Non-convex online learning via algorithmic equivalence

Udaya Ghai[1 2], Zhou Lu [1 2], Elad Hazan[1 2],
[1]Google AI Princeton [2]Department of Computer Science, Princeton University

Motivation

To explain the success of modern deep learning, the study of **global convergence** of gradient descent for **non-convex** objectives is increasingly important, because in practice gradient descent and its variants can achieve zero error on a highly non-convex loss function of a deep neural network.

Inspired by recent results in continuous-time, we investigate a **algorithmic equivalence** methodology for proving convergence of non-convex functions that are **reparameterizations** of convex functions.

Continuous-time Reparameterization

[1] analyzes equivalence of **gradient flow** and **mirror flow**. In particular, the **ODE** for mirror flow on f with regularizer R

$$\dot{\nabla R}(x(t)) = -\eta \nabla f(x(t))$$

is equivalent to gradient flow on $\tilde{f} = f \circ u$ with $x(t) = q(u(t)$

$$\dot{u(t)} = -\eta \tilde{f}(u(t)) \ ,$$

where

$$[\nabla^2 R(x)]^{-1} = J_q(u) J_q(u)^\top .$$

mirror flow on **convex** f

⇕

gradient flow on **nonconvex** function $\tilde{f}$

In a follow-up work [2], the analysis for continuous-time was extended to discrete-time, on some specific algorithms with relative-entropy regularization.

Canonical example: **Exponentiated Gradient (EG)**

- R is negative entropy, $q(u) = u \odot u$
- Analyzed in discrete online settings in [2]

Open question by [1,2]: can we extend this reparameterization approach to **general online convex optimization**, in the discrete-time setting?

Our Result

We show that in the **discrete-time** setting, online gradient descent applied to **non-convex** functions is an **approximation** of online mirror descent applied to convex functions under reparameterization, through a new **algorithmic equivalence** technique.

2 Algorithm

Algorithm 1 Online Mirror Descent

```
Input: Initialization x_1 ∈ K, regularizer R.
for t = 1, ..., T do
    Predict x_t, observe ∇f_t(x_t)
    Update
        y_{t+1} = (∇R)^{-1}(∇R(x_t) − η∇f_t(x_t))
        x_{t+1} = Π^R_K(y_{t+1})
end for
```

$$y_{t+1} = (\nabla R)^{-1}(\nabla R(x_t) - \eta \nabla f_t(x_t))$$
$$x_{t+1} = \Pi_{\mathcal{K}}^{R}(y_{t+1})$$

⇕

Algorithm 2 Online Gradient Descent

```
Input: Initialization u_1 ∈ K' = q^{-1}(K).
for t = 1, ..., T do
    Predict u_t, observe ∇f̃_t(u_t) = ∇f_t(q(u_t))
    Update
        v_{t+1} = u_t − η∇f̃_t(u_t)
        u_{t+1} = Π_{K'}(v_{t+1})
end for
```

$$v_{t+1} = u_t - \eta \nabla \tilde{f}_t(u_t)$$
$$u_{t+1} = \Pi_{\mathcal{K}'}(v_{t+1})$$

Main Theorem

Theorem: Given an instance of **convex OMD** (Alg. 1) which satisfies some assumptions on the smoothness of q, q^{-1}, R, and

$$[\nabla^2 R(x)]^{-1} = J_q(u) J_q(u)^\top ,$$

the **regret** of **Alg. 2** is bounded by $O(T^{2/3})$ by setting $\eta = \Theta(T^{-2/3})$.

Algorithmic Equivalence Analysis

- MD Bregman divergence approximately equivalent to **Euclidean** in reparameterized space

$$D_R(x||y) \approx \frac{1}{2}\|q^{-1}(x) - q^{-1}(y)\|_2^2$$

- The OMD and OGD **iterates are close** after a **single step**:

$$x_t = q(u_t) \Rightarrow \|x_{t+1} - q(u_{t+1})\|_2 = O(\eta^{3/2}).$$

- View the OGD update as a **perturbed** version of OMD, and combine it with the fact that the OMD algorithm can tolerate bounded noise per trial.

Reverse Direction

The other direction from OGD to OMD is even more interesting: given a non-convex OGD, can we show its **global convergence** by showing the **existence** of a convex OMD which corresponds to OGD **implicitly?**

A necessary condition:

- There exists a function q such that $\tilde{f}_t(u)$ can be written as $f_t(q(u))$ where f_t is convex.
- q is a C^3-**diffeomorphism**, and $J_q(u)$ is **diagonal**.
- $q(\mathcal{K}')$ is convex and compact.

Theorem: running Algorithm 2 on loss $\tilde{f}_t(u)$ has regret upper bound $\tilde{O}(T^{\frac{2}{3}})$.

Open Problem

Can this technique get **optimal** $O(\sqrt{T})$ regret bounds? Closeness of MD and GD are not close enough by existing analysis because of **projection**. Tighter analysis may be possible.

References

[1] Ehsan Amid and Manfred Warmuth Reparameterizing mirror descent as gradient descent Neurips 2020

[2] Ehsan Amid and Manfred Warmuth Winnowing with gradient descent COLT 2020