

Follow the leader if you can, Hedge if you must

Tim van Erven

NIPS, 2013

Joint work with: Steven de Rooij
Peter Grünwald
Wouter Koolen

Outline

- Follow-the-Leader:
 - works well for **'easy' data**: few leader changes, i.i.d.
 - but not robust to worst-case data
- Exponential weights with simple tuning:
 - **robust**, but does not exploit easy data
- Second-order bounds:
 - robust against **worst case** + can exploit **i.i.d. data**
 - but do not exploit few leader changes in general
- FlipFlop: **robust + as good as FTL**

Sequential Prediction with Expert Advice

- K experts sequentially predict data x_1, x_2, \dots
- Goal: predict (almost) as well as the best expert on average
- Applications:
 - online convex optimization
 - predicting electricity consumption
 - predicting air pollution levels
 - spam detection
 - ...

Set-up: Repeated Game

- Every round $t = 1, \dots, T$:
 1. Predict probability distribution $w_t = (w_{t,1}, \dots, w_{t,K})$ on experts
 2. Observe expert losses $\ell_t = (\ell_{t,1}, \dots, \ell_{t,K}) \in [0, 1]^K$
 3. Our loss is $w_t \cdot \ell_t = \sum_k w_{t,k} \ell_{t,k}$

Goal: minimize **regret**

$$\sum_{t=1}^T w_t \cdot \ell_t - L^*$$

where

Loss of the best expert

$$L^* = \min_k \sum_{t=1}^T \ell_{t,k}$$

Follow-the-Leader

- Deterministically choose the expert that has predicted best in the past:

$$w_{t,k^*} = 1 \text{ where } k^* = \arg \min_k \sum_{s=1}^{t-1} \ell_{t,k}$$

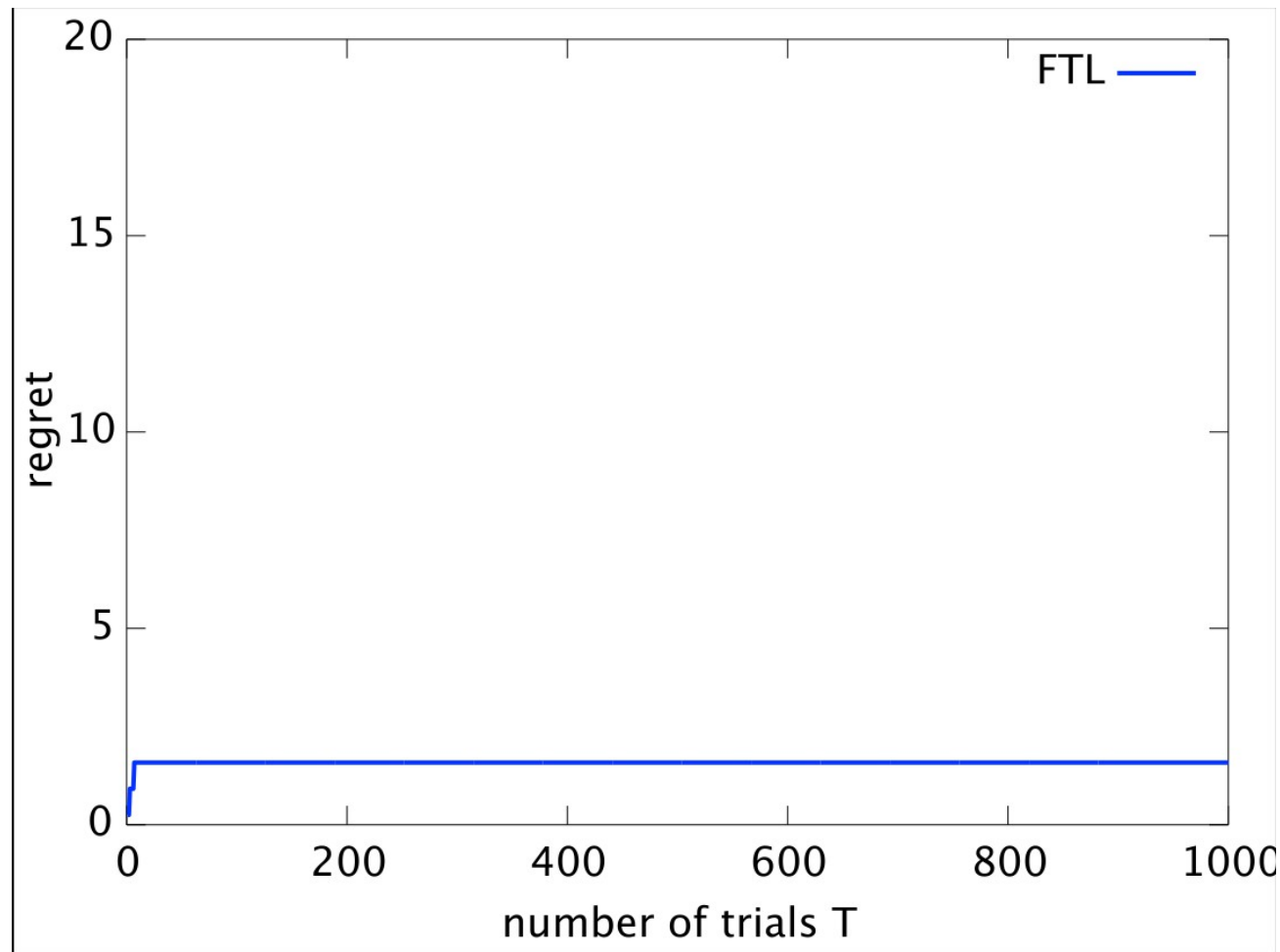
- Equivalently:

$$w_t = \arg \min_w \mathbb{E}_{k \sim w} \left[\sum_{s=1}^{t-1} \ell_{t,k} \right]$$

