CVXOPT II: Pset 3

Justin Lewis

February 2019

1 Problem 1

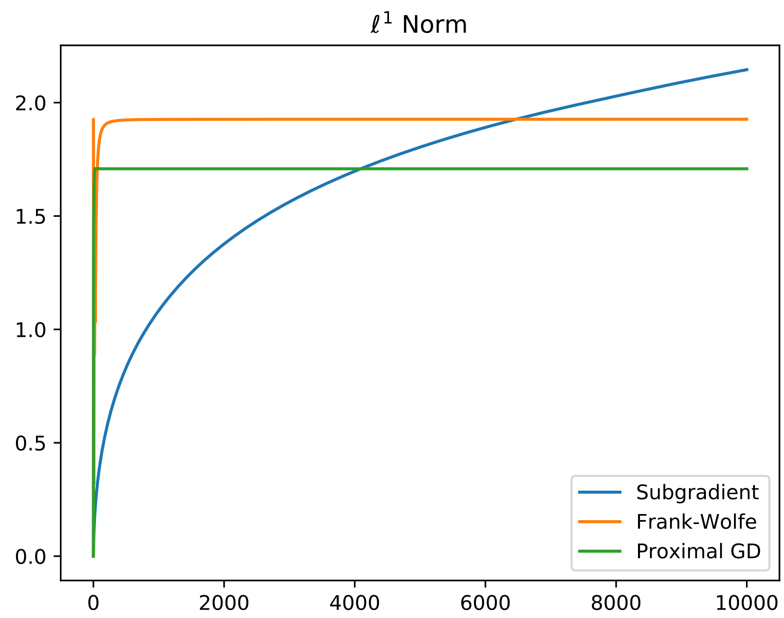

Figure 1: ℓ_1 norm vs iterations for solving LASSO

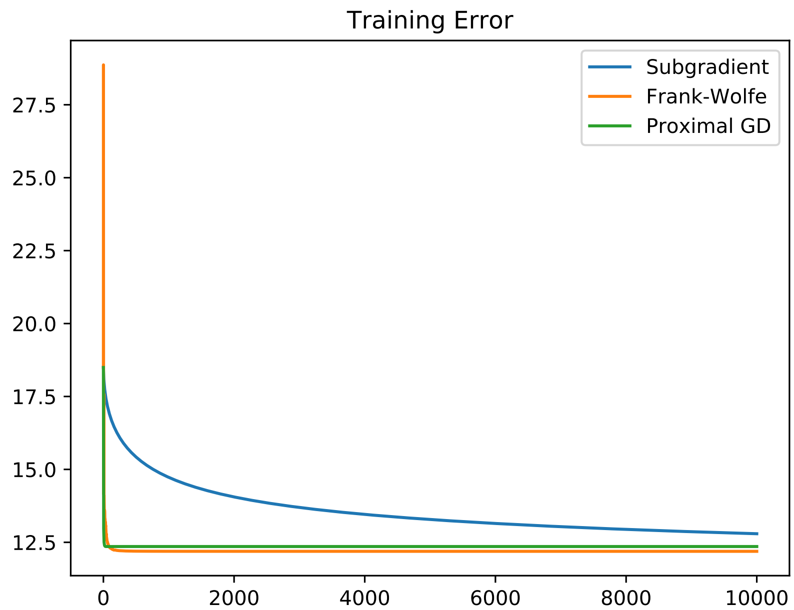

Figure 2: Training error vs iterations for solving LASSO

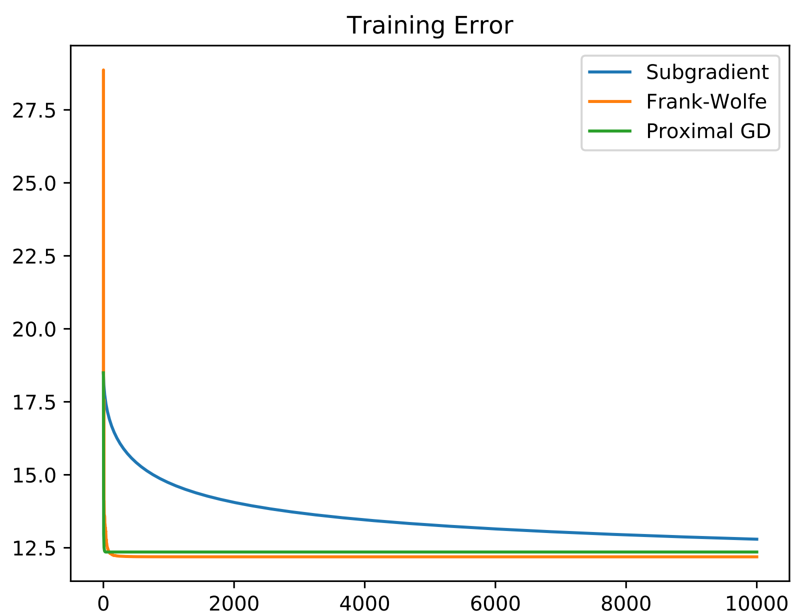

Figure 3: Test error vs iterations for solving LASSO

