Stochastic Mirror Descent

A Brief Overview

Nithin Shrigyan Arsh

Indian Institute of Technology Madras

Group 1

Gradient Descent : a Proximal View

- Gradient Descent Update Rule

$$x_{t+1} = x_t - \eta \cdot \nabla f_t(x_t) \tag{1}$$

- Proximal view

$$x_{t+1} \leftarrow argmin_x \{\eta \cdot \langle \nabla f_t(x_t), x_t \rangle + \frac{1}{2}||x - x_t||^2\} \tag{2}$$

- Taking the gradient and setting to zero

$$\eta \cdot \nabla f_t(x_t) + (x_{t+1} - x_t) = 0 \tag{3}$$

$$\Rightarrow x_{t+1} = x_t - \eta \cdot \nabla f_t(x_t) \tag{4}$$

Interpretation of the Proximal View

$$x_{t+1} \leftarrow argmin_x \{ f_t(x_t) + \eta \cdot \langle \nabla f_t(x_t), x - x_t \rangle + \frac{1}{2} ||x - x_t||^2 \} \quad (5)$$

- The goal is to minimise the function $f_t(x)$ iteratively using small steps.
- So minimise it's linear approximation $f_t(x_t) + \langle \nabla f_t(x_t), x - x_t \rangle$
- Regularise the step size $\frac{1}{2}||x - x_t||^2$ to make the linear approximation valid.

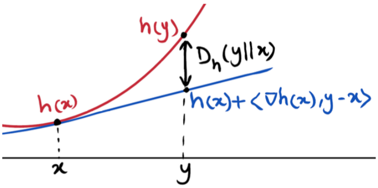
Bregman Divergence

Definition (Bregman Divergence)

Given a strictly convex function $h(\cdot)$,The Bregman divergence from x to y with respect to function $h(\cdot)$ is :

$$D_h(y||x) = h(y) - h(x) - \langle \nabla h(x), y - x \rangle \tag{6}$$

- Interpret $D_h(y||x)$ as the error in the first order approximation.



$D_h(y||x)$ for function $h : R \to R$.[1]

Bregman Divergence examples

1. For $h(x) = \frac{1}{2}||x||^2$ from $R^n \to R$, associated Bregman Divergence is squared Euclidean Distance.

$$D_h(y||x) = \frac{1}{2}||y - x||^2 \tag{7}$$

2. For $h(x) = \sum_{i=1}^{n} (x_i \ln x_i - x_i)$

$$D_h(y||x) = \sum_{i=1}^{n}(y_i \ln \frac{y_i}{x_i} - y_i + x_i) \tag{8}$$

The special case when $\sum_{i=1} y_i = \sum_{i=1} x_i = 1$ gives the Kullback-Leibler (KL) divergence between probability distributions.

Mirror descent Algorithm

$$x_{t+1} \leftarrow argmin_x \{f_t(x_t) + \eta \cdot \langle \nabla f_t(x_t), x - x_t \rangle + D_h(x || x_t)\} \quad (9)$$

- Taking the gradient and setting to zero

$$\eta \cdot \nabla f_t(x_t) + \nabla h(x_{t+1}) - \nabla h(x_t) = 0 \quad (10)$$

$$\nabla h(x_{t+1}) = \nabla h(x_t) - \eta \cdot \nabla f_t(x_t) \quad (11)$$

$$\Rightarrow x_{t+1} = \nabla h^{-1}(\nabla h(x_t) - \eta \cdot \nabla f_t(x_t)) \quad (12)$$

- If $\theta_t = \nabla h(x_t)$ very similar to Gradient Descent.