FTL: the Good News

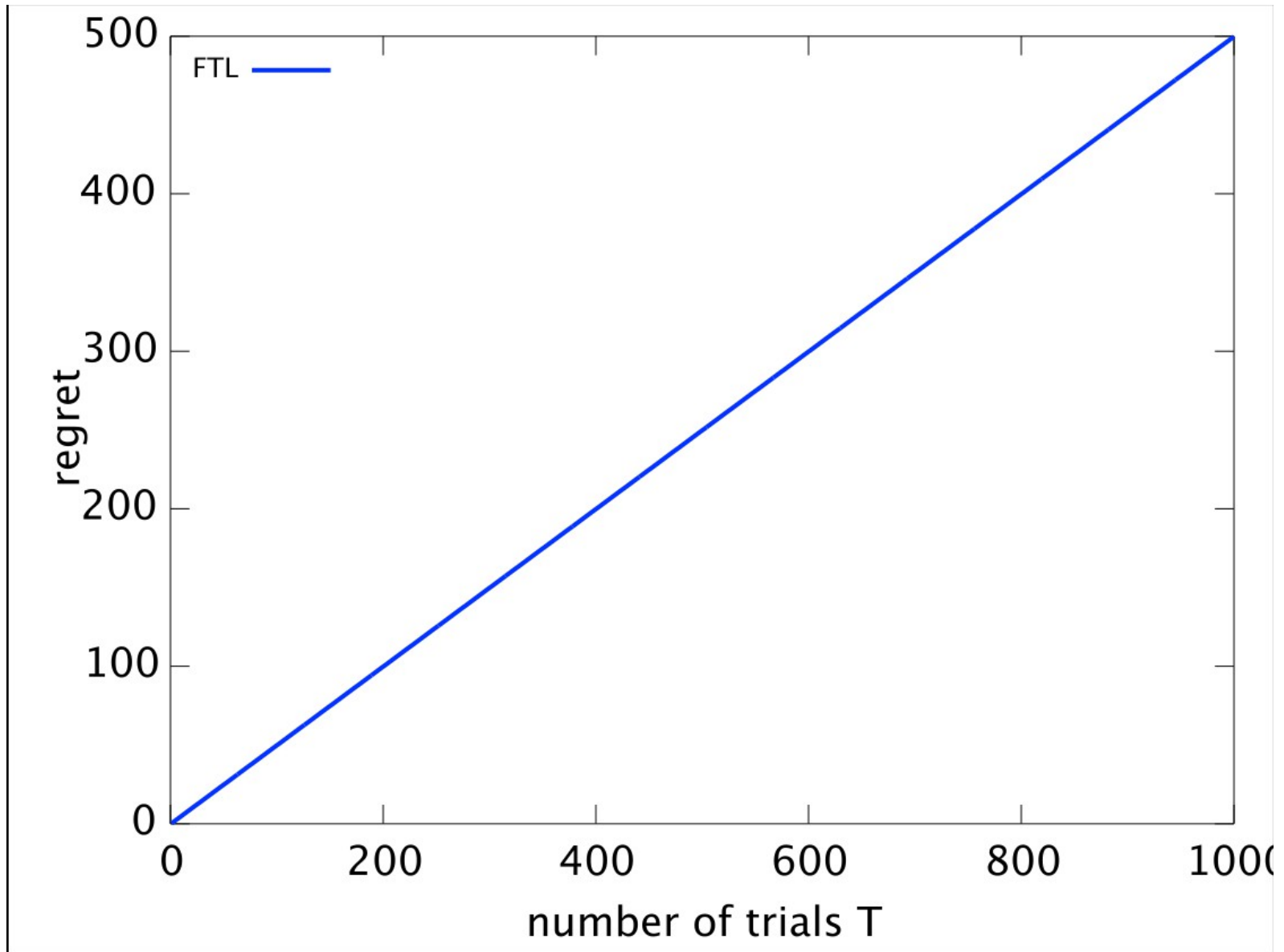
- Regret bounded by nr of leader changes
- Proof sketch:
 - If the leader does not change, our loss is the same as the loss of the leader, so the regret stays the same
 - If the leader does change, our regret increases at most by 1 (range of losses)
- Works well for i.i.d. losses, because the leader changes only finitely many times w.h.p.

FTL on IID Losses



- 4 experts with Bernoulli 0.1, 0.2, 0.3, 0.4 losses

FTL Worst-case Losses



Exponential Weights

- Follow-the-Leader:

$$w_t = \arg \min_w \mathbb{E}_{k \sim w} \left[\sum_{s=1}^{t-1} \ell_{t,k} \right]$$

- **Exponential weights**: add KL divergence from uniform distribution as a regularizer

$$w_t = \arg \min_w \mathbb{E}_{k \sim w} \left[\sum_{s=1}^{t-1} \ell_{t,k} \right] + \frac{1}{\eta} D(w \| u)$$

- $\eta \rightarrow \infty$: recover FTL (aggressive learning)
- As η closer to 0 : closer to uniform distribution (more conservative learning)

Simple Tuning: the Good News

- Worst-case optimal for $\eta = \sqrt{8 \ln(K)/T}$:

$$\text{Regret} \leq \sqrt{T \ln(K)/2}$$

- Proof idea:

- approximate our loss: $w_t \cdot \ell_t = \sum_k w_{t,k} \ell_{t,k}$
- by the **mix loss**:

$$m_t = \frac{-1}{\eta} \ln \sum_k w_{t,k} e^{-\eta \ell_{t,k}}$$

- and bound the **approximation error**:

$$\delta_t = w_t \cdot \ell_t - m_t$$

Simple Tuning: the Good News

our loss = mix loss + approx. error

$$w_t \cdot \ell_t = m_t + \delta_t$$

- Cumulative mix loss is close to L^* :

$$L^* \leq \sum_{t=1}^T m_t \leq L^* + \frac{\ln K}{\eta}$$

- Hoeffding's bound:

