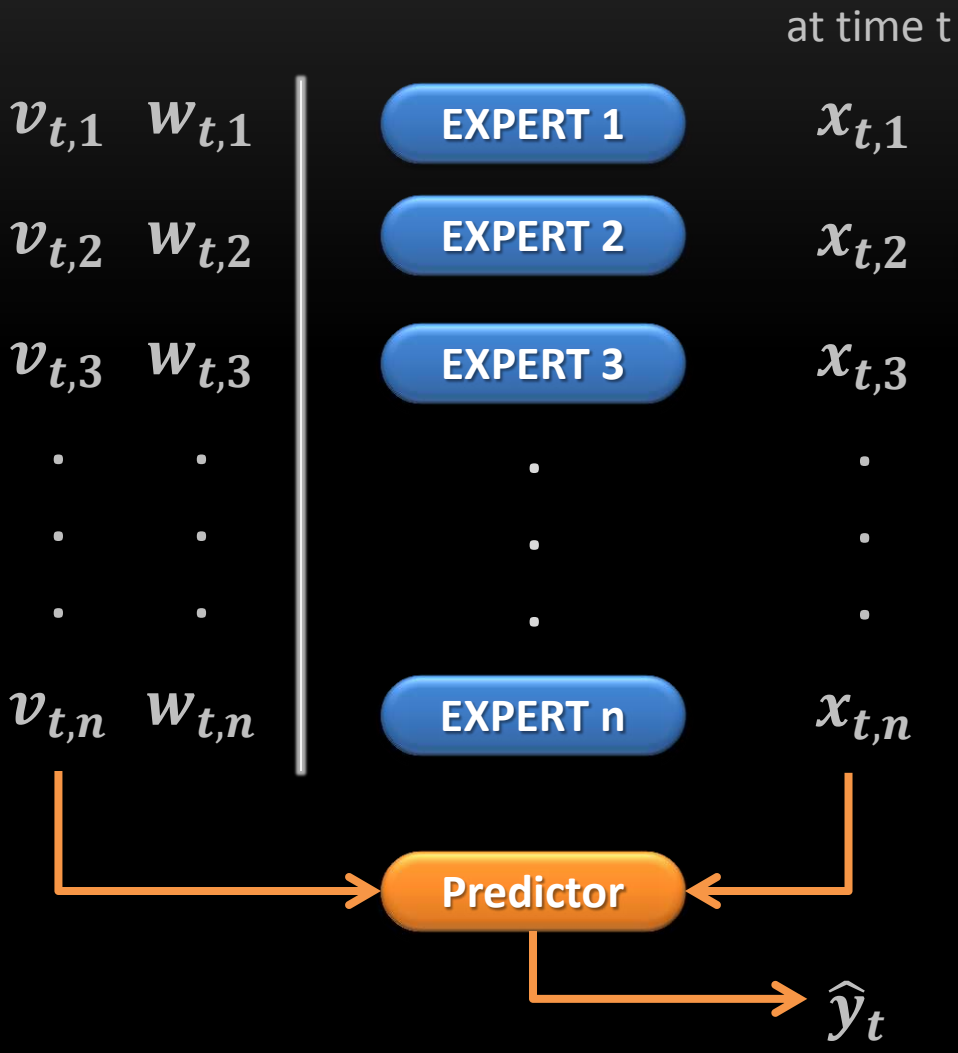
Prediction with expert advice



Prediction with expert advice



Predictor function



1

$$pred_{wmean}(\vec{v}, \vec{x}) = \sum_{i=1}^{n} v_i x_i$$

Predictor Function

- Vovk, V (1998), A game of prediction with expert advice, *Journal of Computer and System Sciences*

- $L_0(z) \triangleq L(0,z)$; $L_1(z) \triangleq L(1,z)$ – both monotonic
- $L_0^{-1}(z)$ & $L_1^{-1}(z)$ – inverse functions

$$\Delta(y) = -c \ln\left(\sum_{i=1}^{n} v_i e^{-\eta L(y,x_i)}\right)$$

2

$$pred_{vovk}(\vec{v},\vec{x}) = \frac{L_0^{-1}\big(\Delta(0)\big) + L_1^{-1}(\Delta(1))}{2}$$

Loss Functions

1. Square loss

$$L_{sp}(p,q) = (p-q)^2$$

2. Relative Entropy loss

$$L_{ent}(p,q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}$$

3. Hellinger loss

$$L_{hel}(p,q) = \frac{1}{2}\left(\left(\sqrt{1-p} - \sqrt{1-q}\right)^2 + \left(\sqrt{p} - \sqrt{q}\right)^2\right)$$

4. Absolute loss

$$L_{abs}(p,q) = |p-q|$$

Setting parameters

- L and $pred$ are (c, η) realizable if

$$L(pred(\vec{v}, \vec{x}), y) \leq -c \ \ln \sum_{i=1}^{n} v_i e^{-\eta L(y, x_i)}$$

	$pred_{wmean}$	$pred_{vovk}$
L_{sq}	1/2	2
L_{ent}	1	1
L_{hel}	1	$\sqrt{2}$
L_{abs}	needs tuning	–

$$c = \frac{1}{\eta}$$

$$c = \frac{1}{1 - e^{-\eta}}$$

η values for loss-prediction function pairings

New algorithms

Basics

- Parameters $0 < \eta, c$ and $0 \leq \alpha \leq 1$
- Initialize weights $w_{1,i}^{s} = 1/n$
- $v_{t,i} = w_{t,i}^{s} \,/\, \Sigma_{i=1}^{n} w_{t,i}^{s}$
- $\hat{y} = pred(\overrightarrow{v_t}, \overrightarrow{x_t})$
- $w_{t,i}^{m} = w_{t,i}^{s} e^{-\eta L(y_t, x_{t,i})}$

