Accelerating Optimization via Adaptive Prediction

Mehryar Mohri
Courant Institute and Google
251 Mercer Street
New York, NY 10012
mohri@cims.nyu.edu

Scott Yang
Courant Institute
251 Mercer Street
New York, NY 10012
yangs@cims.nyu.edu

Abstract

We present a general framework for designing data-dependent optimization algorithms, building upon and unifying recent techniques in adaptive regularization and optimistic gradient predictions. We first provide a general regret guarantee that holds at any time and under minimal assumptions, and then show how different relaxations recover existing algorithms, both basic as well as more recent sophisticated ones. Finally, we show how combining adaptivity and optimism can guide the design of algorithms that benefit from more favorable guarantees than recent state-of-the-art methods.

1 Introduction

In the standard scenario of online convex optimization [11], at each round $t = 1, 2, \ldots$, the learner selects a point x_t out of a compact convex set $\mathcal{K}$ and incurs loss $f_t(x_t)$, where f_t is a convex function defined over $\mathcal{K}$. The learner's objective is to find an algorithm $\mathcal{A}$ that minimizes regret:

$$\mathrm{Reg}_T(\mathcal{A}) = \max_{x \in \mathcal{K}} \mathrm{Reg}_T(\mathcal{A}, x), \text{ where } \mathrm{Reg}_T(\mathcal{A}, x) = \sum_{t=1}^{T} f_t(x_t) - f_t(x)$$

that is the difference between the learner's cumulative loss and the loss of the best x in $\mathcal{K}$. We will assume only that the learner has access to the gradient or an element of the sub-gradient of the loss functions f_t, but that the loss functions f_t can be arbitrarily singular and flat, e.g. not necessarily strongly convex or strongly smooth.

In the scenario just presented, minimax optimal rates can be achieved by standard algorithms such as online gradient descent [11]. However, general minimax optimal rates may be too conservative. Recently, *adaptive regularization* methods have been introduced for standard descent methods to achieve tighter data-dependent regret bounds (see [1], [4], [8], [7], [9]). Specifically, in the "AdaGrad" framework of [4], there exists a sequence of convex functions ψ_t such that the update $x_{t+1} = \mathrm{argmin}_{x \in \mathcal{K}} \eta g_t^\top x + B_{\psi_t}(x, x_t)$ yields regret: $\mathrm{Reg}_T(\mathcal{A}, x) \leq \sqrt{2} \max_t \|x - x_t\|_\infty \sum_{i=1}^n \sqrt{\sum_{t=1}^T |g_{t,i}|^2}$, where $g_t \in \partial f_t(x_t)$ is an element of the subgradient of f_t at x_t, $g_{1:T,i} = \sum_{t=1}^T g_{t,i}$, and B_{ψ_t} is the Bregman divergence defined using the convex function ψ_t. This upper bound on the regret has shown to be within a factor $\sqrt{2}$ of the optimal a posteriori regret. However, this upper bound on the regret can still be very large, even if the functions f_t admit some

Algorithm 1 Composite Adaptive Optimistic Follow-the-Regularized-Leader

```
Input: regularization function $r_0 \geq 0$, composite functions $\{\zeta_t\}_{t=1}^{\infty}$ where $\zeta_t \geq 0$.
Initialize: $\tilde{g}_1 = 0$, $x_1 = \text{argmin}_{x \in \mathcal{K}} r_0(x)$.
for $t = 1, \ldots, T$: do
    Compute $g_t \in \partial f_t(x_t)$.
    Construct regularizer $r_t \geq 0$.
    Predict the next gradient $\tilde{g}_{t+1} = \tilde{g}_{t+1}(g_1, \ldots, g_t, x_1, \ldots, x_t)$.
    Update $x_{t+1} = \text{argmin}_{x \in \mathcal{K}} g_{1:t} \cdot x + \tilde{g}_{t+1} \cdot x + r_{0:t}(x) + \zeta_{1:t+1}(x)$.
end for
```

favorable properties (e.g. $f_t \equiv f$, linear). This is because the dependence is directly on the norm of g_ts.