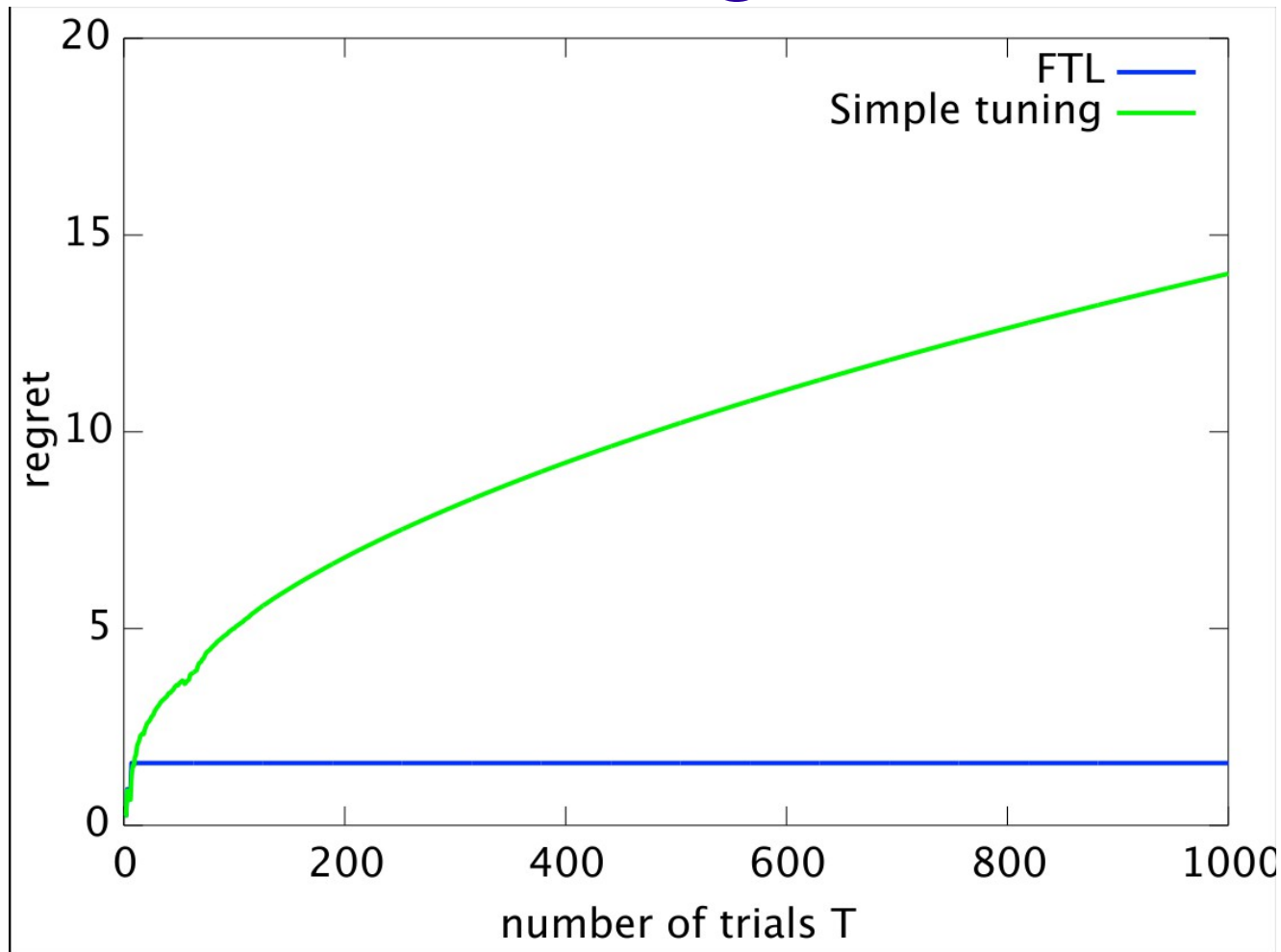
$$\delta_t \leq \frac{\eta}{8}$$

Balances the two terms

- Together:

$$\sum_{t=1}^T w_t \cdot \ell_t - L^* \leq \frac{\ln K}{\eta} + \frac{\eta T}{8} \xrightarrow{\eta = \sqrt{8 \ln K / T}} \sqrt{T \ln(K) / 2}$$

Lost Advantages of FTL



- Simple tuning does much worse than FTL on i.i.d. losses

Simple Tuning: the Bad News

- The bad news:
 - $\eta = \sqrt{8 \ln(K)/T}$ = conservative learning
 - In practice, better when learning rate does not go to 0 with T ! [DGGS, 2013]
 - Lost advantages of FTL!
- We want to exploit **luckiness**:
 - robust against worst-case losses; but
 - if the data are 'easy', we should learn faster!

Luckiness: Exploiting Easy Data

- Improvement for small losses:

$$\text{Regret} = O\left(\sqrt{L^* \ln(K)}\right)$$

- Second-order Bounds:

- [CBMS, 2007] and AdaHedge: $O\left(\sqrt{\sum_t v_t \ln(K)}\right)$
- Related bound by [HK, 2008]

variance of w_t



Luckiness: Exploiting Easy Data

- Improvement for small losses:

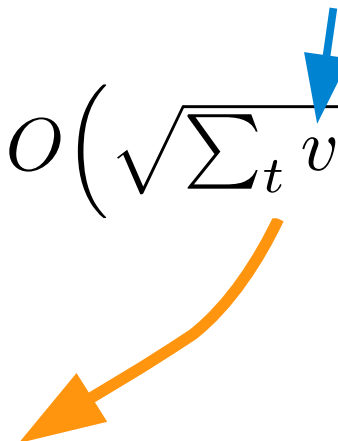
$$\text{Regret} = O\left(\sqrt{L^* \ln(K)}\right)$$

- Second-order Bounds:


