JMLR: Workshop and Conference Proceedings vol 49:1–4, 2016

# Open Problem: Second order regret bounds parametrized by variance across actions and top $\epsilon$ percentile.

**Yoav Freund** YFREUND@UCSD.EDU
*UCSD, San Diego, CA*

## Abstract

We argue that of the two second order bounds given in Cesa-Bianchi et al. (2006) the second type, which is based on variance across actions is superior to the first type, which is based on variance across time. We ask whether there exists an online algorithm with a stronger regret bound of the second type that hold with respect to the $\epsilon$-best actions, simultanously for all values of $\epsilon$. We conjecture that such a regret bound is achieved by NormalHedge Chaudhuri et al. (2009).

## 1. Motivation

The upper bound on the regret of exponential weights algorithms Littlestone and Warmuth (1989); Cesa-Bianchi et al. (1993); Freund and Schapire (1999) have as the leading term $O(\sqrt{n \log N})$ where $n$ is the length of the sequence and $N$ is the number of actions[1] The regret bounds assume that the loss (gain) per iteration is in a bounded range, typically $[0, 1]$. Obviously, if the range is restricted (a priori) to $[0, 1/2]$ then the bound will also be halved.

Consider a scenario in which the range is $[0, 1]$ but the actual observed losses are in the range $[0, 1/2]$, we would like to have an algorithm which will perform (almost) as well as an algorithm that knew a-priori that the range is $[0, 1/2]$. The a-priori knowledge is consequential because multiplicative weights algorithms use it to choose the learning rate.

For a more general formulation of the problem suppose that the losses in iteration $t$ are in the range $[0, a_t]$. Then we are seeking an algorithm with a bound whose leading term is

$$O\left(\sqrt{\sum_{t=1}^{n} a_t^2 \log N}\right) \tag{1}$$

## 2. Second order bounds

In Cesa-Bianchi et al. (2006) the authors prove two regret bounds for the multiplicative weights algorithm. Both bounds satisfy condition (1).

The bounds are based on two different quantifications of variance described below. In this description we follow the notation of Cesa-Bianchi et al. (2006). $N$ is the number of actions, $n$ is total number of iterations, $t$ indexes the iterations and $i$ indexes the actions; $x_{i,t}$ is the instantanous payoff of action $i$ at time step $t$.

[1] In this note we use the term “action”, which corresponds to decision theoretic online learning (DTOL) rather than the term “expert”, which corresponds to a loss-based game in which expert predictions are revealed to the algorithm before it makes it’s prediction.

1. **Variance across time:** Theorem 1 ofof Cesa-Bianchi et al. (2006) states a regret bound in which the leading term is $O(\sqrt{Q_n^* \log N})$ where the cumulative variance is

$$Q_n^* = \sum_{t=1}^{n} x_{k_n^*,t}^2$$

where $k_n^*$ is the action with the highest cumulative payoff at time $n$.

2. **Variance across actions:** Section 4 ofof Cesa-Bianchi et al. (2006) introduces another notion of cumulative variance. They define $Z_t$ to be a random variable which takes on the $i$th payoff $x_{i,t}$ with the probability $p_{i,t}$, which is the weight associated with action $i$ at time $t$ by the algorithm.

   The cumulative variance is defined as $V_n = Var(Z_1) + Var(Z_2) + \cdots + Var(Z_n)$. Theorem 5 and 6 prove regret bounds where the leading term is $O(\sqrt{V_n \log N})$

We will use the acronymns VAT and VAA to represent "variance across time" and "variance across actions" respectively.

While VAA term $V_n$ and the VAT term $Q_n^*$ play the same role in the regret bounds and both satisfy condition 1, the resulting bounds are not equal. In the following section we argue that VAA yields much tighter regret bounds than VAT.

## 3. VAA is stronger than VAT

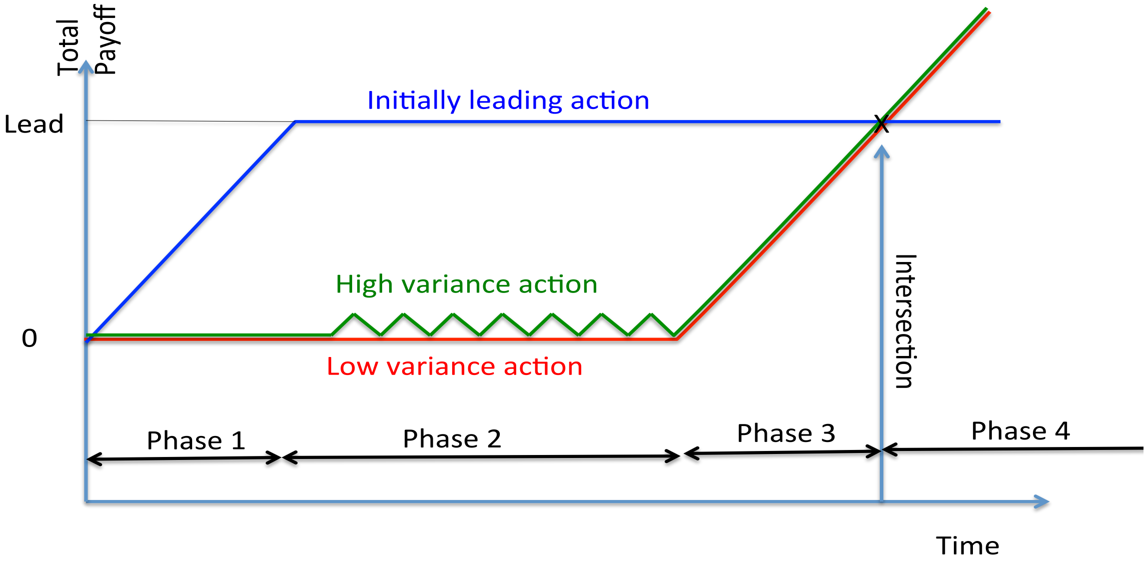
We start with a rough intutitive general argument. We then provide a simple example demonstrating the argument.

