Adaptivity and Optimism: An Improved Exponentiated Gradient Algorithm

Jacob Steinhardt, Percy Liang
Stanford University

Problem Setup

Setting is **learning from experts**:

- n experts, T rounds
- For $t = 1, \ldots, T$:
  - 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

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

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

$$w_{t+1,i} \propto w_{t,i} \exp(-\eta z_{t,i}). \qquad (2)$$

Two Different Updates

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

$$w_{t+1,i} \propto w_{t,i} \exp(-\eta z_{t,i}) \qquad \text{(MW1)}$$
$$w_{t+1,i} \propto w_{t,i}(1 - \eta z_{t,i}) \qquad \text{(MW2)}$$

Regret bounds:

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

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

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).

A Conundrum

- Mirror descent is the current gold standard of online learning algorithms.
- (MW1) is mirror descent with regularizer $\psi(w) = \sum_{i=1}^n w_i \log(w_i)$.
- (MW2) is NOT mirror descent for any regularizer, but has better performance!
- Unsettling: should we abandon our gold standard?
  - Fortunately, no: can cast (MW2) as *adaptive* mirror descent (Orabona et al., 2013).
  - We will show how to approximately recover (MW2); can recover exactly with more complex regularizer.

The first author was supported by the Hertz foundation.

Contributions

- Understanding different variants of multiplicative weights updates as ***adaptive mirror descent***
- Combining adaptive mirror descent (Orabona et al., 2013) and optimistic updates (Rakhlin & Sridharan, 2012) to obtain an **adaptive exponentiated gradient** algorithm achieving **best known bounds**

Review of Mirror Descent

Recall that mirror descent is the (meta-)algorithm

$$w_t = \arg\min_w \psi(w) + \sum_{s=1}^{t-1} w^\top z_s. \qquad (3)$$

Think of as regularized (by ψ) empirical risk minimizer. Well-understood regret bounds for arbitrary convex ψ (Shalev-Shwartz, 2011).

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

But: (MW2) is not mirror descent for *any* choice of regularizer.

Adaptive Mirror Descent to the Rescue

Adaptive mirror descent is the (meta-)algorithm

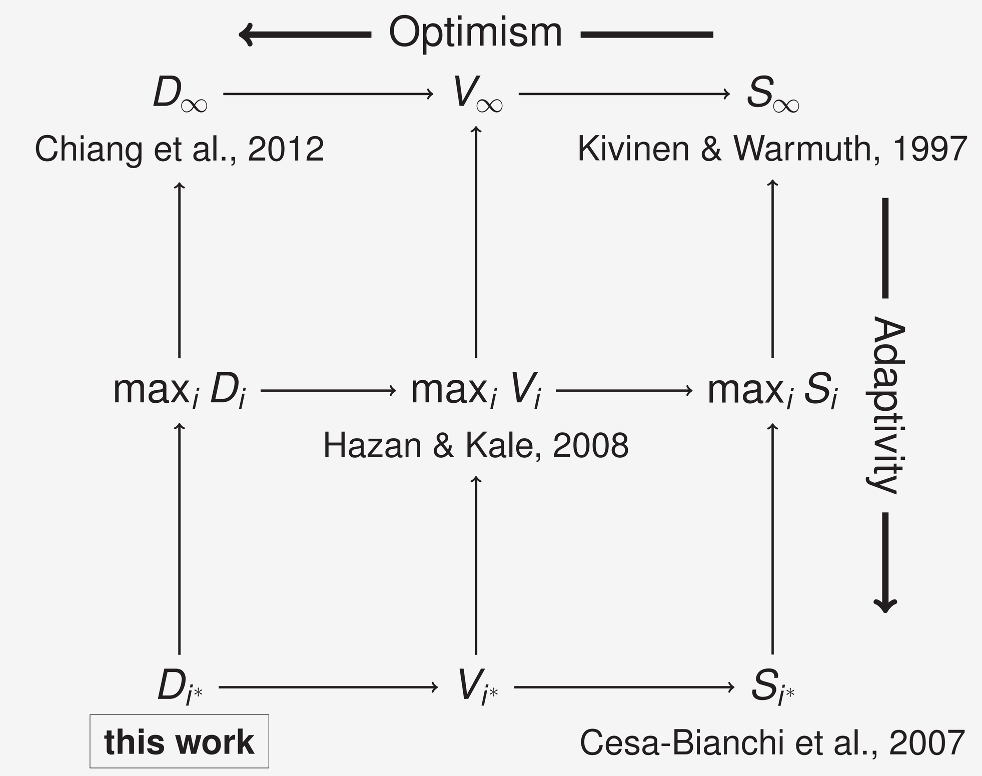
$$w_t = \arg\min_w \psi_t(w) + \sum_{s=1}^{t-1} w^\top z_s. \qquad (4)$$

Difference from mirror descent: regularizer ψ_t can now be *adaptive*. 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):

$$w_{t+1,i} \propto w_{t,i} \exp(-\eta z_{t,i} - \eta^2 z_{t,i}^2) \approx w_{t,i}(1 - \eta z_{t,i}) \qquad (5)$$

- Achieves the improved bound (Regret:MW2).
- Intuition: we penalize experts with high quadratic variation, which makes it easier to find the best expert when it has low variation.