- [CBMS, 2007] and AdaHedge: $O\left(\sqrt{\sum_t v_t \ln(K)}\right)$
- Related bound by [HK, 2008]

$$O\left(\sqrt{\frac{L^*(T - L^*)}{T} \ln(K)}\right)$$

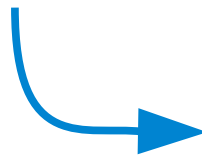
variance of w_t



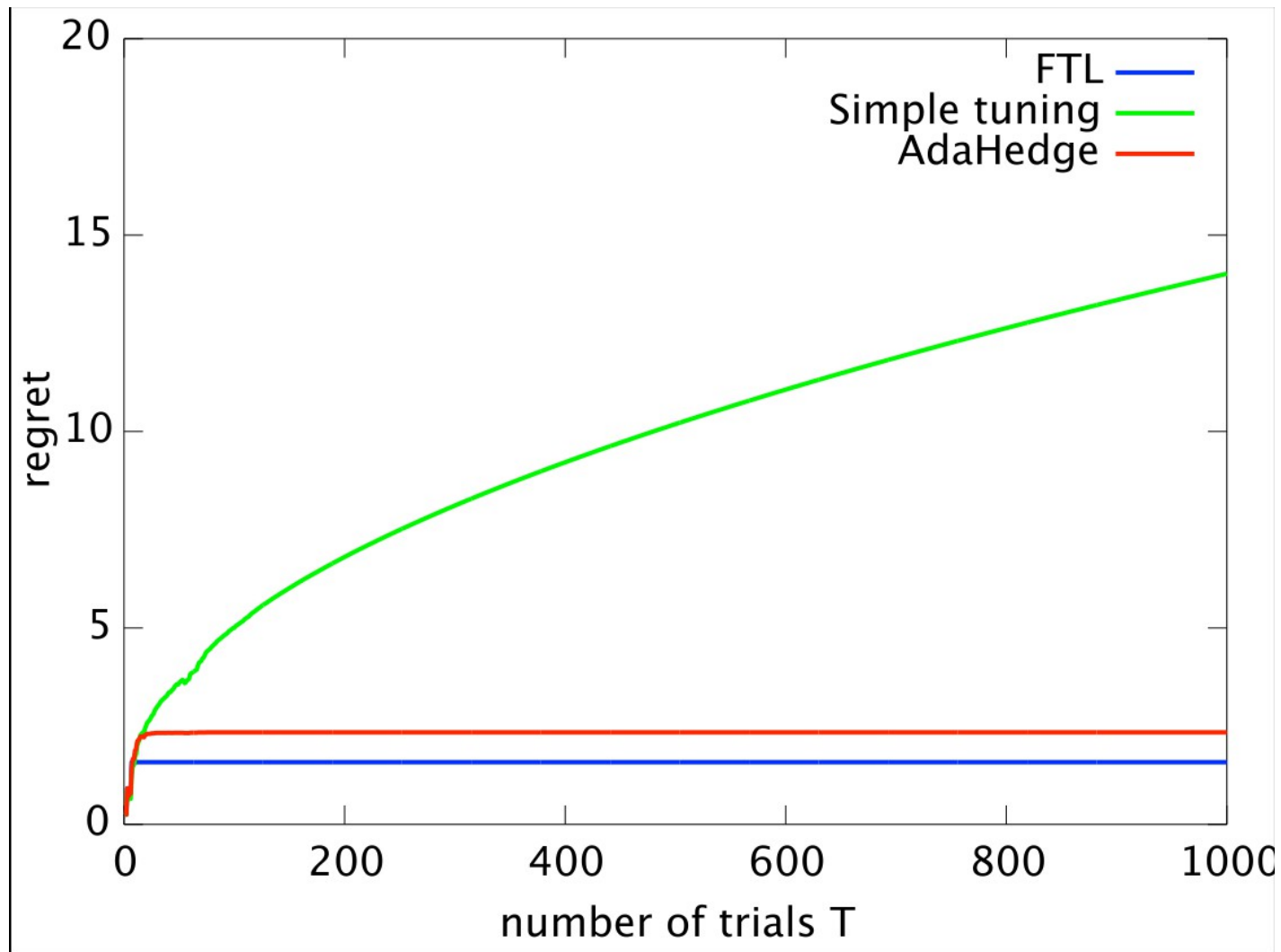
2nd-order Bounds: I.I.D. Data

- Regret bound: $O\left(\sqrt{\sum_t v_t \ln(K)}\right)$  variance of w_t

- For IID data, w_t **concentrates fast** on best expert:


$$\sum_t v_t \leq C \longrightarrow \text{Regret} \leq C'$$

2nd-order Bounds: I.I.D. Data



Recover FTL benefits for i.i.d. data

CBMS: Proof Idea

our loss = mix loss + approx. error

$$w_t \cdot \ell_t = m_t + \delta_t$$

- Cumulative mix loss is close to L^* :

$$L^* \leq \sum_{t=1}^T m_t \leq L^* + \frac{\ln K}{\eta}$$

- **Bernstein's** bound:

$$\delta_t \leq \frac{1}{2}\eta v_t + \text{lower order terms}$$

- Together:

balancing

$$\text{Regret} \leq \frac{\ln K}{\eta} + \frac{1}{2}\eta \sum_{t=1}^T v_t \xrightarrow{\eta = \sqrt{2 \ln(K) / \sum_t v_t}} O\left(\sqrt{\sum_t v_t \ln(K)}\right)$$

AdaHedge: Proof Idea

our loss = mix loss + approx. error

$$w_t \cdot \ell_t = m_t + \delta_t$$

- Cumulative mix loss is close to L^* :

$$L^* \leq \sum_{t=1}^T m_t \leq L^* + \frac{\ln K}{\eta}$$

- **No bound:**

$$\delta_t = \delta_t$$

- Together:

balancing

$$\eta = \frac{\ln(K)}{\sum_t \delta_t}$$

$$\text{Regret} \leq \frac{\ln K}{\eta} + \sum_t \delta_t \longrightarrow O\left(\sum_t \delta_t\right) = O\left(\sqrt{\sum_t v_t \ln K}\right)$$

AdaHedge: Proof Idea

our loss = mix loss + approx. error

$$w_t \cdot \ell_t = m_t + \delta_t$$

- Cumulative mix loss is close to L^* :

$$L^* \leq \sum_{t=1}^T m_t \leq L^* + \frac{\ln K}{\eta}$$

- **No bound:**

$$\delta_t = \delta_t$$

NB Bernstein's bound is pretty sharp, so in practice **CBMS** \approx **AdaHedge** up to constants.

