Adaptivity and Optimism: An Improved Exponentiated Gradient Algorithm

Jacob Steinhardt   Percy Liang

Stanford University

{jsteinhardt,pliang}@cs.stanford.edu

Jun 11, 2013

Setup

Setting is **learning from experts**:

- n experts, T rounds
- For $t = 1, \dots, T$:
  - Learner chooses distribution $w_t \in \Delta_n$ over the experts
  - Nature reveals losses $z_t \in [-1, 1]^n$ of the experts
  - Learner suffers loss $w_t^\top z_t$
- Goal: minimize

$$\text{Regret} \stackrel{\text{def}}{=} \sum_{t=1}^{T} w_t^\top z_t - \sum_{t=1}^{T} z_{t,i^*},$$

  where i^* is the best fixed expert.
- Typical algorithm: multiplicative weights (aka exponentiated gradient):

$$w_{t+1,i} \propto w_{t,i} \exp(-\eta z_{t,i}).$$

Outline

- Compare two variants of the multiplicative weights (exponentiated gradient) algorithm
- Understand the difference through lens of **adaptive mirror descent** (Orabona et al., 2013)
- Combine with machinery of **optimistic updates** (Rakhlin & Sridharan, 2012) to beat best existing bounds.

Two Types of Updates

In literature, two similar but different updates (Kivinen & Warmuth, 1997; Cesa-Bianchi et al., 2007):

$$w_{t+1,i} \propto w_{t,i} \exp(-\eta z_{t,i}) \tag{MW1}$$

$$w_{t+1,i} \propto w_{t,i}(1 - \eta z_{t,i}) \tag{MW2}$$

The regret is bounded as

$$\text{Regret} \leq \frac{\log(n)}{\eta} + \eta \sum_{t=1}^{T} \|z_t\|_\infty^2 \tag{Regret:MW1}$$

$$\text{Regret} \leq \frac{\log(n)}{\eta} + \eta \sum_{t=1}^{T} z_{t,i^*}^2 \tag{Regret:MW2}$$

If best expert i^* has loss close to zero, then second bound better than first.
Gap can be $\Theta(\sqrt{T})$ (in actual performance, not just upper bounds).

Two Types of Updates

In literature, two similar but different updates (Kivinen & Warmuth, 1997; Cesa-Bianchi et al., 2007):

$$w_{t+1,i} \propto w_{t,i} \exp(-\eta z_{t,i}) \quad \text{(MW1)}$$

$$w_{t+1,i} \propto w_{t,i}(1 - \eta z_{t,i}) \quad \text{(MW2)}$$

Mirror descent is the gold standard meta-algorithm for online learning. How do (MW1, MW2) relate to it?

Two Types of Updates

In literature, two similar but different updates (Kivinen & Warmuth, 1997; Cesa-Bianchi et al., 2007):

$$w_{t+1,i} \propto w_{t,i} \exp(-\eta z_{t,i}) \quad \text{(MW1)}$$

$$w_{t+1,i} \propto w_{t,i}(1 - \eta z_{t,i}) \quad \text{(MW2)}$$

Mirror descent is the gold standard meta-algorithm for online learning. How do (MW1, MW2) relate to it?

- (MW1) is mirror descent with regularizer $\frac{1}{\eta} \sum_{i=1}^{n} w_i \log(w_i)$

Two Types of Updates

In literature, two similar but different updates (Kivinen & Warmuth, 1997; Cesa-Bianchi et al., 2007):

$$w_{t+1,i} \propto w_{t,i} \exp(-\eta z_{t,i}) \tag{MW1}$$

$$w_{t+1,i} \propto w_{t,i}(1 - \eta z_{t,i}) \tag{MW2}$$

Mirror descent is the gold standard meta-algorithm for online learning. How do (MW1, MW2) relate to it?