In Cesa-Bianchi et al. (2006) the authors argue that VAT is more natural than VAA, becuase VAA depends on the prediction algorithm while VAT does not. We argue that such a dependence is necessary, regret depends on the number of times that the cumulative payoffs of different actions intersect. For example, follow-the-leader algorithm has no regret if the same action is the leader at all time steps. VAT depends by definition, only on the payoff sequence of the action that is currently the best and is therefor insensitive to the number of cross-overs that occured.

To demonstrate the problem consider following example (See Figure 3 below). Consider a game with two actions and four phases. One action is the 'initially leading action' (ILA) and the other action is either "low variance action" (LVA) or "high variance action" (HVA). I will use the term "gain" to refer to the cumulative payoff.

At the end of the first phase the gain of ILA is large (we call it "the lead") while the gain of HVA or LVA is zero. In phase 2 the lead remains constant. Any reasonable hedging algorithm will assign most of the weight to ILA, and all but ignore the trailing action. In particular, if the lead is sufficiently large, NormalHedge will assign *zero weight* to the second action throughout phase 2. In any case, it does not matter to the algorithm (and should not matter to the bounds) whether the second action is LVA or HVA. In phases 3 and 4 the fates of the two actions reverses. The gain of ILA remains constant, but the gain of the other action increases by one at each step. The length of phase 3 is equal to that of phase 1. When phase 3 ends, the two gain sequences intersect. In phase 4 the gain of the second action keeps increasing. A bound based on VAT will measure the regret with respect to the second action and the bound will strongly depend on whether the second action is LVA or HVA. At the same time, the behaviour of the online algorithm depends only weakly on whether the second action is LVA or HVA.

In Cesa-Bianchi et al. (2006) only the bound is effected by $Q_n^*$, the algorithm does not depend on it. In other work, such as Squint Koolen and van Erven (2015) the value of the VAT effects the weight assigned to the actions and the argument stated above adversely effects the actual perforance of the algorithm.



## 4. Getting rid of both $N$ and $n$

Bounds with a leading term $\emptyset(\sqrt{n \ln N})$ have been improved to bound where the leading term is $O(\sqrt{n \ln 1/\epsilon})$, where $\epsilon$ corresponds to the top $\epsilon$-fraction of the actions Chaudhuri et al. (2009); Chernov and Vovk (2010); Luo and Schapire (2015); Koolen and van Erven (2015). This bound is stronger then the old bound because the value if $\epsilon$ can be chosen in hindsight.

The VAA bounds of Cesa-Bianchi et al. (2006) replace the number of steps, $n$, by the cumulative variance $V_n$.

My conjecture is that there is a hedging algorithm whose regret bound has a leading term of the form $O(\sqrt{V_n \log 1/\epsilon})$ (possibly with lower order logarithmic terms).

**I am offering a prize of $500 to anyone who gives an algorithm with a proven bound of this form, or, alternatively, a proof that such a bound is impossible.**

My conjecture is that NormalHedge has a regret bound of this form.

Beyond solving a riddle, an algorithm with this type of bound would define a new parameterization of online learning which replaces the external parameters $n$ and $N$ with parameters $V_n$ and $\epsilon$ which reflect the internal structure of the payoff sequences. One upshot of using these parameters is that the analysis extends to continuous sets of actions and to continuous time.

## 5. Thanks and apologies

I would like to thank Gergely Neu for pointing out that both types of variance are analyzed in Cesa-Bianchi et al. (2006).

I wish to also apologize in advance for any omittions, errors or misrepresentation I made in this note. I welcome any comments.

## References

Nicolò Cesa-Bianchi, Yoav Freund, David P. Helmbold, David Haussler, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. In *Proceedings of the Twenty-Fifth Annual ACM Symposium on the Theory of Computing*, pages 382–391, 1993.

Nicolò Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2):321–352, 2006. ISSN 1573-0565. doi: 10.1007/s10994-006-5001-7. URL http://dx.doi.org/10.1007/s10994-006-5001-7.

Kamalika Chaudhuri, Yoav Freund, and Daniel J. Hsu. A parameter-free hedging algorithm. *CoRR*, abs/0903.2851, 2009. URL http://arxiv.org/abs/0903.2851.

Alexey V. Chernov and Vladimir Vovk. Prediction with advice of unknown number of experts. *CoRR*, abs/1006.0475, 2010. URL http://arxiv.org/abs/1006.0475.

Yoav Freund and Robert E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29:79–103, 1999.

Wouter M. Koolen and Tim van Erven. Second-order quantile methods for experts and combinatorial games. *CoRR*, abs/1502.08009, 2015. URL http://arxiv.org/abs/1502.08009.

Nick Littlestone and Manfred Warmuth. The weighted majority algorithm. In *30th Annual Symposium on Foundations of Computer Science*, pages 256–261, October 1989.

Haipeng Luo and Robert E. Schapire. Achieving all with no parameters: Adaptive normalhedge. *CoRR*, abs/1502.05934, 2015. URL http://arxiv.org/abs/1502.05934.