- Together:

balancing

$$\eta = \frac{\ln(K)}{\sum_t \delta_t}$$

$$\text{Regret} \leq \frac{\ln K}{\eta} + \sum_t \delta_t \longrightarrow O\left(\sum_t \delta_t\right) = O\left(\sqrt{\sum_t v_t \ln K}\right)$$

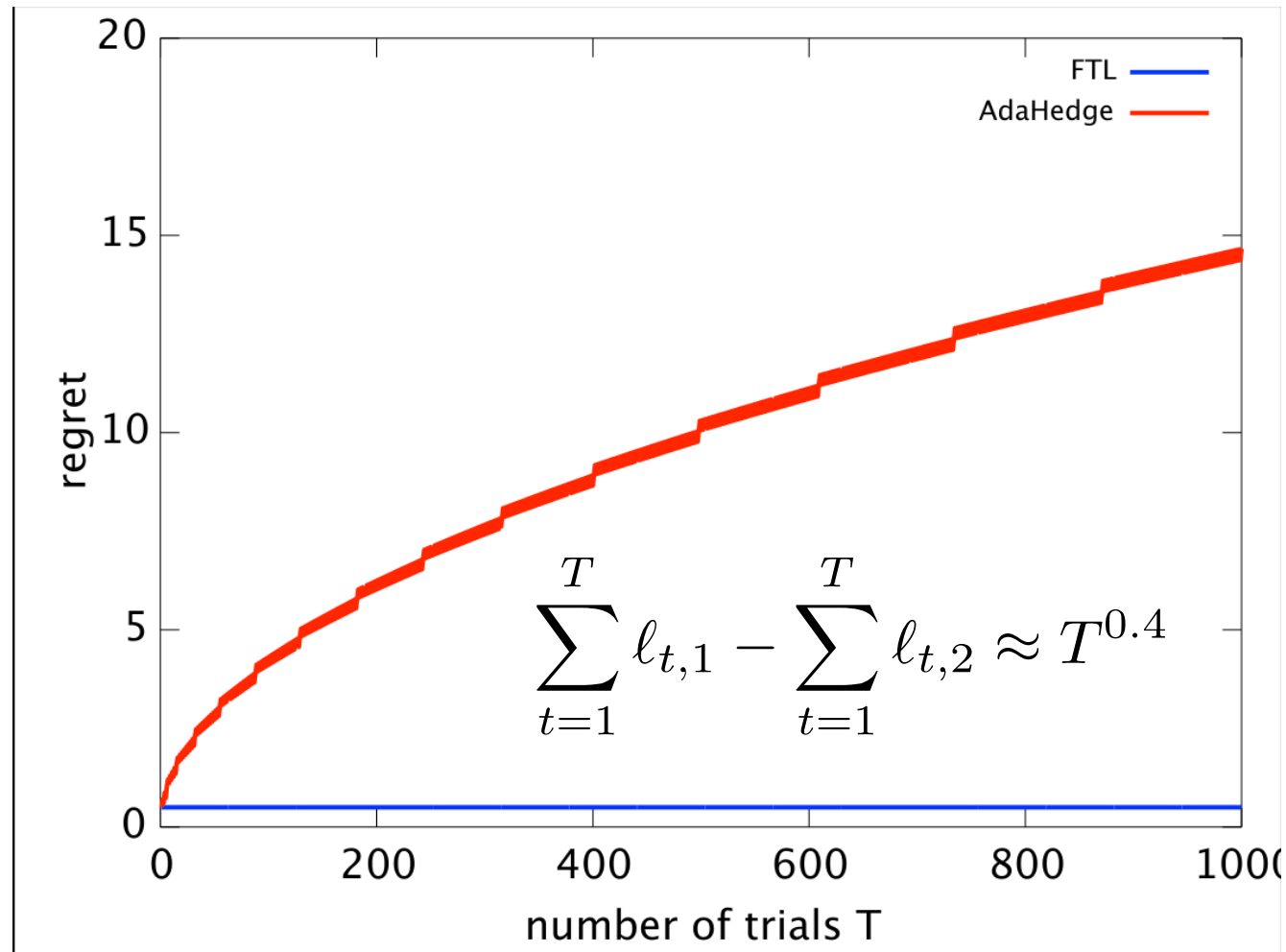
Tuning η Online

- Balancing η in CBMS and AdaHedge depends on **unknown** quantities
- Solve this by changing $\eta = \eta_t$ with t
- Problem: $\sum_t m_t \leq L^* + \ln K / \eta$ breaks

Lemma [KV, 2005]: If $\eta_1 \geq \eta_2 \geq \eta_3 \geq \dots$, then

$$\sum_{t=1}^T m_t \leq L^* + \ln(K) / \eta_T$$

2nd-order Bounds: the Bad News



- Do **not recover** FTL benefits for other 'easy' data with a small number of leader changes

Luckiness: Exploiting Easy Data

- Improvement for small losses:

$$\text{Regret} = O\left(\sqrt{L^* \ln(K)}\right)$$

- Second-order Bounds:

- [CBMS, 2007] and AdaHedge: $O\left(\sqrt{\sum_t v_t \ln(K)}\right)$
- Related bound by [HK, 2008]

- FlipFlop:

- “Follow the leader if you can, Hedge if you must”
- **Regret** \leq **best** of **AdaHedge** and **FTL**

FlipFlop

- FlipFlop bound:

$$\text{Regret} \leq \begin{cases} 6 \cdot \text{FTL Regret} \\ 3 \cdot \text{AdaHedge Regret Bound} \end{cases}$$

- Alternate Flip and Flop regimes
 - Flip: Tune $\eta_t = \infty$ like FTL
 - Flop: Tune η_t like AdaHedge
- (No restarts of the algorithm, like in 'doubling trick'!)