- (MW1) is mirror descent with regularizer $\frac{1}{\eta} \sum_{i=1}^{n} w_i \log(w_i)$
- (MW2) is NOT mirror descent for any fixed regularizer

Two Types of Updates

In literature, two similar but different updates (Kivinen & Warmuth, 1997; Cesa-Bianchi et al., 2007):

$$w_{t+1,i} \propto w_{t,i} \exp(-\eta z_{t,i}) \quad \text{(MW1)}$$

$$w_{t+1,i} \propto w_{t,i}(1 - \eta z_{t,i}) \quad \text{(MW2)}$$

Mirror descent is the gold standard meta-algorithm for online learning. How do (MW1, MW2) relate to it?

- (MW1) is mirror descent with regularizer $\frac{1}{\eta} \sum_{i=1}^{n} w_i \log(w_i)$
- (MW2) is NOT mirror descent for any fixed regularizer
- Unsettling: should we abandon mirror descent as a gold standard?

Two Types of Updates

In literature, two similar but different updates (Kivinen & Warmuth, 1997; Cesa-Bianchi et al., 2007):

$$w_{t+1,i} \propto w_{t,i} \exp(-\eta z_{t,i}) \quad \text{(MW1)}$$

$$w_{t+1,i} \propto w_{t,i}(1 - \eta z_{t,i}) \quad \text{(MW2)}$$

Mirror descent is the gold standard meta-algorithm for online learning. How do (MW1, MW2) relate to it?

- (MW1) is mirror descent with regularizer $\frac{1}{\eta} \sum_{i=1}^{n} w_i \log(w_i)$
- (MW2) is NOT mirror descent for any fixed regularizer
- Unsettling: should we abandon mirror descent as a gold standard?
  - No: can cast (MW2) as ***adaptive*** mirror descent (Orabona et al., 2013)

Adaptive Mirror Descent to the Rescue

- Recall that mirror descent is the (meta-)algorithm

$$w_t = \operatorname{argmin}_w \psi(w) + \sum_{s=1}^{t-1} w^\top z_s.$$

For $\psi(w) = \frac{1}{\eta}\sum_{i=1}^n w_i \log(w_i)$, we recover (MW1).

Adaptive Mirror Descent to the Rescue

- Recall that mirror descent is the (meta-)algorithm

$$w_t = \operatorname{argmin}_w \psi(w) + \sum_{s=1}^{t-1} w^\top z_s.$$

  For $\psi(w) = \frac{1}{\eta} \sum_{i=1}^{n} w_i \log(w_i)$, we recover (MW1).
- *Adaptive* mirror descent (Orabona et al., 2013) is the meta-algorithm

$$w_t = \operatorname{argmin}_w \psi_t(w) + \sum_{s=1}^{t-1} w^\top z_s.$$

Adaptive Mirror Descent to the Rescue

- Recall that mirror descent is the (meta-)algorithm

$$w_t = \text{argmin}_w \, \psi(w) + \sum_{s=1}^{t-1} w^\top z_s.$$

For $\psi(w) = \frac{1}{\eta} \sum_{i=1}^{n} w_i \log(w_i)$, we recover (MW1).

- *Adaptive* mirror descent (Orabona et al., 2013) is the meta-algorithm

$$w_t = \text{argmin}_w \, \psi_t(w) + \sum_{s=1}^{t-1} w^\top z_s.$$

Adaptive Mirror Descent to the Rescue

- Recall that mirror descent is the (meta-)algorithm

$$w_t = \text{argmin}_w\, \psi(w) + \sum_{s=1}^{t-1} w^\top z_s.$$

For $\psi(w) = \frac{1}{\eta} \sum_{i=1}^{n} w_i \log(w_i)$, we recover (MW1).