2 Problem 2

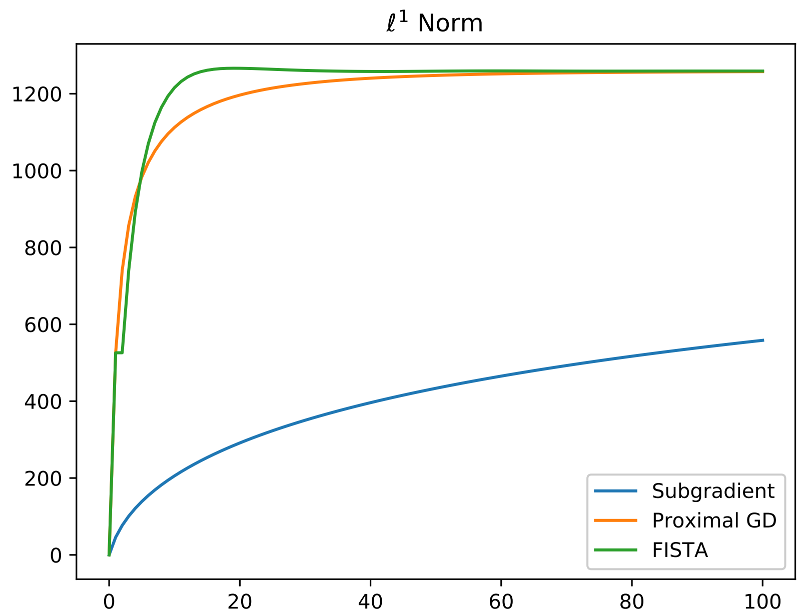

Figure 4: ℓ_1 norm vs iterations. The difference between FISTA and ProxGrad was only meaningful in the first 100 iterations

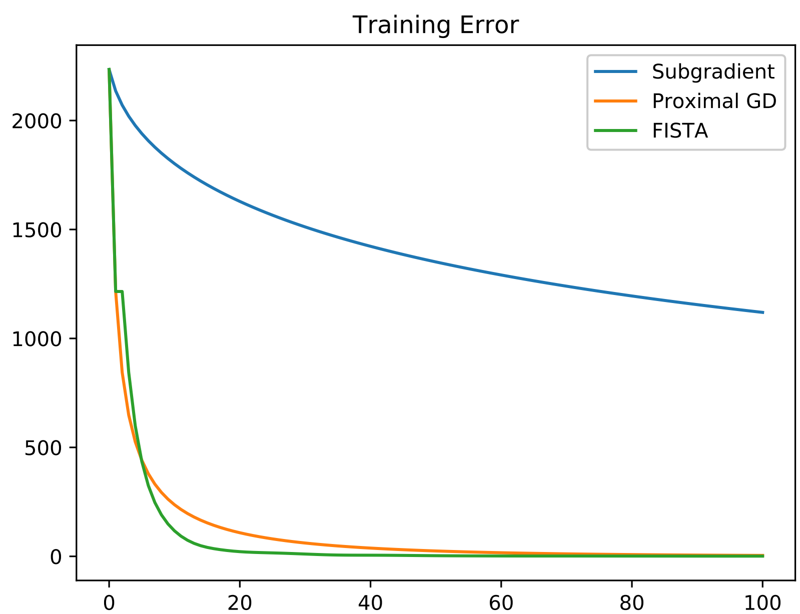

Figure 5: Training vs iterations. The difference between FISTA and ProxGrad was only meaningful in the first 100 iterations