FlipFlop: Proof Ideas

- Alternate Flip and Flop regimes
 - Flip: Tune $\eta_t = \infty$ **like FTL**
 - Flop: Tune η_t **like AdaHedge**
- Analysing two regimes:
 1. Relate mix loss for Flip to mix loss for Flop
 2. Keep approximation errors balanced between regimes

1. Relating Mix Losses

- We violate condition of KV-lemma:

$$\eta_1 \geq \eta_2 \geq \eta_3 \geq \dots$$

- But:

$$\begin{aligned} \sum_t m_t &\leq \sum_t m_t^{\text{flop}} + C \sum_{t \in \text{flop}} \delta_t \\ &\leq L^* + \frac{\ln K}{\eta_T^{\text{flop}}} + C \sum_{t \in \text{flop}} \delta_t \\ &= L^* + (C + 1) \sum_{t \in \text{flop}} \delta_t \end{aligned}$$

2. Balance Approximation Errors

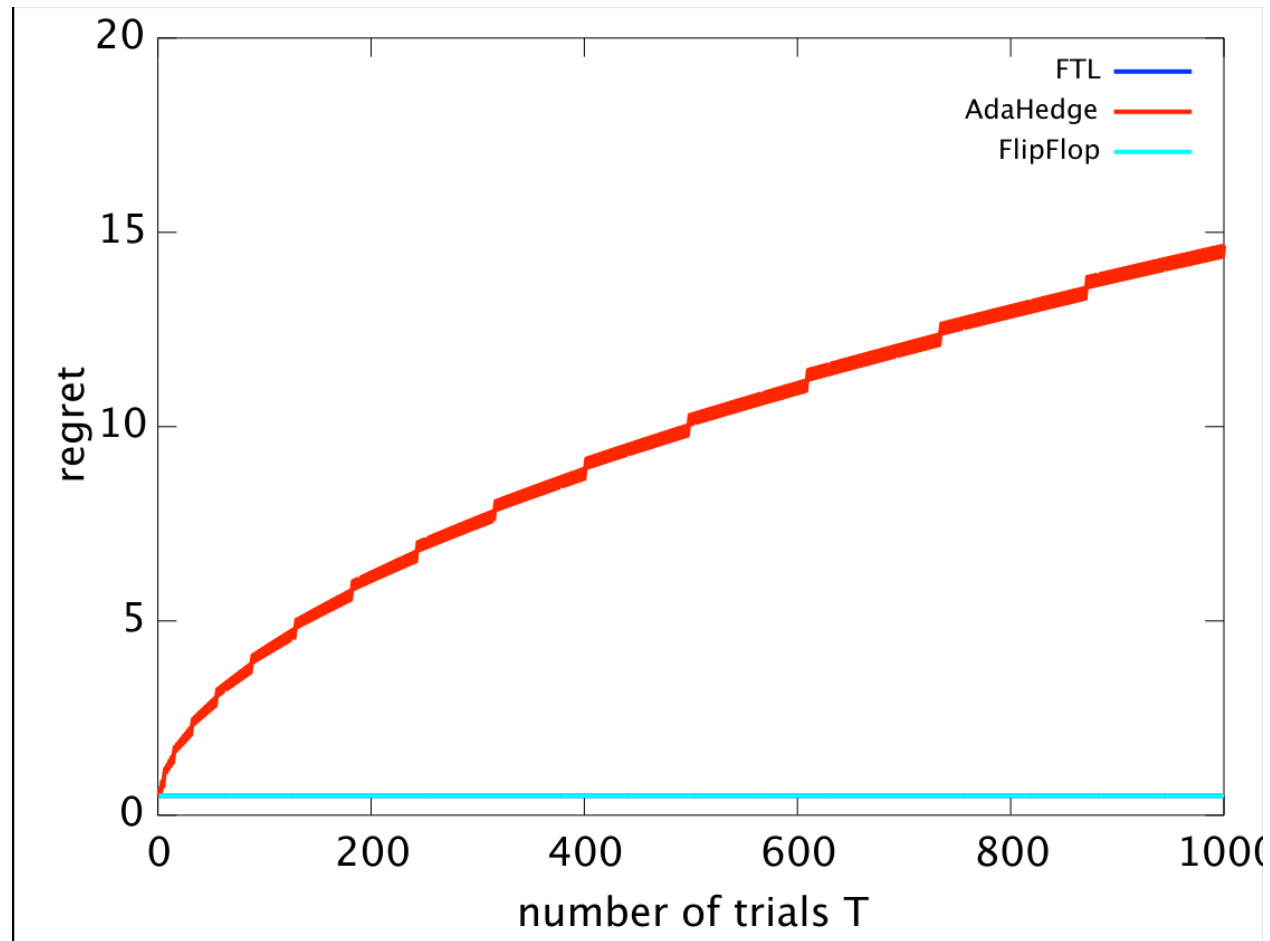
- Alternate regimes to keep approximation errors balanced:

$$\sum_{t \in \text{flip}} \delta_t \propto \sum_{t \in \text{flop}} \delta_t$$

$$\mathbf{Regret} = \sum_t m_t - L^* + \sum_t \delta_t \leq (C + 2) \sum_{t \in \text{flop}} \delta_t + \sum_{t \in \text{flip}} \delta_t$$