- *Adaptive* mirror descent (Orabona et al., 2013) is the meta-algorithm

$$w_t = \text{argmin}_w\, \psi_t(w) + \sum_{s=1}^{t-1} w^\top z_s.$$

For $\psi_t(w) = \frac{1}{\eta} \sum_{i=1}^{n} w_i \log(w_i) + \eta \sum_{i=1}^{n} \sum_{s=1}^{t-1} w_i z_{s,i}^2$, we approximately recover (MW2).

  - Update: $w_{t+1,i} \propto w_{t,i} \exp(-\eta z_{t,i} - \eta^2 z_{t,i}^2) \approx w_{t,i}(1 - \eta z_{t,i})$
  - Enough to achieve better regret bound.
  - Can recover (MW2) exactly with more complicated ψ_t.

Advantages of Our Perspective

- So far, we have cast (MW2) as *adaptive* mirror descent, with regularizer $\psi_t(w) = \sum_{i=1}^{n} w_i \left[\frac{1}{\eta}\log(w_i) + \eta \sum_{s=1}^{t-1} z_{s,i}^2\right]$.
- Explains the better regret bound while staying within the mirror descent framework, which is nice.
- Our new perspective also allows us to apply lots of modern machinery:
  - optimistic updates (Rakhlin & Sridharan, 2012)
  - matrix multiplicative weights (Tsuda et al., 2005; Arora & Kale, 2007)
- By "turning the crank", we get results that beat state of the art!

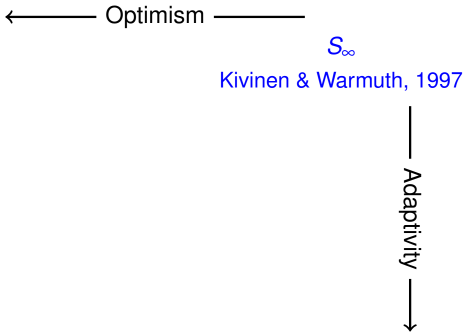
Beating State of the Art

Optimism

Adaptivity

In the above we let

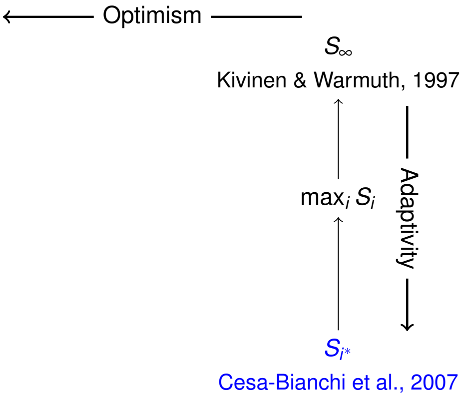
Beating State of the Art



In the above we let

$$S_\infty \stackrel{\text{def}}{=} \sum_{t=1}^{T} \|z_t\|_\infty^2$$

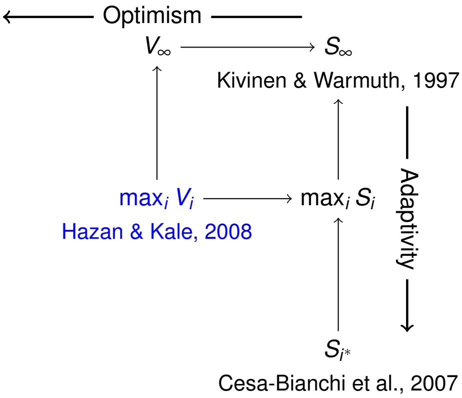
Beating State of the Art



In the above we let

$$S_\infty \stackrel{\text{def}}{=} \sum_{t=1}^{T} \|z_t\|_\infty^2$$
$$S_i \stackrel{\text{def}}{=} \sum_{t=1}^{T} z_{t,i}^2$$