3 Problem 3

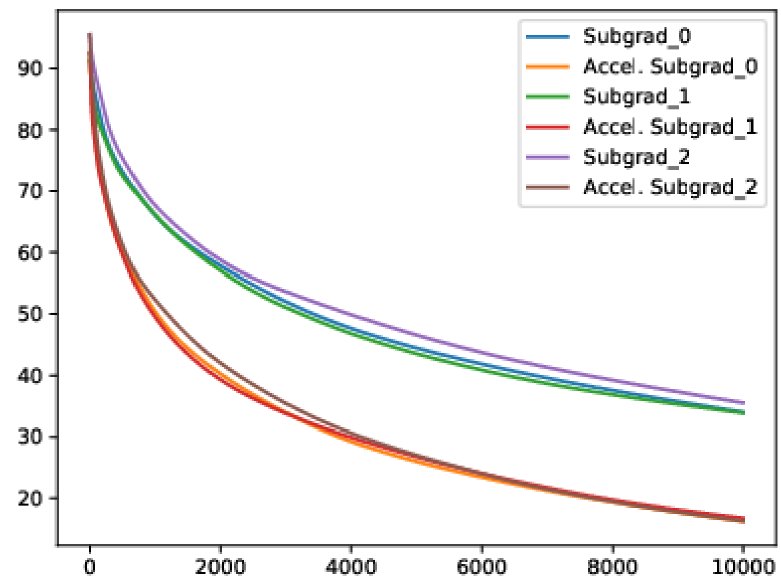

Figure 6: Training Error vs iterations. Acceleration improved rate of convergence on average.

4 Problem 4

4.1 Part a)

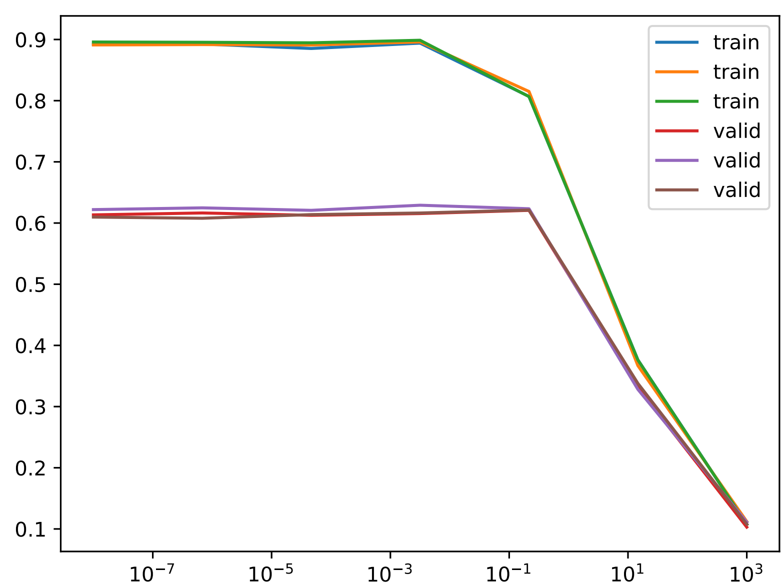

Figure 7: Train + Test performance vs. value of parameter μ for Vanilla gradient descent. Validation curves show test performance.

