Leiden, November 18, 2014

OPEN
PROBLEM

Tuning the Learning Rate: Contradictory Results

Tim van Erven

Universiteit
Leiden

Adapting to Easy Data

- Online convex optimization: [Hazan,Agarwal,Kale,2007], [Bartlett,Hazan,Rakhlin,2007]
  - Easy: objective function is **strongly convex**
  - Exploit by using extra **small learning rate**

Adapting to Easy Data

- Online convex optimization: [Hazan,Agarwal,Kale,2007], [Bartlett,Hazan,Rakhlin,2007]
  - Easy: objective function is **strongly convex**
  - Exploit by using extra **small learning rate**
- 2^{nd}-order Bounds in Prediction with expert advice: [Cesa-Bianchi,Mansour,Stoltz,2007], [de Rooij,vE,Grünwald,Koolen,2014]
  - Easy: **variance** of exponential weights is **small**
  - Exploit by using extra **large learning rate**

Adapting to Easy Data

- Online convex optimization: [Hazan,Agarwal,Kale,2007], [Bartlett,Hazan,Rakhlin,2007]
  - Easy: objective function is **strongly convex**
  - Exploit by using extra **small learning rate**
- 2^{nd}-order Bounds in Prediction with expert advice: [Cesa-Bianchi,Mansour,Stoltz,2007], [de Rooij,vE,Grünwald,Koolen,2014]
  - Easy: **variance** of exponential weights is **small**
  - Exploit by using extra **large learning rate**
- Contradiction: what if both notions of “easy” occur at the same time?

Online Convex Optimization

- For $t = 1, \ldots, T$
  - Choose w_t from convex set $\mathcal{W} \subset \mathbb{R}^{\mathcal{K}}$
  - Observe convex function $f_t : \mathcal{W} \to \mathbb{R}$

- Regret:

$$\mathcal{R}_T = \sum_{t=1}^{T} f_t(w_t) - \min_{w \in \mathcal{W}} \sum_{t=1}^{T} f_t(w)$$

“Easy”: f_t is Strongly Convex

- H-strong convexity:

$$f_t(w) \geq f_t(u) + \nabla f_t(u)^\top (w - u) + \frac{H}{2} \|w - u\|_2^2$$

"Easy": f_t is Strongly Convex

- H-strong convexity:

$$f_t(w) \geq f_t(u) + \nabla f_t(u)^\top (w-u) + \frac{H}{2}\|w-u\|_2^2$$

- More general: H-strong convexity w.r.t. convex function h:

$$f_t(w) \geq f_t(u) + \nabla f_t(u)^\top (w-u) + \frac{H}{2} B_h(w,u)$$

where B_h is the Bregman divergence w.r.t. h. E.g. $B_h(w,u) = \|w-u\|_2^2$ or $B_h(w,u) = \mathrm{KL}(w\|u)$

Algorithm: Mirror Descent

- Gradient at our predictions:
  $g_t = \nabla f_t(w_t)$ assumed bounded $\|g_t\| \leq G$
- Mirror descent:

$$w_{t+1} = \arg\min_{w \in \mathcal{W}} \langle w, g_t \rangle + \frac{1}{\eta_{t+1}} B_h(w, w_t)$$

Algorithm: Mirror Descent

- Gradient at our predictions:
  $g_t = \nabla f_t(w_t)$ assumed bounded $\|g_t\| \leq G$
- Mirror descent:

$$w_{t+1} = \arg\min_{w \in \mathcal{W}} \langle w, g_t \rangle + \frac{1}{\eta_{t+1}} B_h(w, w_t)$$

- For H-strongly convex w.r.t. h: small learning rate

$$\eta_t = \frac{2}{Ht}$$

  gives small $O(\log T)$ regret

Prediction with Expert Advice*

- Instance of OCO
- K experts get loss $\ell_t \in [0,1]^K$ in round t
- $w \in \mathcal{W} = \Delta^K$ is probability vector on experts

- $f_t(w) = \langle w, \ell_t \rangle$
  - is linear in w
  - gradient $g_t = \ell_t$

* Original definition used by Vovk is more general

Algorithm: FTRL for KL Divergence

- Follow-the-Regularized-Leader:

$$w_{t+1} = \arg\min_{w \in \mathcal{W}} \langle w, \sum_{s \leq t} g_s \rangle + \frac{1}{\eta_{t+1}} B_h(w, w_0)$$

- Take $B_h(w, w_0) = \mathrm{KL}(w \| w_0)$

- Almost same as mirror descent except update over all data instead of from last data-point.
- Equivalent if η_t would not vary with t.

