Accelerated Mirror descent methods for overparametrized networks

Kartik Srinivas*
Department of Computer Science
Indian Institute of Technology, Hyderabad
Sangareddy, Telangana
arkartik@student.ubc.ca

Abstract

An analysis of Stochastic Mirror descent, Proximal Mirror descent and their respective accelerated versions on overparametrized networks

Theoretical Results

Mirror Descent

The function $\Phi(x) = \frac{1}{p}\|x\|_p^p$ is strongly convex with a parameter.

The map used in the problem takes the gradients to a different space namely the dual space of the system. The question is, what exactly is the norm that is being used on both of the sides.

The update equation is based on the fenchel conjugate of the function and is as follows

$$\nabla w^*(y) = \text{argmax}_{x \in X} \langle x, y \rangle - w(x) \tag{1}$$

Using this we get

$$\nabla w^*(y) = \text{argmax}_{x \in X} \langle x, y \rangle - \frac{1}{p}\|x\|_p^p$$

$$y = \nabla_x \frac{\sum |x_j^*|^p}{p}$$

$$y_j = |x_j^*|^{p-1}\text{sgn}(x_j)$$

$$x_j^* = |y_j|^{1/(p-1)}\text{sgn}(y_j)$$

Therefore the update step is

$$x_{t+1} = |\nabla w(x_t) - \eta \nabla f(x_t)|^{\frac{1}{p-1}}\text{sgn}(\nabla w(x_t) - \eta \nabla f(x_t)) \tag{2}$$

Proximal Mirror Descent

For the proximal case the only additional update in the equation comes through the regularized norm $\mu\|x\|_1$ The additional gradient is μ sgn (x), which gets added to yield

*Also affiliated with the University of British Columbia, Vancouver at the time of this work

Preprint. Under review.

$$y_j = \text{sgn}(x_j)(\mu + |x_j^*|^{p-1})$$

Hence we get

$$x_j = \text{sgn}(y_j)\max(0, |y_j| - \mu)^{\frac{1}{p-1}}$$

Note how the optimal solution to the problem can be seen component wise, since the function $\frac{1}{p}\|x\|^p$ can be broken component-wise into several parts. The argmax would be a cartesian product of each component wise optimal x_j^*. The stationary point will exist only when $|y_j| \leq \mu$

$$\begin{aligned}
f'(x_j) &= \nabla_{x_j}(x_j y_j - \frac{1}{p}|x_j|^p - |x_j|) \\
&= y_j - |x_j|^{p-1} - \text{sgn}(x_j) \\
&= \begin{cases} < 0 & x_j \leq 0, |y_j| < \mu \\ > 0 & x_j \geq 0, |y_j| < \mu \end{cases} \\
&= 0 \text{ if } |y_j| \geq \mu \text{ and } y_j = \text{sgn}(x_j)(\mu + |x_j|^{p-1})
\end{aligned}$$

In our case the proxy for μ is the multilication of the step size η and the regularization weight λ, i.e $\mu = \lambda\eta$

$$x_{t+1} = \text{sgn}(\nabla w(x_t) - \eta\nabla f(x_t))\max(0, |\nabla w(x_t) - \eta\nabla f(x_t)| - \lambda\eta)^{\frac{1}{p-1}} \tag{3}$$

where $\nabla w(x) = |x|^{p-1}\text{sgn}(x)$