Beating State of the Art

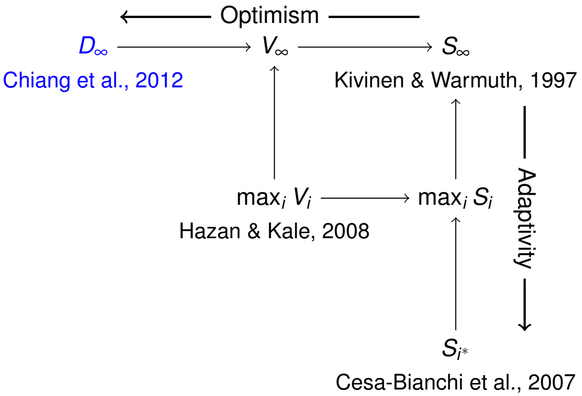


In the above we let

$$V_\infty \stackrel{\text{def}}{=} \sum_{t=1}^{T} \|z_t - \bar{z}\|_\infty^2, \quad S_\infty \stackrel{\text{def}}{=} \sum_{t=1}^{T} \|z_t\|_\infty^2$$

$$V_i \stackrel{\text{def}}{=} \sum_{t=1}^{T} (z_{t,i} - \bar{z}_i)^2, \quad S_i \stackrel{\text{def}}{=} \sum_{t=1}^{T} z_{t,i}^2$$

Beating State of the Art

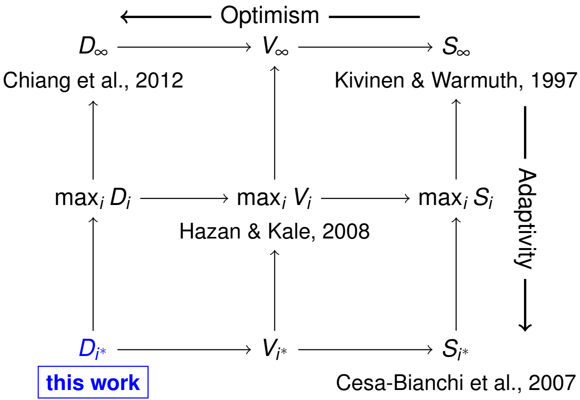


In the above we let

$$D_\infty \stackrel{\text{def}}{=} \sum_{t=1}^{T} \|z_t - z_{t-1}\|_\infty^2, \quad V_\infty \stackrel{\text{def}}{=} \sum_{t=1}^{T} \|z_t - \bar{z}\|_\infty^2, \quad S_\infty \stackrel{\text{def}}{=} \sum_{t=1}^{T} \|z_t\|_\infty^2$$

$$V_i \stackrel{\text{def}}{=} \sum_{t=1}^{T} (z_{t,i} - \bar{z}_i)^2, \quad S_i \stackrel{\text{def}}{=} \sum_{t=1}^{T} z_{t,i}^2$$

Beating State of the Art



In the above we let

$$D_\infty \stackrel{\text{def}}{=} \sum_{t=1}^{T} \|z_t - z_{t-1}\|_\infty^2, \quad V_\infty \stackrel{\text{def}}{=} \sum_{t=1}^{T} \|z_t - \bar{z}\|_\infty^2, \quad S_\infty \stackrel{\text{def}}{=} \sum_{t=1}^{T} \|z_t\|_\infty^2$$

$$D_i \stackrel{\text{def}}{=} \sum_{t=1}^{T} (z_{t,i} - z_{t-1,i})^2, \quad V_i \stackrel{\text{def}}{=} \sum_{t=1}^{T} (z_{t,i} - \bar{z}_i)^2, \quad S_i \stackrel{\text{def}}{=} \sum_{t=1}^{T} z_{t,i}^2$$

Optimistic Updates: A Brief Review

- (Normal) mirror descent:

$$w_t = \operatorname{argmin}_w \psi(w) + w^\top \left[\sum_{s=1}^{t-1} z_s \right]$$

Optimistic Updates: A Brief Review

- (Normal) mirror descent:

$$w_t = \text{argmin}_w \, \psi(w) + w^\top \left[\sum_{s=1}^{t-1} z_s \right]$$

- Optimistic mirror descent (Rakhlin & Sridharan, 2012); add a *hint* m_t:

$$w_t = \text{argmin}_w \, \psi(w) + w^\top \left[m_t + \sum_{s=1}^{t-1} z_s \right]$$

  Guesses (m_t) the next term (z_t) in the cost function.
  Pay regret in terms of $\mathbf{z_t} - \mathbf{m_t}$ rather than $\mathbf{z_t}$.

