On the optimality of anytime Hedge in the stochastic regime

Jaouad Mourtada, Stéphane Gaïffas

CMAP, École polytechnique

CMStatistics 2018
Pisa, 15/12/18

Reference: "On the optimality of the Hedge algorithm in the stochastic regime", J. Mourtada & S. Gaïffas, arXiv preprint arXiv:1809.01382.

Hedge setting

Experts $i = 1, \ldots, M$; can be thought of as sources of predictions.
Aim is to predict almost as well as the best expert in hindsight.

Hedge problem (= online linear optimization on the simplex)

At each time step $t = 1, 2, \ldots$

1. *Forecaster chooses probability distribution* $\boldsymbol{v}_t = (v_{i,t})_{1 \leqslant i \leqslant M} \in \Delta_M$ *on the experts;*
2. *Environment chooses loss vector* $\boldsymbol{\ell}_t = (\ell_{i,t})_{1 \leqslant i \leqslant M} \in [0,1]^M$*;*
3. *Forecaster incurs loss* $\ell_t := \langle \boldsymbol{v}_t, \boldsymbol{\ell}_t \rangle = \sum_{i=1}^{M} v_{i,t} \ell_{i,t}$.

Goal: Control, for **every** loss vectors $\boldsymbol{\ell}_t \in [0,1]^M$, the **regret**

$$R_T = \sum_{t=1}^{T} \ell_t - \min_{1 \leqslant i \leqslant M} \sum_{t=1}^{T} \ell_{i,t} \,.$$

Hedge algorithm and regret bound

First observation: Follow the Leader (FTL) / ERM, $v_{i_t,t} = 1$ where $i_t \in \mathrm{argmin}_i \sum_{s=1}^{t-1} \ell_{i,s} \Rightarrow$ **no sublinear regret** !

Indeed, let

$$(\ell_{1,1}, \ell_{2,1}), (\ell_{1,2}, \ell_{2,2}), (\ell_{1,3}, \ell_{2,3}), \dots = (1/2, 0), (0, 1), (1, 0), \dots$$

Then, $\sum_{t=1}^{T} \langle \boldsymbol{v}_t, \boldsymbol{\ell}_t \rangle = T - \frac{1}{2}$, but $\sum_{t=1}^{T} \ell_{2,t} \leqslant \frac{T-1}{2}$, hence $R_T \geqslant \frac{T-1}{2} \neq o(T)$.

Hedge algorithm and regret bound

First observation: Follow the Leader (FTL) / ERM, $v_{i_t,t} = 1$ where $i_t \in \mathrm{argmin}_i \sum_{s=1}^{t-1} \ell_{i,s} \Rightarrow$ **no sublinear regret** !

Hedge algorithm (Constant learning rate)

$$v_{i,t} = \frac{e^{-\eta L_{i,t-1}}}{\sum_{j=1}^{M} e^{-\eta L_{j,t-1}}}$$

where $L_{i,t} = \sum_{s=1}^{t} \ell_{i,s}$, η *learning rate.*

Regret bound [Freund & Schapire 1997; Vovk, 1998]:

$$R_T \leqslant \frac{\log M}{\eta} + \frac{\eta T}{8} \leqslant \sqrt{(T/2)\log M}$$

for $\eta = \sqrt{8(\log M)/T}$ tuned knowing **fixed time horizon** T.

$O(\sqrt{T \log M})$ regret bound is **minimax (worst-case) optimal**.

Hedge algorithm and regret bound

Hedge algorithm (Time-varying learning rate)

$$v_{i,t} = \frac{e^{-\eta_t L_{i,t-1}}}{\sum_{j=1}^{M} e^{-\eta_t L_{j,t-1}}}$$

where $L_{i,t} = \sum_{s=1}^{t} \ell_{i,s}$, η_t *learning rate.*

Regret bound: if η_t decreases,

$$R_T \leqslant \frac{\log M}{\eta_T} + \frac{1}{8}\sum_{t=1}^{T} \eta_t \leqslant \sqrt{T \log M}$$

for $\eta_t = \sqrt{2(\log M)/t}$, valid **for every horizon T (anytime)**.

$O(\sqrt{T \log M})$ regret bound is **minimax (worst-case) optimal**.

Beyond worst case: adaptivity to easy stochastic instances

- Hedge with $\eta \asymp \sqrt{(\log M)/T}$ (constant) or $\eta_t \asymp \sqrt{(\log M)/t}$ (anytime) achieve **optimal worst case** $O(\sqrt{T \log M})$ **regret**.

[1] E.g., van Erven et al., 2011; Gaillard et al., 2014; Luo & Schapire, 2015.