Static expert

- $w_{t,i}^{m} = w_{t,i}^{s} e^{-\eta L(y_t, x_{t,i})}$

$$w_{t+1,i}^{s} = w_{t,i}^{m}$$

Fixed share

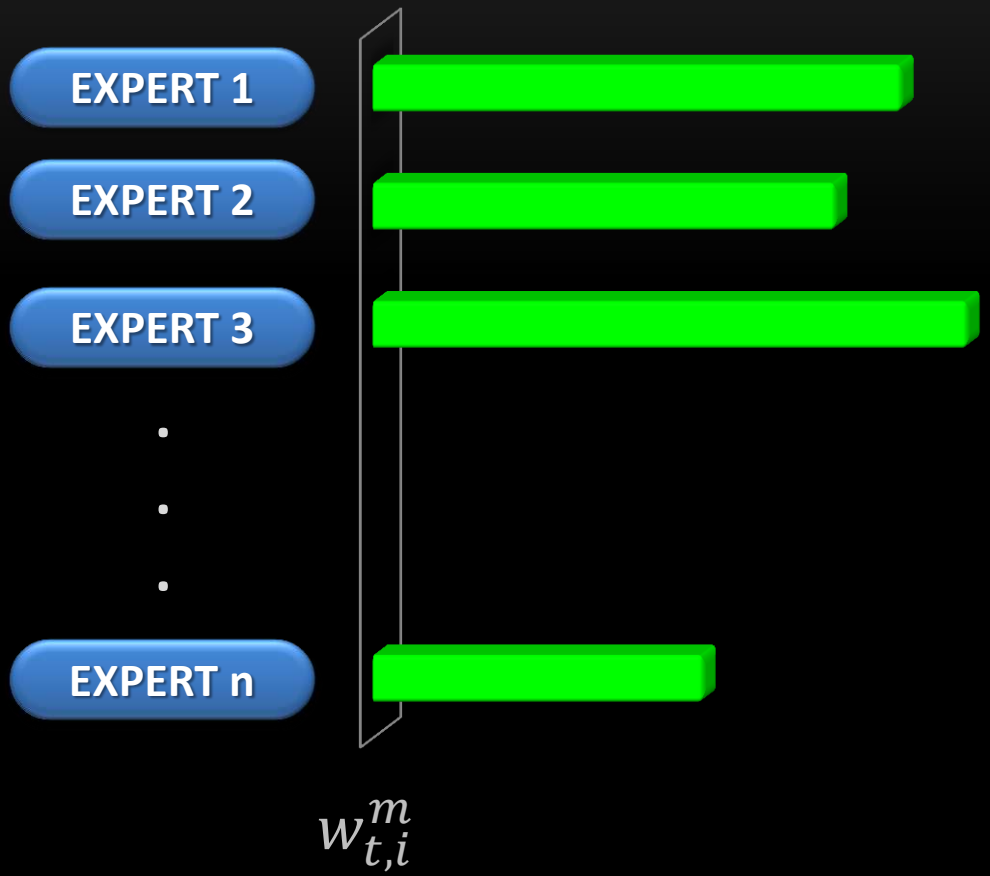
Fixed share

- $w_{t,i}^{m} = w_{t,i}^{s} e^{-\eta L(y_t, x_{t,i})}$

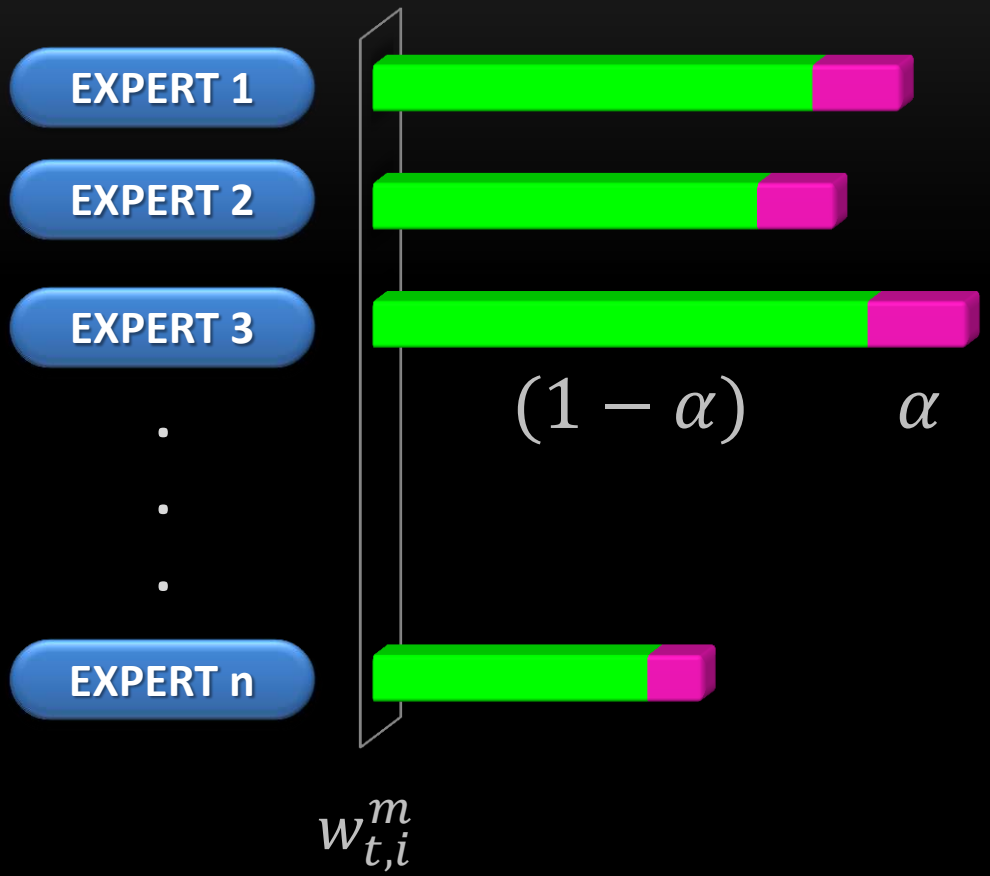
$$pool = \sum_{i=1}^{n} \alpha w_{t,i}^{m}$$