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla f_t(x_t) \quad (13)$$

Mirror descent Algorithm

$$x_{t+1} \leftarrow argmin_x\{f_t(x_t) + \eta \cdot \langle \nabla f_t(x_t), x - x_t \rangle + D_h(x||x_t)\} \quad (9)$$

- Taking the gradient and setting to zero

$$\eta \cdot \nabla f_t(x_t) + \nabla h(x_{t+1}) - \nabla h(x_t) = 0 \quad (10)$$

$$\nabla h(x_{t+1}) = \nabla h(x_t) - \eta \cdot \nabla f_t(x_t) \quad (11)$$

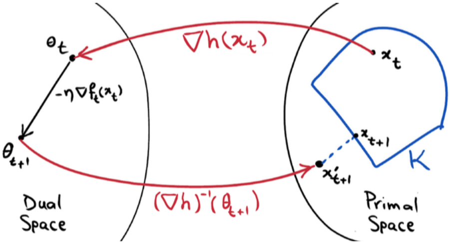
$$\Rightarrow x_{t+1} = \nabla h^{-1}(\nabla h(x_t) - \eta \cdot \nabla f_t(x_t)) \quad (12)$$

- If $\theta_t = \nabla h(x_t)$ very similar to Gradient Descent.

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla f_t(x_t) \quad (13)$$

Is there a connection ? Yes !

Mirror Descent : Mirror Map View



The four basic steps in each iteration of the mirror descent algorithm

- The Dual Space acts as a Mirror Image of the Primal Space[2].

Connection to Preconditioned Gradient Descent

- Mirror Descent Update Rule :

$$x_{t+1} = \nabla h^{-1}(\nabla h(x_t) - \eta \cdot \nabla f_t(x_t)) \tag{14}$$

$$x_{t+1} = x_t - \nabla h^{-1}(\eta \cdot \nabla f_t(x_t)) \tag{15}$$

- Very similar to Newton Method if Hessian $H_f = \nabla h$.

$$x_{t+1} = x_t - H_f^{-1}(\eta \cdot \nabla f_t(x_t)) \tag{16}$$

- We trade-off between robustness and rate of convergence.

Application to Deep Learning

Problem Setup:

- For highly over parameterized Neural Networks: number of parameters $\gg$ number of training data points
- Training loss may have infinitely many global minima!
- The interpolating solution we converge depends on the initialization point and the algorithm used.
- From these multiple solutions, we want to find the solution that performs the best on unseen data.

Application to Deep Learning

Problem Setup:

- For highly over parameterized Neural Networks: number of parameters $\gg$ number of training data points
- Training loss may have infinitely many global minima!
- The interpolating solution we converge depends on the initialization point and the algorithm used.
- From these multiple solutions, we want to find the solution that performs the best on unseen data.

How does Stochastic Mirror Descent (SMD) help?

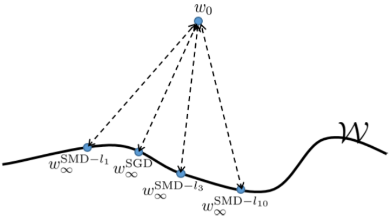
- Helps us compare generalization performance using different potential functions.

Implicit Regularisation of SMD[3]

Theorem

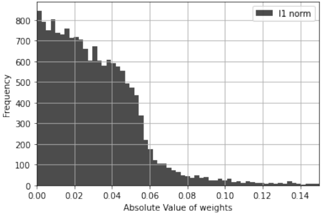
For highly overparameterized nonlinear models, under reasonable assumptions:

1. *the SMD algorithm for any particular potential converges to the global minimum.*
2. *the global minimum obtained by SMD is approximately the closest one to the initialization in the Bregman divergence corresponding to the potential.*