An alternative line of research has been investigated by a series of recent publications that have analyzed online learning in "slowly-varying" scenarios [5, 3, 10, 2]. If $\mathcal{R}$ is a self-concordant function, $\|\cdot\|_{\nabla^2\mathcal{R}(x_t)}$ is the semi-norm induced by its Hessian at the point x_t,[1] and $\tilde{g}_{t+1} = \tilde{g}_{t+1}(g_1, \ldots, g_t, x_1, \ldots, x_t)$ is a "prediction" of a time $t+1$ subgradient g_{t+1} based on information up to time t, then one can obtain regret bounds of the following form: $\text{Reg}_T(\mathcal{A}, x) \leq \frac{1}{\eta}\mathcal{R}(x) + 2\eta \sum_{t=1}^{T} \|g_t - \tilde{g}_t\|_{\nabla^2\mathcal{R}(x_t),*}$. Here, $\|\cdot\|_{\nabla^2\mathcal{R}(x_t),*}$ denotes the dual norm of $\|\cdot\|_{\nabla^2\mathcal{R}(x_t)}$: for any x, $\|x\|_{\nabla^2\mathcal{R}(x_t),*} = \sup_{\|y\|_{\nabla^2\mathcal{R}(x_t)} \leq 1} x^T y$. This guarantee can be very favorable in the *optimistic* case where $\tilde{g}_t \approx g_t$ for all t. Nevertheless, it admits the drawback that much less control is available over the induced norm since it is difficult to predict, for a given self-concordant function $\mathcal{R}$, the behavior of its Hessian at the points x_t selected by an algorithm. Moreover, there is no guarantee of "near-optimality" with respect to an optimal a posteriori regularization as there is with the adaptive algorithm.

2 Adaptive and Optimistic Follow-the-Regularized-Leader algorithm

In view of the discussion in the previous section, we present an adaptive and optimistic version of the Follow-the-Regularized-Leader (FTRL) family of algorithms. In each round of standard FTRL, a point is chosen that is the minimizer of the average linearized loss incurred plus a regularization term. In our new version of FTRL, we will find a minimizer of not only the average loss incurred, but also a prediction of the next round's loss. In addition, we will define a dynamic time-varying sequence of regularization functions that can be used to optimize against this new loss term. Algorithm 1 shows the pseudocode of our Composite Adaptive and Optimistic Follow-the-Regularized-Leader (CAO-FTRL) algorithm, where we also allow for objective functions with composite terms that we may wish to optimize directly.

The following result provides a regret guarantee for the algorithm when one uses proximal regularizers, i.e. functions r_t such that $\text{argmin}_{x \in \mathcal{K}} r_t(x) = x_t$.

Theorem 1 (CAO-FTRL-Prox). *Let $\{r_t\}$ be a sequence of proximal non-negative functions, such that $\text{argmin}_{x \in \mathcal{K}} r_t(x) = x_t$, and let $\tilde{g}_t$ be the learner's estimate of g_t given the history of functions $f_1, \ldots, f_{t-1}$ and points $x_1, \ldots, x_{t-1}$. Let $\{\zeta_t\}_{t=1}^{\infty}$ be a sequence of non-negative convex functions, such that $\zeta_1(x_1) = 0$. Assume further that the function $h_{0:t} : x \mapsto g_{1:t} \cdot x + \tilde{g}_{t+1} \cdot x + r_{0:t}(x) + \zeta_{1:t+1}(x)$ is 1-strongly convex with respect to some norm $\|\cdot\|_{(t)}$. Then the following regret bounds*

[1] The norm induced by a symmetric positive definite (SPD) matrix A is defined for any x by $\|x\|_A = \sqrt{x^\top A x}$.

hold for CAO-FTRL (Algorithm 1):

$$\sum_{t=1}^{T} f_t(x_t) - f_t(x) \leq \zeta_{1:T-1}(x) + r_{0:T-1}(x) + \sum_{t=1}^{T} \|g_t - \tilde{g}_t\|_{(t-1),*}^2$$
$$\sum_{t=1}^{T} [f_t(x_t) + \zeta_t(x_t)] - [f_t(x) + \zeta_t(x)] \leq r_{0:T}(x) + \sum_{t=1}^{T} \|g_t - \tilde{g}_t\|_{(t),*}^2 .$$

The regret bound just presented can be vastly superior to the adaptive methods of [4], [8], and others. For instance, one common choice of gradient prediction is $\tilde{g}_{t+1} = g_t$, so that for slowly varying gradients (e.g. nearly "flat" functions), $g_t - \tilde{g}_t \approx 0$, but $\|g_t\|_{(t)} = \|g\|_{(t)}$. Moreover, for reasonable gradient predictions, $\|\tilde{g}_{t+1}\|_{(t)} \approx \|g_t\|_{(t)}$ generally, so that in the worst case, Algorithm 1's regret will be at most a factor of two more than standard methods. At the same time, the use of non self-concordant regularization allows one to more explicitly control the induced norm in the regret bound as well as provide more efficient updates than those of [10].