$$w_{t+1,i}^{s} = (1-\alpha) w_{t,i}^{m} + \frac{1}{n-1}(pool - \alpha w_{t,i}^{m})$$

Fixed share



Fixed share



Fixed share

$$w_{t+1,i}^{S}$$

Fixed share

- Loss bound

$$L(S,A) \leq c\ln(n) + c\eta L\left(P_{l,n,k,\vec{t},\vec{e}}(S)\right) + c(l-k-1)\ln(1/(1-\alpha)) + ck[\ln(1/\alpha) + \ln(n-1)]$$

- Grows with sequence length
- $\alpha = \frac{k}{l-1}$
  - Determined experimentally
  - "Rate of change of expert per trial"
  - $\hat{k} \leq k, \hat{l} \geq l : \alpha = \frac{\hat{k}}{\hat{l}-1}$

Fixed share

✘ Regret bound $\sim O\left((k+1)\ln n + k\ln\frac{l}{k}\right)$

✓ n weights

– O(n) time to update all weights in a trial

$$pool = \sum_{i=1}^{n} \alpha w_{t,i}^{m}$$

$$w_{t+1,i}^{s} = (1-\alpha)w_{t,i}^{m} + \frac{1}{n-1}\left(pool - \alpha w_{t,i}^{m}\right)$$

Fixed share

- Good expert keeps sharing weight

Variable share

Variable share

- $w_{t,i}^{m} = w_{t,i}^{s} e^{-\eta L(y_t, x_{t,i})}$

$$pool = \sum_{i=1}^{n} \left(1 - (1-\alpha)^{L(y_t, x_{t,i})}\right) w_{t,i}^{m}$$

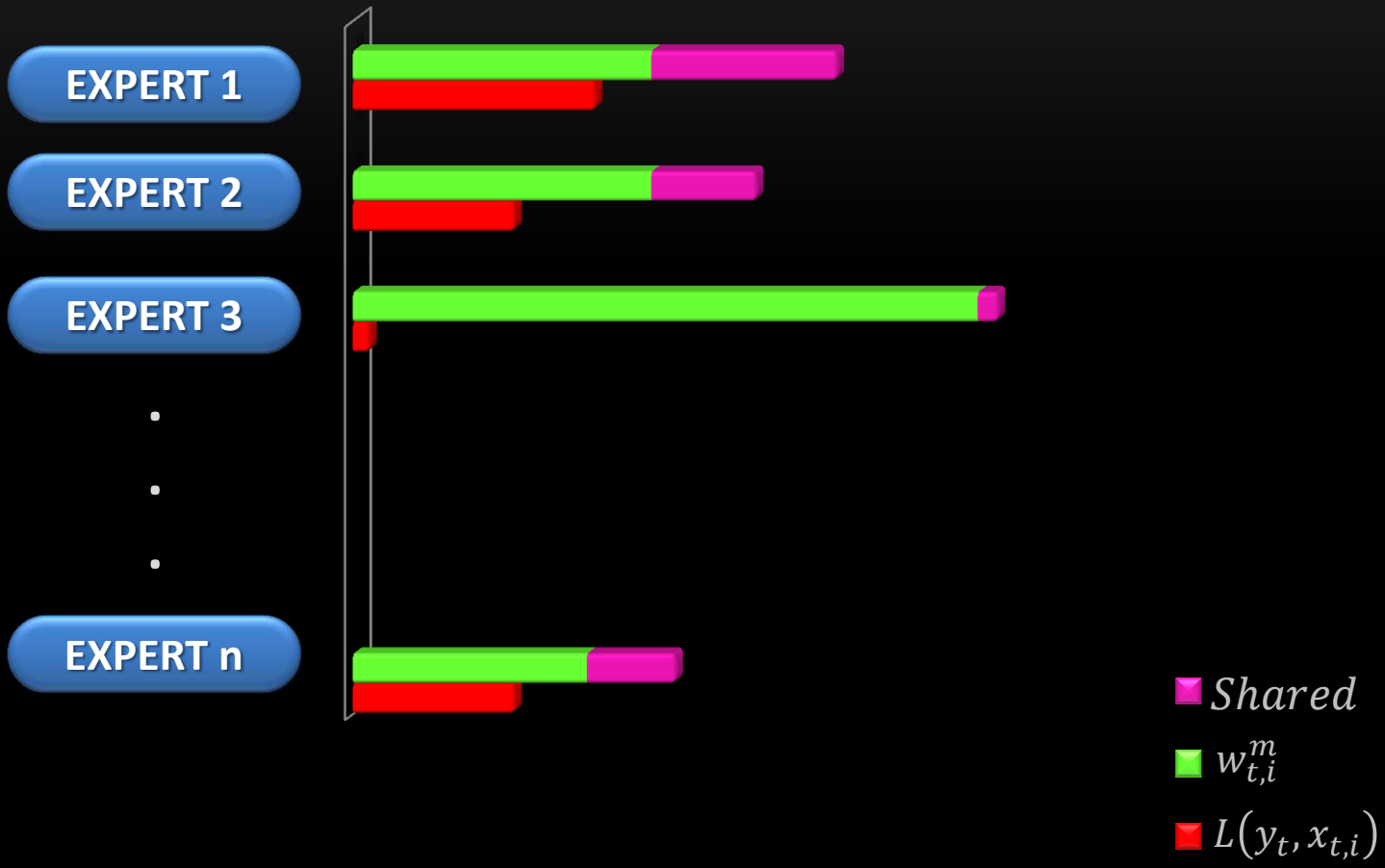
$$w_{t+1,i}^{s} = (1-\alpha)^{L(y_t, x_{t,i})} w_{t,i}^{m} + \frac{1}{n-1}\left(pool - \left(1 - (1-\alpha)^{L(y_t, x_{t,i})}\right) w_{t,i}^{m}\right)$$