Optimal μ value: between 10^{-3} to 10^{-2}. Although value less than 10^{-2} is essentially the same

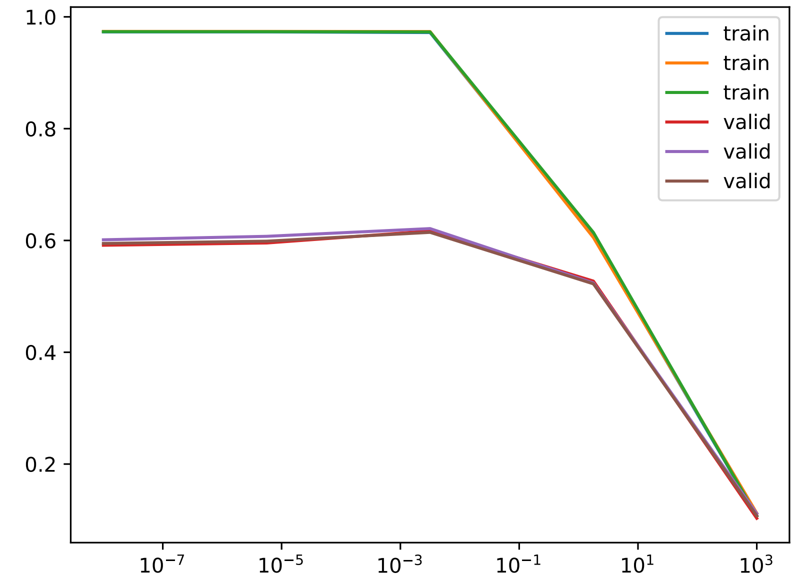

Figure 8: Train + Test performance vs. value of parameter μ for Nesterov accelerated gradient descent. Validation curves show test performance.

4.2 Part b)

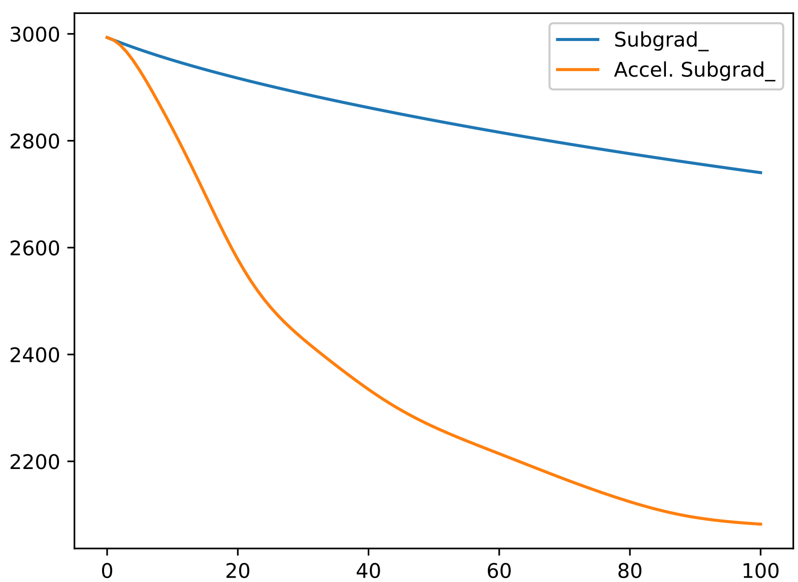

Figure 9: Figure demonstrating descent using subgradient vs accelerated subgradient descent.

4.3 Part c)

Clearly, as μ is increased, the importance of the regularizing term begins to dominate. This lead to poor performance on fitting the dataset. As the parameter goes to zero, accelerated gradient descent seems to have a slight benefit in fitting the data (0.975 accuracy vs 0.9). But, that said, I did not notice a significant difference in the final performance between the two methods as a function of μ.

4.4 Part d)

The findings aside, I would have independently believed (from what I understand about accelerated gradient descent), that small μ would lead to a benefit for acceleration. This is because for small μ, the condition number of the problem $\frac{\alpha}{\beta}$ would be closer to zero than for large μ (as the problem is μ strongly convex). Thus, the loss surface would potentially have a long narrow valley which gradient descent does not handle well. In contrast, accelerated GD utilizes its momentum to take much larger step sizes once inside such valleys.