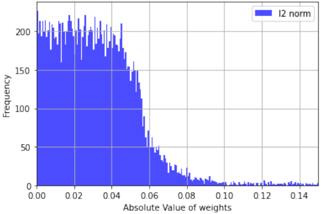


Multiple Global Minima

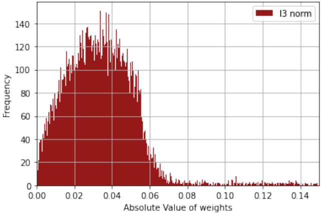
Performance on MNIST using Vanilla CNN



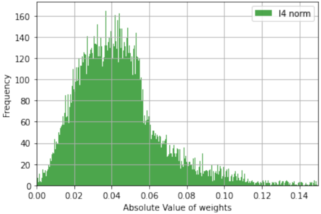
l_1 Test acc. $= 79\%$



l_2 Test acc. $= 80\%$



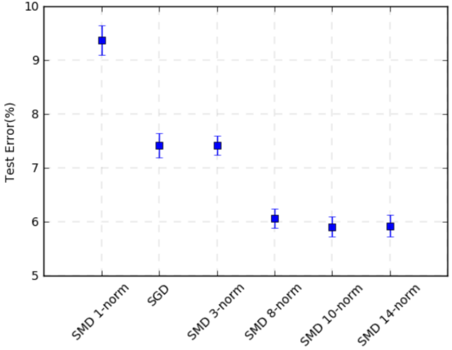
l_3 Test acc. $= 83\%$



l_4 Test acc. $= 82\%$

l_3-norm is the best choice of Bregman Potential for this example.

Performamce on CIFAR-10 using ResNet-18



Test Error of SMD under differnt norms

- l_{10} ans l_{14} norms generalise the best.
- Choice of Bregman Potential clearly is problem dependent.

Take Away Points

- The Learning Problem is not equivalent to just optimizing training loss but we need to arrive at a solution that generalises well to unseen data.

Take Away Points

- The Learning Problem is not equivalent to just optimizing training loss but we need to arrive at a solution that generalises well to unseen data.
- SMD converges to a global minimum, which is approximately the closest to the initialization in Bregman divergence sense, under reasonable assumptions.

Take Away Points

- The Learning Problem is not equivalent to just optimizing training loss but we need to arrive at a solution that generalises well to unseen data.
- SMD converges to a global minimum, which is approximately the closest to the initialization in Bregman divergence sense, under reasonable assumptions.
- Choosing a suitable Bregman Divergence will help our model generalise better.

Take Away Points

- The Learning Problem is not equivalent to just optimizing training loss but we need to arrive at a solution that generalises well to unseen data.
- SMD converges to a global minimum, which is approximately the closest to the initialization in Bregman divergence sense, under reasonable assumptions.
- Choosing a suitable Bregman Divergence will help our model generalise better.
- Domain Knowledge may help in choosing the right Bregman Potential for our Optimisation Problem.

  Eg: One might choose l_1-norm for Compressed Sensing Applications.

Open Problems to think about

- What is the right choice of Bregman Potential
  - Is it dependent on architecture of Neural Network?
  - Is it dependent on the training data ?

[1] Y. Li and Y. Liang, "Learning overparameterized neural networks via stochastic gradient descent on structured data," in Proc. Adv. Neural Inf.Process. Syst., 2018,

Open Problems to think about

- What is the right choice of Bregman Potential
  - Is it dependent on architecture of Neural Network?
  - Is it dependent on the training data ?
- Why do over parameterised Neural Networks generalise better?[1]
  - Contradicts our ideas about over-fitting.

[1] Y. Li and Y. Liang, "Learning overparameterized neural networks via stochastic gradient descent on structured data," in Proc. Adv. Neural Inf.Process. Syst., 2018,

References

[1] Anupam Gupta. "15-850: Advanced Algorithms, Fall 2020 Notes". In: CMU (2020).

[2] A. Nemirovski and D. Yudin. "Problem Complexity and Method Efficiency in Optimization.". In: John Wiley, New York. (1983).

[3] Navid Azizan Ruhi, Sahin Lale, and Babak Hassibi. "Stochastic Mirror Descent on Overparameterized Nonlinear Models: Convergence, Implicit Regularization, and Generalization". In: *CoRR* abs/1906.03830 (2019). arXiv: 1906.03830.