Variable share

- Fraction of weight shared depends on loss
  - $f = \left(1 - (1-\alpha)^{L(y_t, x_{t,i})}\right)$
  - Recall : $0 \leq \alpha \leq 1$
  - High loss → Bad expert → large f
  - Low loss → Good expert → small f

Variable share

Variable share



Variable share

Variable share

- Loss bound ($L \in [0,1]$)

$$L(S,A) \leq c\ln(n) + c\left[\eta + \ln\frac{1}{1-\alpha}\right] L\left(P_{l,n,k,\vec{t},\vec{e}}(S)\right) + ck\left[\eta + \ln\frac{1}{\alpha} + \ln\frac{1}{1-\alpha} + \ln(n-1)\right]$$

- $\alpha = \frac{\hat{k}}{2\hat{k}+\hat{L}}$: $L\left(P_{l,n,k,\vec{t},\vec{e}(S)}\right) \leq \hat{L}; \quad \hat{k} \in \mathbb{R}^+$

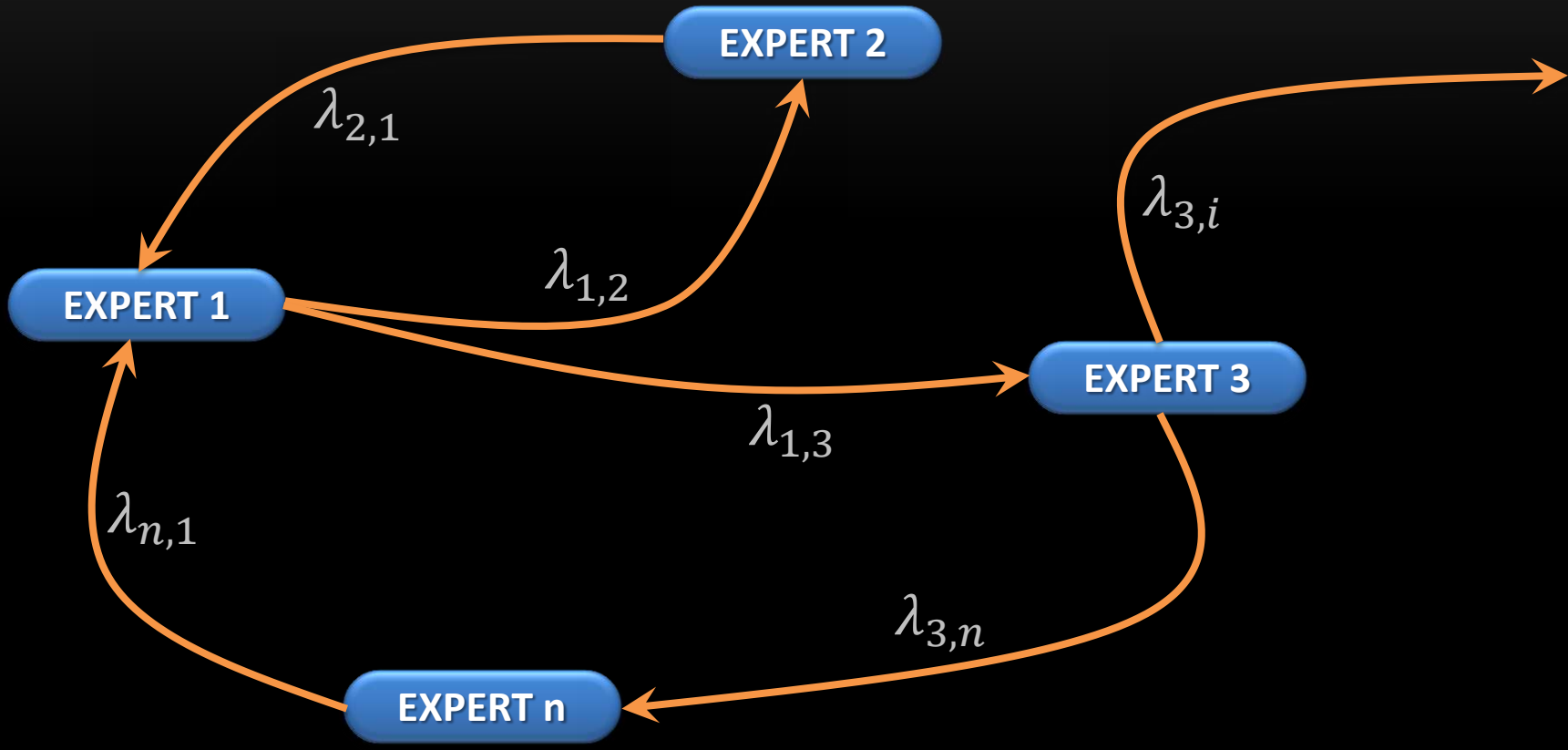
Variable share

- ✓ $L(S,A) \sim O\left((k+1)\ln n + k\ln\frac{L}{k}\right)$
  - L – loss of best k-partition
- ✓ n weights - $O(n)$ time to update all per trial

Variable share

Proximity variable share

Proximity variable share



Proximity variable share

- Parameters
  - $\lambda_i^0 > 0$: Initial weights
  - $\lambda_{j,k} > 0;\ \ \Sigma_{k \neq j}^{n} \lambda_{j,k} = 1$: Shared from expert j to k
- $w_{1,1}^S = \lambda_1^0, \dots, w_{1,n}^S = \lambda_n^0$
- Same as before
  - $pred(\overrightarrow{v_t}, \overrightarrow{x_t})$
  - $w_{t,i}^m = w_{t,i}^S e^{-\eta L(y_t, x_{t,i})}$

Proximity variable share

- Variable share

$$w^{s}_{t+1,i} = (1-\alpha)^{L(y_t, x_{t,i})} w^{m}_{t,i} + \frac{1}{n-1}\left(pool - \left(1-(1-\alpha)^{L(y_t, x_{t,i})}\right) w^{m}_{t,i}\right)$$

Proximity variable share

- Variable share

$$w^{s}_{t+1,i} = (1-\alpha)^{L(y_t, x_{t,i})} w^{m}_{t,i} + \Sigma^{n}_{j \neq i} \frac{1}{n-1} \left(1 - (1-\alpha)^{L(y_t, x_{t,j})}\right) w^{m}_{t,j}$$

Proximity variable share

- Proximity variable share

$$w^{s}_{t+1,i} = (1-\alpha)^{L(y_t, x_{t,i})} w^{m}_{t,i} + \Sigma^{n}_{j \neq i} \lambda_{j,i} \left(1 - (1-\alpha)^{L(y_t, x_{t,j})}\right) w^{m}_{t,j}$$

✘ $O(n^2)$ to update all weights in a trial

Proximity variable share

- Loss bound ($L \in [0,1]$)

$$L(S,A) \leq \ln \lambda_{e_0}^0 + \sum_{i=1}^{k} \ln \lambda_{e_{i-1},e_i}$$
$$+c\left[\eta + \ln \frac{1}{1-\alpha}\right] L\left(P_{l,n,k,\vec{t},\vec{e}}(S)\right)$$
$$+ck\left[\eta + \ln \frac{1}{\alpha} + \ln \frac{1}{1-\alpha} + \cancel{\ln(n-1)}\right]$$

- $\alpha = \frac{\hat{k}}{2\hat{k}+\hat{L}}$: $L\left(P_{l,n,k,\vec{t},\vec{e}(S)}\right) \leq \hat{L}; \quad \hat{k} \in \mathbb{R}^+$

  (Same as Variable share)

Proximity variable share

- $\lambda_{e_0} = \frac{1}{n}$; $\lambda_{e_{i-1},e_i} = \frac{1}{n-1}$

$$L(S,A) \sim O\left((k+1)\ln n + k \ln \frac{L}{k}\right)$$

- Can be $O(k)$

Comparison

Comparison

- Static expert
  - Exponentially many weights
  - Loss bound depends on sequence length
- Fixed share
  - Number of weights = Number of experts
  - Loss bound depends on sequence length
  - Good expert keeps sharing weight