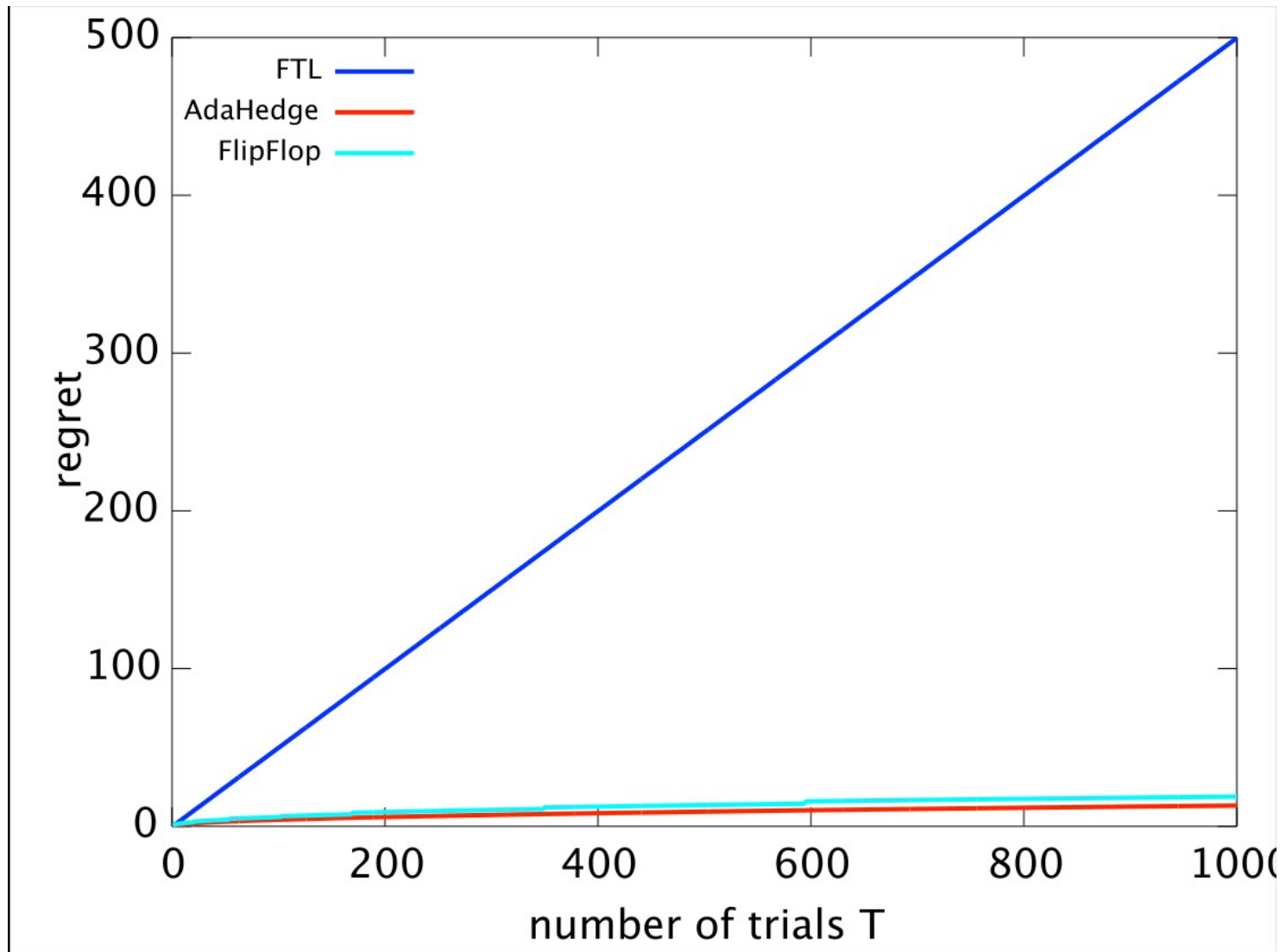
$$\propto \begin{cases} \sum_{t \in \text{flip}} \sum_t \delta_t & \longrightarrow \mathbf{FTL \ Bound} \\ \sum_{t \in \text{flop}} \sum_t \delta_t & \longrightarrow \mathbf{AdaHedge \ Bound} \end{cases}$$

Small Nr Leader Changes Again



- FlipFlop exploits **easy data**, AdaHedge does not

FTL Worst-case Again



Summary

- Follow-the-Leader:
 - works well for **'easy' data**: i.i.d., few leader changes
 - but not robust to worst-case data
- Second-order bounds (e.g. CBMS, AdaHedge):
 - robust against **worst case** + can exploit **i.i.d. data**
 - but do not exploit few leader changes in general
- FlipFlop: **best of both worlds**

Luckiness: What's Missing?

- FlipFlop:
 - “Follow the leader if you can, Hedge if you must”
 - **Regret** \leq **best** of **AdaHedge** and **FTL**
- But what if optimal η is in between AdaHedge and FTL?
- Can we compete with the best possible η chosen in hindsight?

References


- Cesa-Bianchi and Lugosi. *Prediction, learning, and games*. 2006.
- Cesa-Bianchi, Mansour, Stoltz. *Improved second-order bounds for prediction with expert advice*. *Machine Learning*, 66(2/3):321–352, 2007.
- Devaine, Gaillard, Goude, Stoltz. *Forecasting electricity consumption by aggregating specialized experts*. *Machine Learning*, 90(2):231-260, 2013.
- Van Erven, Grünwald, Koolen and De Rooij. *Adaptive Hedge*. NIPS 2011.
- Hazan, Kale. *Extracting certainty from uncertainty: Regret bounded by variation in costs*. COLT 2008.
- De Rooij, Van Erven, Grünwald, Koolen. **Follow the Leader If You Can, Hedge If You Must**. Accepted by the *Journal of Machine Learning Research*, 2013.

EXTRA SLIDES

No Need to Pre-process Losses

- Common assumption $\ell_{t,k} \in [0, 1]$ requires **translating** and **rescaling** the losses
- CBMS:
 - Extension so this is **not necessary**.
Important when range of losses is unknown!
- AdaHedge and FlipFlop:
 - Invariant under rescaling and translation of losses, so get this **for free**.

