Accelerated Mirror descent methods

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

Dishank[†]
The wonderful world of One piece
The Shakky's Rip-off Bar
Sabaody Archipelago, Grand Line
dishank@onepiece.edu

Abstract

An analysis of Stochastic Mirror descent, Proximal Mirror descent and their respective accelerated versions.

Theoretical Results

Strong convexity of $\frac{1}{p}\|x\|_p^p$

$$\begin{aligned}\Phi(x) &= \frac{1}{p}\|x\|_p^p \\ \nabla_i\Phi(x) &= \nabla_i\frac{\sum|x_j|^p}{p} \\ &= |x_i|^{p-1}\text{sgn}(x_i) \\ \nabla_{ij}\Phi(x) &= \begin{cases}(p-1)|x_i|^{p-2} \text{ if i=j} \\ 0 \text{ o/w}\end{cases}\end{aligned}$$

For $p > 1$, Hessian is positive semi-definite. If $x_i \neq 0\ \forall i$, then none of the eigen-values of the Hessian are 0. So $\Phi(x)$ is strongly convex.

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

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

[*]Also affiliated with the University of British Columbia, Vancouver at the time of this work
[†]Also affiliated to IIT Hyderabad at the time of this work

Preprint. Under review.

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

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

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

Accelerated Proximal Mirror Descent

We can accelerate the Proximal Mirror Descent Algorithm using the Nesterov Trick. Here the weight update equation is

$$w_{t+1} = w_t + \gamma_t \Delta w_{t-1} - \eta_t \nabla \mathcal{L}(w_t + \gamma_t \Delta w_{t-1}) \tag{4}$$

Here, $\Delta w_{t-1} = w_t - w_{t-1}$ and γ_t is the momentum parameter. In Dual space this equation is

$$\nabla \psi(w_{t+1}) = \nabla \psi(w_t) + \gamma_t \Delta z_{t-1} - \eta_t \nabla \mathcal{L}(w_t + \gamma_t \Delta w_{t-1}) \tag{5}$$

Here, $\Delta z_{t-1} = \nabla\psi(w_t) - \nabla\psi(w_{t-1})$ and $\nabla\psi(w_t)$ are the dual variables.

Using previous equations for proximal mirror descent, we can find the new update equation as

$$\begin{aligned} y_t &= \nabla w(x_t) + \gamma(\nabla w(x_t) - \nabla w(x_{t-1})) - \eta\nabla f(x_t + \gamma(x_t - x_{t-1})) \\ x_{t+1} &= \text{sgn}(y_t)\max(0, |y_t| - \lambda\eta)^{\frac{1}{p-1}} \end{aligned}$$

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

We observed that pytorch implementation approximates the Nesterov momentum equations as

$$\begin{aligned} v_{t+1} &= \gamma v_t + \nabla f(x_t) \\ x_{t+1} &= x_t - \eta v_{t+1} \end{aligned}$$

However, we used the original Nesterov equations given as

$$\begin{aligned} v_{t+1} &= \gamma v_t - \eta\nabla f(x_t + \gamma v_t) \\ x_{t+1} &= x_t + v_{t+1} \end{aligned}$$

In dual space, these equations become

$$\begin{aligned} z_{t+1} &= \gamma z_t - \eta\nabla f(x_t + \gamma(x_t - x_{t-1})) \\ y_{t+!} &= y_t + z_{t+1} \\ x_{t+1} &= \text{sgn}(y_{t+1})\max(0, |y_{t+1}| - \lambda\eta)^{\frac{1}{p-1}} \end{aligned}$$

Since we want to compute the gradient at $x_t + \gamma(x_t - x_{t-1})$, we update the weights to $x_t + \gamma(x_t - x_{t-1})$ in practice and only use the actual update at the last update step.

At the first update, we ensure initialisation is such that the first step is same as Proximal Mirror Descent update step (momentum is 0).