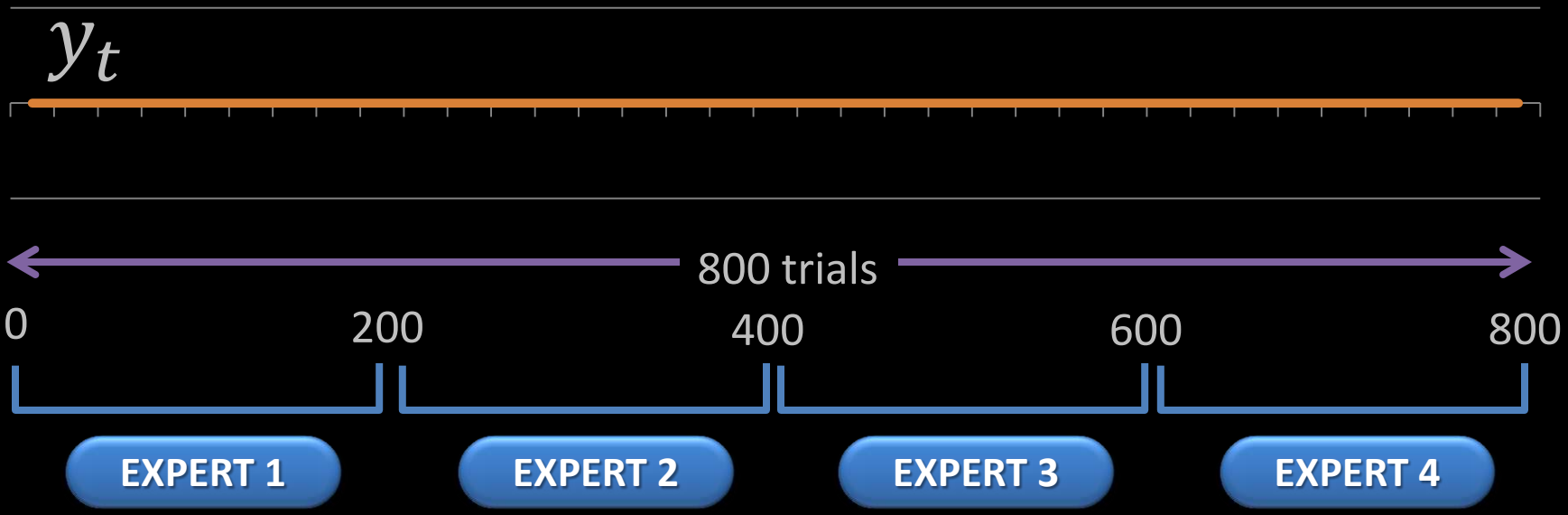
Comparison

- Variable share
  - Loss bound independent of sequence length
  - Expert distributes its weight equally to others
- Proximity variable share
  - Selective weight sharing
  - Knowledge of expert transitions

Experiments

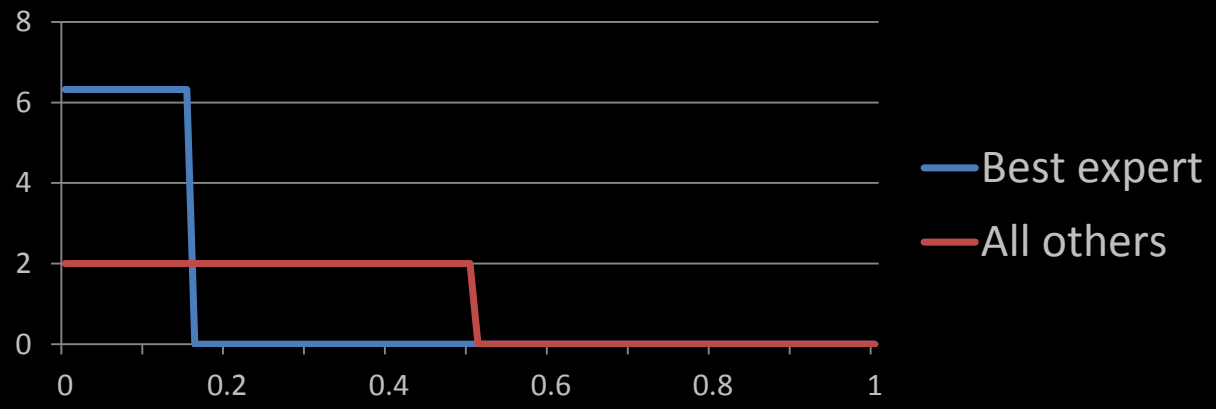
Experiments

- 64 experts
- 800 trials – 4 segments (200 trials each)
- For all trials, $y_t = 0$



Experiments

- Prediction $x_{t,i}$ sampled (uniform distribution)
  - Best expert $\left(0, 1/2\sqrt{10}\right)$
  - All others $(0, 1/2)$



- Expected loss per trial ($L_{sq}\left(y_t, x_{t,i}\right)$)
  - Best expert : 1/120
  - All others : 1/12