Beyond worst case: adaptivity to easy stochastic instances

- Hedge with $\eta \asymp \sqrt{(\log M)/T}$ (constant) or $\eta_t \asymp \sqrt{(\log M)/t}$ (anytime) achieve **optimal worst case** $O(\sqrt{T \log M})$ **regret**.
- However, worst-case is **pessimistic** and can lead to **overly conservative** algorithms.
- **"Easy" problem instance**: stochastic case. If the loss vectors $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \dots$ are i.i.d. (e.g., $\ell_{i,t} = \ell(f_i(X_t), Y_t)$), FTL/ERM achieves **constant** $O(\log M)$ regret $\Rightarrow$ **fast rate**.

[1] E.g., van Erven et al., 2011; Gaillard et al., 2014; Luo & Schapire, 2015.

Beyond worst case: adaptivity to easy stochastic instances

- Hedge with $\eta \asymp \sqrt{(\log M)/T}$ (constant) or $\eta_t \asymp \sqrt{(\log M)/t}$ (anytime) achieve **optimal worst case** $O(\sqrt{T \log M})$ **regret**.
- However, worst-case is **pessimistic** and can lead to **overly conservative** algorithms.
- **"Easy" problem instance**: stochastic case. If the loss vectors $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \dots$ are i.i.d. (e.g., $\ell_{i,t} = \ell(f_i(X_t), Y_t)$), FTL/ERM achieves **constant** $O(\log M)$ regret $\Rightarrow$ **fast rate**.
- Recent line of work[1]: algorithms that combine **worst-case** $O(\sqrt{T \log M})$ regret with **faster rate** on "easier" instances.
- Example: **AdaHedge** algorithm [van Erven et al., 2011,2015]. **Data-dependent learning rate** η_t.
  - Worst-case: "safe" $\eta_t \asymp \sqrt{(\log M)/t}$, $O(\sqrt{T \log M})$ regret;
  - Stochastic case: $\eta_t \asymp cst$ ($\approx$ FTL), $O(\log M)$ regret.

[1] E.g., van Erven et al., 2011; Gaillard et al., 2014; Luo & Schapire, 2015.

Optimality of anytime Hedge in the stochastic regime

Our result: anytime Hedge with "conservative" $\eta_t \asymp \sqrt{(\log M)/t}$ is actually optimal in the easy stochastic regime!

- **Stochastic instance**: i.i.d. loss vectors $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \dots$ such that $\mathbb{E}[\ell_{i,t} - \ell_{i^*,t}] \geqslant \Delta$ for $i \neq i^*$ (where $i^* = \mathrm{argmin}_i \mathbb{E}[\ell_{i,t}]$).

Proposition (M., Gaïffas, 2018)

On any stochastic instance with sub-optimality gap Δ, anytime Hedge with $\eta_t \asymp \sqrt{(\log M)/t}$ achieves, for every $T \geqslant 1$:

$$\mathbb{E}[R_T] \lesssim \frac{\log M}{\Delta} \, .$$

Remark: $\frac{\log M}{\Delta}$ regret is optimal under the gap assumption.

Anytime Hedge vs. Fixed horizon Hedge

Theorem (M., Gaïffas, 2018)

*On any stochastic instance with sub-optimality gap Δ, **anytime Hedge** with $\eta_t \asymp \sqrt{(\log M)/t}$ achieves, for every $T \geqslant 1$:*

$$\mathbb{E}[R_T] \lesssim \frac{\log M}{\Delta} \,.$$

Proposition (M., Gaïffas, 2018)

*If $\ell_{i^*,t} = 0$, $\ell_{i,t} = 1$ for $i \neq i^*$, $t \geqslant 1$, a stochastic instance with gap $\Delta = 1$, **constant Hedge** with $\eta_t \asymp \sqrt{(\log M)/T}$ achieves*

$$R_T \asymp \sqrt{T \log M} \,.$$

- Seemingly similar Hedge variants behave very differently on stochastic instances!
- Even if horizon T is known, anytime variant is preferable.

Some proof ideas

- Divide time **two phases** $[1, \tau]$ (dominated by noise) and $[\tau, T]$ (weights concentrate fast to i^*), with $\tau \asymp \frac{\log M}{\Delta^2}$.
- Early phase: worst-case regret $R_\tau \lesssim \sqrt{\tau \log M} \lesssim \frac{\log M}{\Delta}$.
- At the beginning of late phase, *i.e.* $t \approx \tau \approx \frac{\log M}{\Delta^2}$, two things occur simultaneously:
  1. i^* linearly dominates the other experts: for every $i \neq i^*$, $L_{i,t} - L_{i^*,t} \gtrsim \frac{1}{2}\Delta t$. Hoeffding: it suffices that $Me^{-t\Delta^2} \lesssim 1$.
  2. Expert i^* receives at least $1/2$ of the weights: under previous condition, it suffices that $Me^{-\Delta\sqrt{t \log M}} \lesssim 1$.
- Condition (2) eliminates potentially linear dependence on M in the bound. To control regret in the second phase, we then use (1) and the fact that for $c > 0$, $\sum_{t \geqslant 0} e^{-c\sqrt{t}} \lesssim \frac{1}{c^2}$.