Optimistic Updates: A Brief Review

- (Normal) mirror descent:

$$w_t = \text{argmin}_w \, \psi(w) + w^\top \left[\sum_{s=1}^{t-1} z_s\right]$$

- Optimistic mirror descent (Rakhlin & Sridharan, 2012); add a *hint* m_t:

$$w_t = \text{argmin}_w \, \psi(w) + w^\top \left[m_t + \sum_{s=1}^{t-1} z_s\right]$$

  Guesses (m_t) the next term (z_t) in the cost function.
  Pay regret in terms of $\mathbf{z_t} - \mathbf{m_t}$ rather than $\mathbf{z_t}$.

- E.g.: $m_t = z_{t-1}$, $m_t = \frac{1}{t}\sum_{s=1}^{t-1} z_s$

Multiplicative Weights with Optimism

Name	Auxiliary Update	Prediction (w_t)
MW1	$\beta_{t+1,i} = \beta_{t,i} - \eta z_{t,i}$	$w_{t,i} \propto \exp(\beta_{t,i})$

Multiplicative Weights with Optimism

Name	Auxiliary Update	Prediction (w_t)
MW1	$\beta_{t+1,i} = \beta_{t,i} - \eta z_{t,i}$	$w_{t,i} \propto \exp(\beta_{t,i})$
MW2	$\beta_{t+1,i} = \beta_{t,i} - \eta z_{t,i} - \eta^2 z_{t,i}^2$	$w_{t,i} \propto \exp(\beta_{t,i})$

Multiplicative Weights with Optimism

Name	Auxiliary Update	Prediction (w_t)
MW1	$\beta_{t+1,i} = \beta_{t,i} - \eta z_{t,i}$	$w_{t,i} \propto \exp(\beta_{t,i})$
MW2	$\beta_{t+1,i} = \beta_{t,i} - \eta z_{t,i} - \eta^2 z_{t,i}^2$	$w_{t,i} \propto \exp(\beta_{t,i})$
MW3	$\beta_{t+1,i} = \beta_{t,i} - \eta z_{t,i} - \eta^2 (z_{t,i} - z_{t-1,i})^2$	$w_{t,i} \propto \exp(\beta_{t,i} - \eta z_{t-1,i})$

Multiplicative Weights with Optimism

Name	Auxiliary Update	Prediction (w_t)
MW1	$\beta_{t+1,i} = \beta_{t,i} - \eta z_{t,i}$	$w_{t,i} \propto \exp(\beta_{t,i})$
MW2	$\beta_{t+1,i} = \beta_{t,i} - \eta z_{t,i} - \eta^2 z_{t,i}^2$	$w_{t,i} \propto \exp(\beta_{t,i})$
MW3	$\beta_{t+1,i} = \beta_{t,i} - \eta z_{t,i} - \eta^2 (z_{t,i} - z_{t-1,i})^2$	$w_{t,i} \propto \exp(\beta_{t,i} - \eta z_{t-1,i})$

Regret of MW3:

$$\text{Regret} \leq \frac{\log(n)}{\eta} + \eta \sum_{t=1}^{T} (z_{t,i^*} - z_{t-1,i^*})^2$$

Dominates all existing bounds in this setting!

Summary

- Cast multiplicative weights algorithm as *adaptive* mirror descent
- Applied machinery of optimistic updates to beat best existing bounds
- Also in paper:
  - extension to general convex losses
  - extension to matrices
  - generalization of FTRL lemma to convex cones