We now present a series of reductions of AO-FTRL to known algorithms.

Corollary 1. *With the following suitable choices of the parameters in Theorem 1, the following regret bounds can be recovered:*

1. *Adaptive FTRL-Prox of [7] (up to a constant factor of 2):* $\tilde{g} \equiv 0$, $\zeta_t \equiv 0$.
2. *Primal-Dual AdaGrad of [4]:* $r_{0:t} = \psi_t$, $\tilde{g} \equiv 0$.
3. *Optimistic FTRL of [10]:* $r_0 = \eta\mathcal{R}$ *where* $\eta > 0$ *and* $\mathcal{R}$ *a self-concordant function,* $r_t = 0, \forall t \geq 1$, $\zeta_t \equiv 0$.

We now demonstrate how one can build an "optimistic" version of the well-known AdaGrad algorithm, which can be viewed as an adaptive version of Gradient Descent. In our version, our algorithm will *accelerate* when the gradient predictions become more accurate, as opposed to when the gradients are simply smaller.

Corollary 2 (Adaptive and Optimistic Gradient Descent)**.** *Let* $\mathcal{K} \subset \times_{i=1}^{n}[-R_i, R_i]$ *be an* n*-dimensional rectangle, and denote* $\Delta_{s,i} = \sqrt{\sum_{a=1}^{s}(g_{a,i} - \tilde{g}_{a,i})^2}$. *Set* $r_{0:t} = \sum_{i=1}^{n}\sum_{s=1}^{t}\frac{\Delta_{s,i}-\Delta_{s-1,i}}{2R_i}(x_i - x_{s,i})^2$. *Then, if we use the martingale-type gradient prediction* $\tilde{g}_{t+1} = g_t$, *the following regret bound holds:* $Reg_T(\text{AO-GD}, x) \leq 4\sum_{i=1}^{n} R_i\sqrt{\sum_{t=1}^{T}(g_{t,i} - g_{t-1,i})^2}$.

Moreover, this regret bound is nearly equal to the optimal a posteriori regret bound: $R_i\sum_{i=1}^{n}\sqrt{\sum_{t=1}^{T}(g_{t,i} - g_{t-1,i})^2} = \max_i R_i\sqrt{n \inf_{s\succcurlyeq 0,\langle s,1\rangle\leq n}\sum_{t=1}^{T}\|g_t - g_{t-1}\|_{diag(s)^{-1}}^2}$.

Notice that the regularization function is minimized when the gradient predictions become more accurate. Thus, if we interpret our regularization as an implicit learning rate, our algorithm uses a larger learning rate and *accelerates* as our gradient predictions become more accurate. This is in stark contrast to other adaptive regularization methods, such as AdaGrad, where learning rates are simply inversely proportional to the norm of the gradient.

Moreover, since the regularization function decomposes over the coordinates, this acceleration can occur on a per-coordinate basis. If our gradient predictions are more accurate in some coordinates than others, then our algorithm will be able to adapt accordingly. Under the simple martingale prediction scheme, this means that our algorithm will be able to accelerate when only certain coordinates of the gradient are slowly-varying, even if the entire gradient is not.

Algorithm 2 CAOS-Reg-ERM-Epoch

```
Input: scaling constant $\alpha > 0$, composite term $\zeta$, $r_0 = 0$.
Initialize: initial point $x_1 \in \mathcal{K}$, distribution $p_1$.
Sample $j_1$ according to $p_1$, and set $t = 1$.
for $s = 1, \ldots, k$: do
    Compute $\bar{g}_s^j = \nabla f_j(x_1) \; \forall j \in \{1, \ldots, m\}$.
    for $a = 1, \ldots, T/k$: do
        If $T \mod k = 0$, compute $g^j = \nabla f_j(x_t) \; \forall j$.
        Set $\hat{g}_t = \frac{g_t^{j_t}}{p_{t,j_t}}$, and construct $r_t \geq 0$.
        Sample $j_{t+1} \sim p_{t+1}$ and set $\tilde{g}_{t+1} = \frac{\bar{g}_s^{j_t}}{p_{t,j_t}}$.
        Update $x_{t+1} = \text{argmin}_{x \in \mathcal{K}} \hat{g}_{1:t} \cdot x + \tilde{g}_{t+1} \cdot x + r_{0:t}(x) + (t+1)\alpha\zeta(x)$ and $t = t + 1$.
    end for
end for
```