2nd-order Bounds: I.I.D. Data

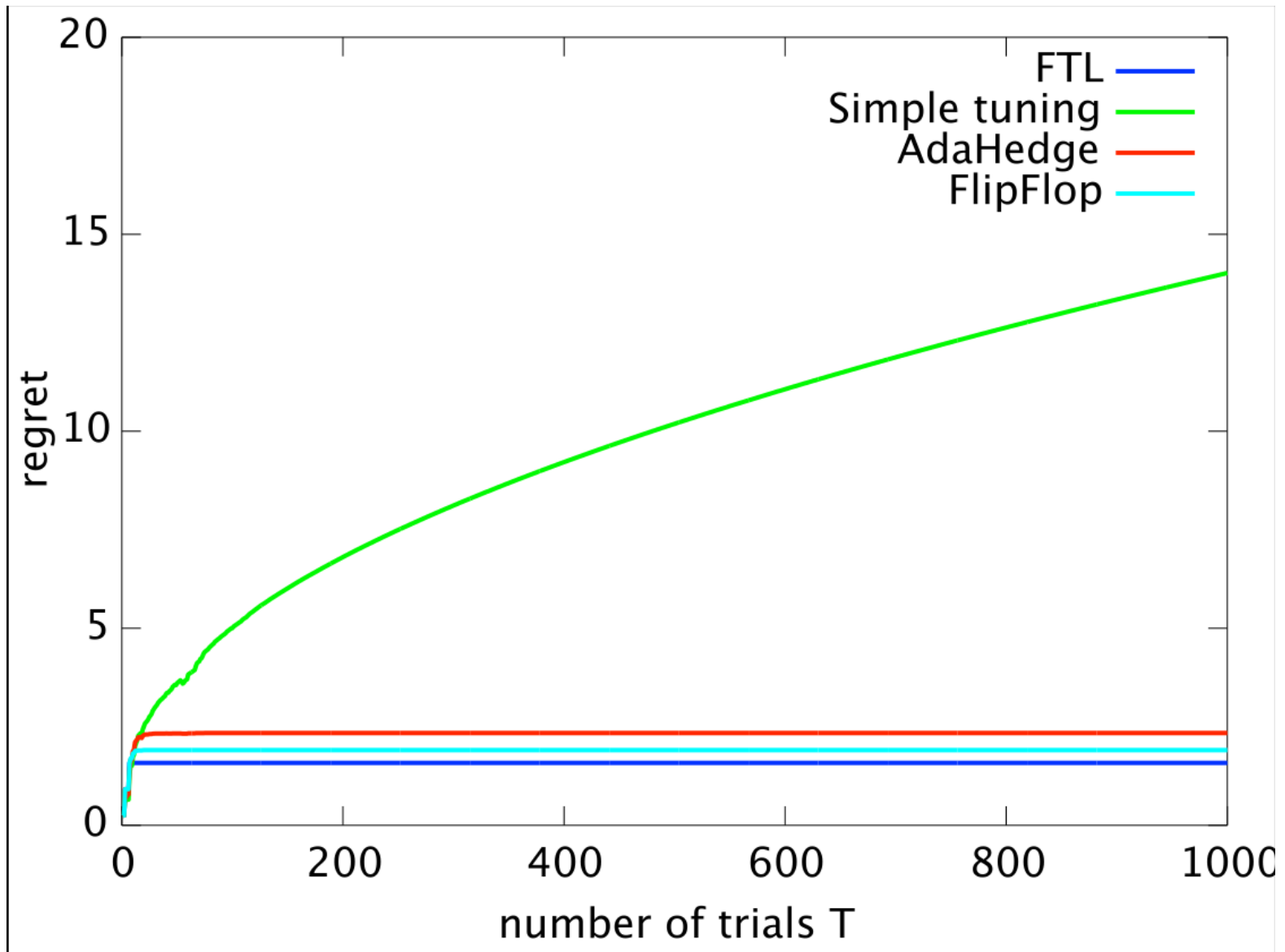
- Regret bound: $O\left(\sqrt{\sum_t v_t \ln(K)}\right)$  variance of w_t
- If w_t **concentrates fast** on best expert, then

$$\sum_t v_t \leq C \longrightarrow \text{Regret} \leq C'$$

- IID data:

1. Balancing $\eta_t = \sqrt{\frac{2 \ln(K)}{\sum_s^{t-1} v_s}}$ is large for all $t \leq T$
2. w_t concentrates fast
3. Then 1. also holds for $T + 1$

FlipFlop on I.I.D. Data



Example: Spam Detection

Subject	From	
☑ <i>Gratis Turkije. . .</i>	<i>Reizen Center</i>	$y_1 = 1$
☑ <i>uitnodiging hoorzitting reorganisatie FEW dinsdag 20 se...</i>	<i>Ivo van Stokkum</i>	$y_2 = 0$
☑ <i>Re: Urgent Business Inquiry.</i>	<i>Ubc Ltd</i>	$y_3 = 1$
☑ <i>Reminder: first colloquium</i>	<i>Jeu, R.M.H. de</i>	$y_4 = 0$
📧 <i>Informatie over VUnet</i>	<i>College van Bestuur</i>	$y_5 = 0$
☑ <i>USD 500 Free Deposit at PartyPoker!</i>	<i>PartyPoker</i>	$y_6 = 1$
📧 <i>YOU ARE A WINNER!!! VERY URGENT NOTIFICATION.</i>	<i>UK INTL. LOTTERY PROMOTION</i>	$y_7 = 1$
📧 <i>bachelor/master diploma uitreiking 14 september</i>	<i>Sotiriou, M.</i>	$y_8 = 0$
☑ <i>HAPPY NEW YEAR 2068</i>	<i>Anil Shilpakar</i>	$y_9 = 1$
📧 <i>Thailand Package</i>	<i>Anil Shilpakar</i>	$y_{10} = 1$

Example: Spam Detection

- Data: (x_t, y_t) with $y_t \in \{0, 1\}$
- Predictions: probability $p_t \in [0, 1]$ that $y_t = 1$
- Loss (probability of wrong label):

$$\ell(y_t, p_t) = \begin{cases} p_t & \text{if } y_t = 0 \\ 1 - p_t & \text{if } y_t = 1 \end{cases}$$

- Experts: K spam detection algorithms
- If expert k predicts $p_{t,k}$, then $\ell_{t,k} = \ell(y_t, p_{t,k})$
- Regret: expected nr. mistakes over expected nr. of mistakes of best algorithm

FTL: the Bad News

- Consider two trivial spam detectors (experts):

$$p_{t,1} = 0 \quad p_{t,2} = 1$$

- If we deterministically choose an expert k^* (like FTL) then we could be wrong all the time:

$$\ell_{t,k^*} = 1 \quad \ell_{t,\neg k^*} = 0$$

Regret:

- Let n denote the number of times expert 1 has loss 1. Then $L^* = \min\{n, T - n\} \leq T/2$
- **Linear** regret = $T - L^* \geq T/2$