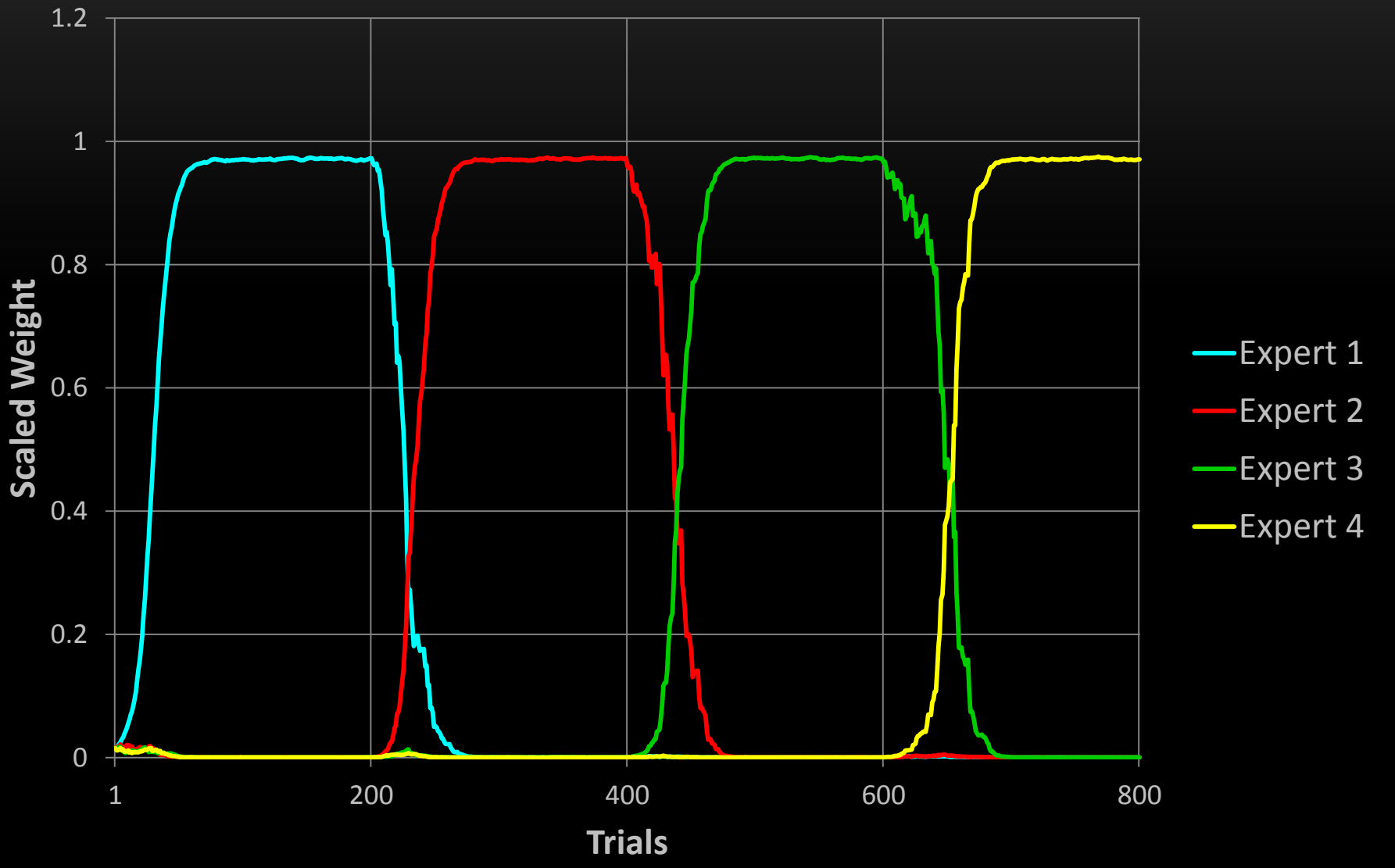
Experiments

- $\eta = 2\ ;\ c = 1/2$
- Fixed share
  - $\alpha = 3/799$ (“Rate of change of expert per trial”)
- Variable share
  - $\hat{L} = 800/120 = 6.73$ (Loss of best partition)
  - $\alpha = k/\left(2k + \hat{L}\right) = 3/(6 + 6.73)$