The advantage of adaptive algorithms

- **Stochastic regime** with **gap Δ** often considered in the literature to show the improvement of adaptive algorithms.
- However, anytime Hedge achieves optimal $O(\frac{\log M}{\Delta})$ regret in this case. **No need to tune η_t ?**

[2] Mammen & Tsybakov, 1999; Bartlett & Mendelson, 2006.

The advantage of adaptive algorithms

- **Stochastic regime** with gap Δ often considered in the literature to show the improvement of adaptive algorithms.
- However, anytime Hedge achieves optimal $O(\frac{\log M}{\Delta})$ regret in this case. **No need to tune** η_t **?**
- (β, B)-Bernstein condition[2] ($\beta \in [0,1], B > 0$): for $i \neq i^*$,

$$\mathbb{E}[(\ell_{i,t} - \ell_{i^*,t})^2] \leqslant B\mathbb{E}[\ell_{i,t} - \ell_{i^*,t}]^{\beta} .$$

Proposition (Koolen, Grünwald & van Erven, 2016)

Algorithms with so-called "second-order regret bounds" (including AdaHedge) achieve on (β, B)-Bernstein stochastic losses:

$$\mathbb{E}[R_T] \lesssim (B \log M)^{\frac{1}{2-\beta}} T^{\frac{1-\beta}{2-\beta}} + \log M .$$

For $\beta = 1$, gives $O(B \log M)$ regret; we can have $B \ll \frac{1}{\Delta}$!

[2] Mammen & Tsybakov, 1999; Bartlett & Mendelson, 2006.

The advantage of adaptive algorithms

- $(1, B)$-Bernstein condition: $\mathbb{E}[(\ell_{i,t} - \ell_{i^*,t})^2] \leqslant B\mathbb{E}[\ell_{i,t} - \ell_{i^*,t}]$.
- In this case, adaptive algorithms achieve $O(B \log M)$ regret.
- We have $B \leqslant \frac{1}{\Delta}$, but potentially $B \ll \frac{1}{\Delta}$ (e.g., low noise).

Proposition

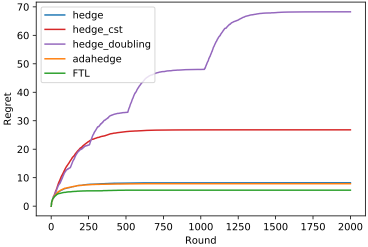
There exists a $(1,1)$-Bernstein stochastic instance on which anytime Hedge satisfies

$$\mathbb{E}[R_T] \gtrsim \sqrt{T \log M}\,.$$

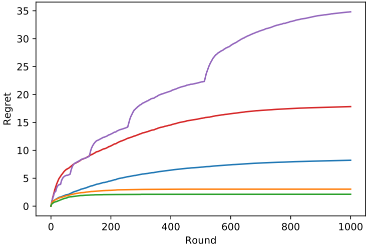
In fact, gap Δ (essentially) characterizes anytime Hedge's regret on ***any*** *stochastic instance: for $T \gtrsim 1/\Delta^2$,*

$$\mathbb{E}[R_T] \gtrsim \frac{1}{(\log M)^2 \Delta}\,.$$

Experiments



(a)



(b)

Figure: Cumulative regret of Hedge algorithms on two stochastic instances. (a) Stochastic instance with a gap, independent losses across experts ($M = 20, \Delta = 0.1$); (b) Bernstein instance with small Δ, but small B ($M = 10, \Delta = 0.04, B = 4$).

Conclusion and perspectives

- Despite **conservative** learning rate (*i.e.*, large penalization), anytime Hedge achieves $O(\frac{\log M}{\Delta})$ regret, **adaptively in the gap Δ**, in the easy stochastic case.
- **Not the case** with fixed-horizon $\eta_t \asymp \sqrt{(\log M)/T}$ instead of $\eta_t \asymp \sqrt{(\log M)/t}$.
- Tuning the learning rate does help in some situations.
- Result of a similar flavor in **stochastic optimization**[3]: SGD with step size $\eta_t \asymp \frac{1}{\sqrt{t}}$ achieves $O(\frac{1}{\mu T})$ excess risk after averaging on μ-strongly convex problems (**adaptively** in μ). Not directly related, in fact **"opposite"** phenomenon.

[3] Moulines & Bach, 2011.

Thank you!