An Improved Algorithm and Bound



Name	Auxiliary Updates	Prediction (w_t)	Source
EG (MW1)	$\beta_{t+1} = \beta_t - \eta z_t$	$\exp(\beta_t)$	Kivinen and Warmuth [1997]
MW2	$\beta_{t+1,i} = \beta_{t,i} + \log(1 - \eta z_{t,i})$	$\exp(\beta_t)$	Cesa-Bianchi et al. [2007]
Variation-MW	$\beta_{t+1,i} = \beta_{t,i} - \eta z_{t,i} - 4\eta^2(z_{t,i} - m_{t,i})^2$ $m_t = \frac{1}{t}\sum_{s=1}^{t-1} z_s$	$\exp(\beta_t)$	Hazan and Kale [2008]
Optimistic MW	$\beta_{t+1,i} = \beta_{t,i} - \eta z_{t,i}$	$\exp(\beta_t - \eta z_{t-1})$	Chiang et al. [2012]
AEG-Path	$\beta_{t+1,i} = \beta_{t,i} - \eta z_{t,i} - \eta^2(z_{t,i} - z_{t-1,i})^2$	$\exp(\beta_t - \eta z_{t-1})$	this work
AMEG-Path	$B_{t+1} = B_t - \eta Z_t - \eta^2(Z_t - Z_{t-1})^2$	$\exp(B_t - \eta Z_{t-1})$	this work

$A \to B$ means that A is strictly better than B. In the above we let

$$D_\infty \stackrel{\text{def}}{=} \sum_{t=1}^T \|z_t - z_{t-1}\|_\infty^2,\ V_\infty \stackrel{\text{def}}{=} \sum_{t=1}^T \|z_t - \bar{z}\|_\infty^2,\ S_\infty \stackrel{\text{def}}{=} \sum_{t=1}^T \|z_t\|_\infty^2$$
$$D_i \stackrel{\text{def}}{=} \sum_{t=1}^T (z_{t,i} - z_{t-1,i})^2,\ V_i \stackrel{\text{def}}{=} \sum_{t=1}^T (z_{t,i} - \bar{z}_i)^2,\ S_i \stackrel{\text{def}}{=} \sum_{t=1}^T z_{t,i}^2$$

Final bound:

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

With our new perspective, we get the result simply by "turning the crank" on modern online learning machinery!

Proof Technique: Optimistic Updates

We achieve our bound using optimistic updates (Rakhlin & Sridharan, 2012).

Optimistic mirror descent incorporates a *hint* m_t:

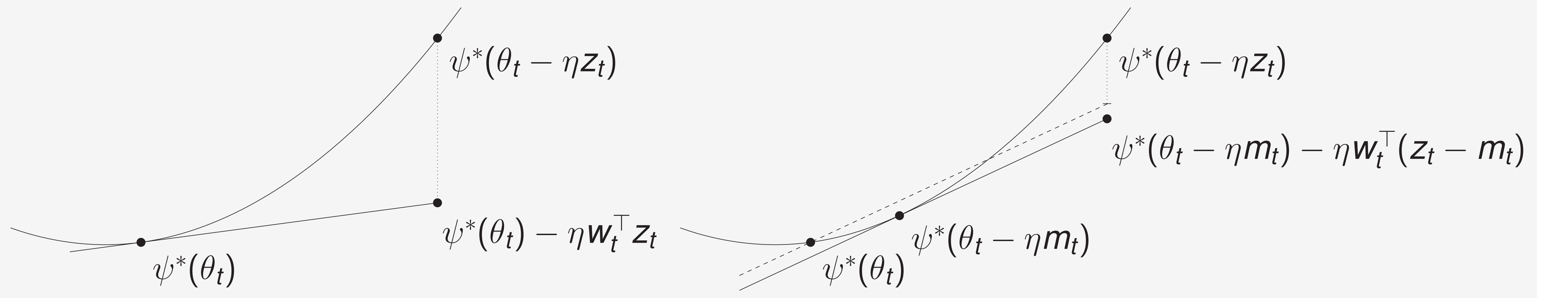
$$w_t = \arg\min_w \psi(w) + w^\top \left[m_t + \sum_{s=1}^{t-1} z_s \right] \qquad (6)$$

Compare to normal mirror descent (3).

Adds a guess (m_t) of the next term (z_t) in the empirical loss. Pay regret in terms of $z_t - m_t$ rather than z_t.

- Usually: $m_t = z_{t-1}$, or $m_t = \frac{1}{t}\sum_{s=1}^{t-1} z_s$

Geometric illustration:



Explanation: bounds for mirror descent are obtained by a potential function argument on the minimum regularized empirical risk, denoted $\psi^*(\theta_t)$ where $\theta_t = -\eta \sum_{s=1}^{t-1} z_s$. Regret is bounded by Bregman divergence (dotted line) of ψ^* between θ_t and $\theta_t - \eta z_t$. Moving in the direction of m_t can decrease this divergence.

Extensions

Matrices

Extension to matrix setting is important for certain combinatorial approximation algorithms (Tsuda et al. 2005; Arora & Kale, 2007).

Now nature plays *matrices* $Z_1, \ldots, Z_T$, $-I \preceq Z_t \preceq I$. Similarly, learner chooses matrices $W_1, \ldots, W_T$, $W_t \succeq 0$, $\text{tr}(W_t) = 1$. If W_t, Z_t are diagonal, recover vector setting.

If we replace $\psi(w) = \sum_{i=1}^n w_i \log(w_i)$ with the *von-Neumann entropy* $\psi(W) = \text{tr}(W \log(W))$, then same ideas from before still work.

Convex Losses and Unconstrained Optimization

All the algorithms extend to general convex loss functions $f_1, \ldots, f_T$ by setting $z_t = \partial f_t(w_t)$ and applying the same updates as before.

We can extend the algorithms to the non-negative orthant by simply not renormalizing to the simplex (still need to initialize weights to $\frac{1}{n}$).

We can go from the non-negative orthant to $\mathbb{R}^n$ by keeping track of *two* weight vectors w_+ and w_-, with loss $(w_+ - w_-)^\top z$.