3 Application to Stochastic Regularized Empirical Risk Minimization

Many learning algorithms can be viewed as instances of regularized empirical risk minimization (e.g. SVM, Logistic Regression, Lasso), where the goal is to minimize an objective function of the following form: $H(x) = \sum_{j=1}^m f_j(x) + \alpha\zeta(x)$.

We present here a refinment of the commonly used stochastic gradient descent. For simplicity, we elect to use as gradient prediction the last gradient of the current function being sampled, f_j. However, we may run into the problem of never seeing a function before. A logical modification would be to separate optimization into epochs and do a full batch update over all functions f_j at the start of each epoch. This is similar to the technique used in the Stochastic Variance Reduced Gradient (SVRG) algorithm of [6]. However, we do not assume extra function regularity as they do in their paper, so the bounds are not comparable. The algorithm is presented in Algorithm 2 and comes with the following guarantee:

Corollary 3. *Assume $\mathcal{K} \subset \times_{i=1}^n [-R_i, R_i]$. Denote $\Delta_{s,i} = \sqrt{\sum_{a=1}^s (\hat{g}_{a,i} - \tilde{g}_{a,i})^2}$, and let $r_{0:t} = \sum_{i=1}^n \sum_{s=1}^t \frac{\Delta_{s,i} - \Delta_{s-1,i}}{2R_i} (x_i - x_{s,i})^2$ be the adaptive regularization. Then the regret of Algorithm 2 is bounded by: $\mathbb{E}\left[\sum_{t=1}^T f_t(x_t) + \alpha\zeta(x_t) - f_t(x) - \alpha\zeta(x)\right] \leq \sum_{i=1}^n 4R_i \sqrt{\sum_{s=1}^k \sum_{t=(s-1)(T/k)+1}^{(s-1)(T/k)+T/k} \sum_{j=1}^m \frac{\left|g_{t,i}^j - \bar{g}_{s,i}^j\right|^2}{p_{t,j}}}$*

Moreover, if $\|\nabla f_j\|_\infty \leq L_j \; \forall j$, then setting $p_{t,j} = \frac{L_i}{\sum_{j=1}^m L_j}$ yields a worst-case bound of: $8 \sum_{i=1}^n R_i \sqrt{T \left(\sum_{j=1}^m L_j\right)^2}$.

4 Conclusion

We presented a general framework for developing efficient adaptive and optimistic algorithms for online convex optimization. Building upon recent advances in adaptive regularization and predictable online learning, we improved upon each method. We demonstrated the power of this approach by deriving algorithms with guarantees than can perform much better on easier data than those commonly used in practice.

References

[1] Peter L. Bartlett, Elad Hazan, and Alexander Rakhlin. Adaptive online gradient descent. In *NIPS*, pages 65–72, 2007.

[2] Chao-Kai Chiang, Chia-Jung Lee, and Chi-Jen Lu. Beating bandits in gradually evolving worlds. In *COLT*, pages 210–227, 2013.

[3] Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *COLT*, pages 6.1–6.20, 2012.

[4] John C. Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. In *COLT*, pages 257–269, 2010.

[5] Elad Hazan and Satyen Kale. Better algorithms for benign bandits. In *SODA*, pages 38–47, 2009.

[6] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pages 315–323, 2013.

[7] H. Brendan McMahan. Analysis techniques for adaptive online learning. *CoRR*, 2014.

[8] H. Brendan McMahan and Matthew J. Streeter. Adaptive bound optimization for online convex optimization. In *COLT*, pages 244–256, 2010.

[9] Francesco Orabona, Koby Crammer, and Nicolò Cesa-Bianchi. A generalized online mirror descent with applications to classification and regression. *CoRR*, 2013.

[10] Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *COLT*, pages 993–1019, 2013.

[11] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pages 928–936, 2003.