Accelerating Optimization via Adaptive Prediction

Mehryar Mohri[1] Scott Yang[2]

[1]Google, New York University

[2]New York University

NIPS Easy Data II, Dec 10, 2015

Learning Scenario and Set-Up

Online Convex Optimization

- Sequential optimization problem
- $\mathcal{K} \subset \mathbb{R}^n$ compact action space, f_t convex loss functions
- At time t, learner chooses action x_t, receives loss function f_t, and suffers loss $f_t(x_t)$
- Goal: minimize regret

$$\max_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x_t) - f_t(x)$$

Worst-case vs Data-dependent Methods

Worst-case methods:

1. Algorithms: Mirror Descent, FTRL
2. Regret bounds typically of the form $\mathcal{O}(\sqrt{T})$
3. Algorithms do not give faster rates on "easy data"

Data-dependent methods:

1. Adaptive regularization [Duchi et al 2010]
   Easy data: sparsity
2. Predictable sequences [Rakhlin and Sridharan 2012]
   Easy data: slowly-varying gradients

Adaptive Regularization

AdaGrad algorithm of [Duchi et al 2010] (+ many others):

1. Standard Mirror Descent:
   $x_{t+1} = \text{argmin}_{x \in \mathcal{K}} \, g_t \cdot x + B_\psi(x, x_t)$.
2. Adaptivity: change the regularizer at each time step
   $\psi \longrightarrow \psi_t$.
3. Worst-case optimal data-dependent bound:
   $\mathcal{O}\left(\sum_{i=1}^{n} \sqrt{\sum_{t=1}^{T} |g_{t,i}|^2}\right)$
4. Easy data scenario: sparsity

Predictable Sequences

Optimistic FTRL algorithm of [Rakhlin and Sridharan 2012]
Idea:

- Learner should try to "predict" the next gradient $M_t(g_1, \ldots, g_{t-1}) \approx g_t$.

Consequences:

- Typical regret bound $\mathcal{O}\left(\sqrt{\sum_{t=1}^{T} |g_t - M_t|_2^2}\right)$.
- Often still worst-case optimal
- Easy data scenario: slowly varying gradients

Adaptive Predictions

Motivation:

- Adaptive regularization good for sparsity
- Predictable sequences good for slowly varying gradients

Questions:

- Can we combine both and get the best of both worlds?
- What are the easy data scenarios for such an algorithm?

Adaptive Predictions

Idea:

- Derive an adaptive norm bound for optimistic FTRL: $\mathcal{O}\left(\sum_{t=1}^{T} |g_t - M_t|_{(t),*}\right)$
- Find "best" norm associated to gradient prediction error instead of gradient losses.

Consequences:

- Can view AdaGrad as special case of naively predicting zero
- Can view Optimistic FTRL as naive regularization
- Behaves well under sparsity
- Accelerates faster than Optimistic FTRL when predictions vary in per-coordinate accuracy

Practical Considerations

Extensions:

- Composite terms
- Proximal versus non-proximal regularization
- Large-scale optimization problems: epoch-based variants

For more details, please stop by the poster. Thank you!