“Easy”: Small Variance

- $v(w_t) =$ variance of w_t $\qquad V_T = \sum_{t \leq T} v(w_t)$
- 2nd-order bound on regret:

$$\mathcal{R}_T = O(\sqrt{\log(K) V_T})$$

for $\eta_t \propto \sqrt{\log(K)/V_{t-1}}$

“Easy”: Small Variance

- $v(w_t) =$ variance of w_t $\qquad V_T = \sum_{t \le T} v(w_t)$
- 2nd-order bound on regret:

$$\mathcal{R}_T = O(\sqrt{\log(K) V_T})$$

  for $\eta_t \propto \sqrt{\log(K)/V_{t-1}}$
- If V_T small: regret small, η_t large
- If weights w_t concentrate fast on single expert, then $V_T = O(1)$, $\eta_t \geq$ constant.

PwEA algs do not exploit strong convexity

- Can do linear approximation of f_t:

$$f_t(w^*) \geq f_t(w_t) + g_t^\top (w^* - w_t)$$

PwEA algs do not exploit strong convexity

- Can do linear approximation of f_t:

$$f_t(w^*) \geq f_t(w_t) + g_t^\top (w^* - w_t)$$

$$f_t(w_t) - f_t(w^*) \leq \langle w_t, g_t \rangle - \langle w^*, g_t \rangle$$

PwEA algs do not exploit strong convexity

- Can do linear approximation of f_t:

$$f_t(w^*) \geq f_t(w_t) + g_t^\top (w^* - w_t)$$
$$f_t(w_t) - f_t(w^*) \leq \langle w_t, g_t \rangle - \langle w^*, g_t \rangle$$

- Then 2nd-order FTRL algs work for any convex f_t, not just linear.
- But linear approximation throws away strong convexity.

OCO Algs Do Not Exploit Small Variance

- Worst-case PwEA analysis:
  - **Hoeffding's bound** on cumulant generating function
- To get 2nd-order bounds:
  - Replace by **Bernstein's bound**, which brings in the variance

OCO Algs Do Not Exploit Small Variance

- Worst-case PwEA analysis:
  - **Hoeffding's bound** on cumulant generating function
- To get 2nd-order bounds:
  - Replace by **Bernstein's bound**, which brings in the variance
- OCO analysis:

Mirror descent update without projection

$$\frac{1}{\eta_t} B_h(w_t, \tilde{w}_{t+1}) \leq \eta_t \|g_t\|_\infty^2$$

which is basically Hoeffding's bound

Adapting to Easy Data

- Online convex optimization: [Hazan,Agarwal,Kale,2007], [Bartlett,Hazan,Rakhlin,2007]
  - Easy: objective function is **strongly convex**
  - Exploit by using extra **small learning rate**
- 2nd-order Bounds in Prediction with expert advice: [Cesa-Bianchi,Mansour,Stoltz,2007], [de Rooij,vE,Grünwald,Koolen,2014]
  - Easy: **variance** of exponential weights is **small**
  - Exploit by using extra **large learning rate**
- Can we adapt to both notions of “easy” at the same time?

References

- Hazan, Agarwal, Kale, **Logarithmic regret algorithms for online convex optimization**, Machine Learning, 2007.
- Bartlett, Hazan, Rakhlin, **Adaptive Online Gradient Descent**, NIPS 2007.
- Cesa-Bianchi, Mansour, Stoltz, **Improved second-order bounds for prediction with expert advice**, Machine Learning, 2007.
- De Rooij, Van Erven, Grünwald, Koolen, **Follow the Leader If You Can, Hedge If You Must**, JMLR, 2014.

2nd-order Bounds

- Crucial in PwEA analysis is **mixability gap**:

$$\delta_t = \langle w_t, g_t \rangle + \frac{1}{\eta_t} \log \langle w_t, e^{-\eta_t g_t} \rangle$$

= approximation error when linear objective is approximated by exp-concave objective

- Hoeffding's inequality => slow rates:

$$\delta_t \leq \eta_t \|g_t\|_\infty^2$$

- Bernstein's inequality => fast rates:

$$\delta_t \lesssim \eta_t v_t \|g_t\|_\infty^2$$

leads to 2nd-order bounds

OCO Algs Do Not Exploit Small Variance

- Update one step ahead:

$$\bar{w}_{t+1} = \arg\min_{w \in \mathcal{W}} \langle w, g_t \rangle + \frac{1}{\eta_t} B_h(w, w_t)$$

$$\tilde{w}_{t+1} = \arg\min_{w \in \mathbb{R}^K} \langle w, g_t \rangle + \frac{1}{\eta_t} B_h(w, w_t)$$

$$\eta_t \delta_t = B_h(w_t, \bar{w}_{t+1}) \leq B_h(w_t, \tilde{w}_{t+1})$$

- OCO analysis looks like Hoeffding's bound:

$$\frac{1}{\eta_t} B_h(w_t, \tilde{w}_{t+1}) \leq \eta_t \|g_t\|_\infty^2$$