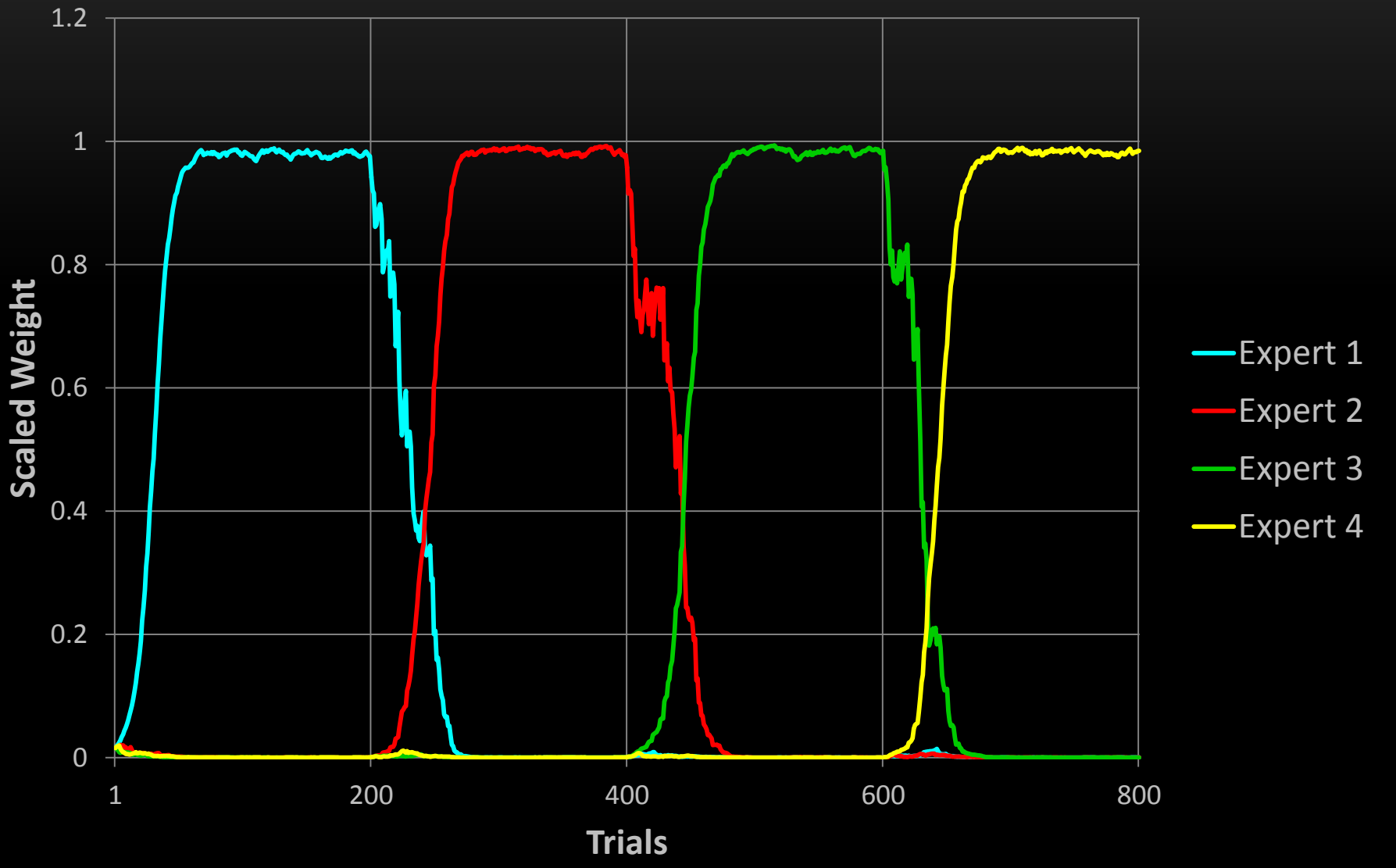
Results

Results – Static expert

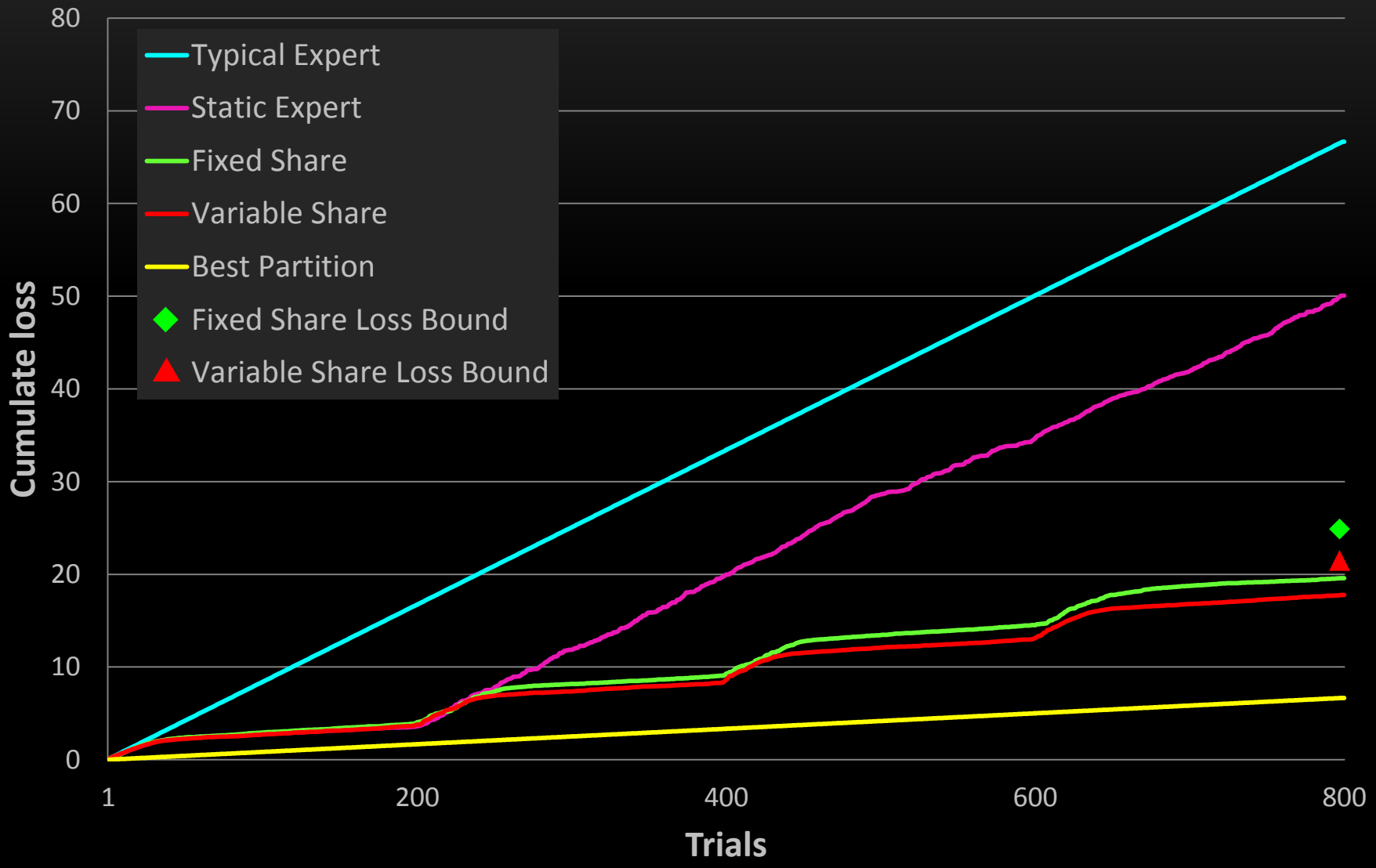
Results – Fixed share



Results – Variable share



Results – Variable share



Conclusion

- Experts change during trials
- Static expert – weight loss only
- Key idea – weight sharing
- Fixed share
  - Share fixed fraction of weight
- Variable share
  - Give up more weight if more loss
- Proximity variable share
  - Donate weight to specific experts
- Will make mistakes at transition point

Questions?

Thank you!