CAP 6617 Advanced Machine Learning, Fall 2022

Homework 2

Due 10/6/2022 11:59PM

1. *Proximal (stochastic) subgradient method?* Consider the problem of minimizing $L(\boldsymbol{\theta}) + \lambda r(\boldsymbol{\theta})$ where both L and r are nonsmooth, but r admits an efficient proximal operator. Instead of simply taking the subgradient of the combined function $L + \lambda r$, one may be tempted to try the "proximal subgradient method" to utilize the efficient proximal operator of r:

$$\begin{aligned} &\text{obtain} \quad \boldsymbol{g}^{(t)} \in \partial L(\boldsymbol{\theta}^{(t)}) \\ &\boldsymbol{\theta}^{(t+1)} \leftarrow \operatorname{Prox}_{\gamma^{(t)}\lambda r}\left(\boldsymbol{\theta}^{(t)} - \gamma^{(t)}\boldsymbol{g}^{(t)}\right) \end{aligned}$$

In this exercise we show that this does not hurt the theoretical convergence of the subgradient method. In fact it requires a slightly milder condition that all subgradients of L are bounded, not those of the combined function $L + \lambda r$ if we were to use plain subgradient method.

(a) Rewrite the algorithm as

$$\begin{cases} \boldsymbol{\theta}^{(t+1)} \leftarrow \operatorname{Prox}_{\gamma^{(t)}\lambda r}\left(\widetilde{\boldsymbol{\theta}}^{(t)}\right) \\ \widetilde{\boldsymbol{\theta}}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t+1)} - \gamma^{(t)}\boldsymbol{g}^{(t+1)} \end{cases}$$

Show that

$$L(\boldsymbol{\theta}^{(t+1)}) + \lambda r(\boldsymbol{\theta}^{(t+1)}) - L(\boldsymbol{\theta}) - \lambda r(\boldsymbol{\theta}) \le \frac{1}{\gamma^{(t)}}(\widetilde{\boldsymbol{\theta}}^{(t+1)} - \widetilde{\boldsymbol{\theta}}^{(t)})^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t+1)}).$$

(b) We then add and subtract the same term on the right-hand-side to get

$$\begin{aligned} L(\boldsymbol{\theta}^{(t+1)}) + \lambda r(\boldsymbol{\theta}^{(t+1)}) - L(\boldsymbol{\theta}) - \lambda r(\boldsymbol{\theta}) &\le \frac{1}{\gamma^{(t)}}(\widetilde{\boldsymbol{\theta}}^{(t+1)} - \widetilde{\boldsymbol{\theta}}^{(t)})^\top(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}^{(t+1)}) - (\widetilde{\boldsymbol{\theta}}^{(t+1)} - \widetilde{\boldsymbol{\theta}}^{(t)})^\top\boldsymbol{g}^{(t+1)} \\ &\le \frac{1}{2\gamma^{(t)}}\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}^{(t)}\|^2 - \frac{1}{2\gamma^{(t)}}\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}^{(t+1)}\|^2 - \frac{1}{2\gamma^{(t)}}\|\widetilde{\boldsymbol{\theta}}^{(t+1)} - \widetilde{\boldsymbol{\theta}}^{(t)}\|^2 \\ &\quad - (\widetilde{\boldsymbol{\theta}}^{(t+1)} - \widetilde{\boldsymbol{\theta}}^{(t)})^\top\boldsymbol{g}^{(t+1)} \\ &= \frac{1}{2\gamma^{(t)}}\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}^{(t)}\|^2 - \frac{1}{2\gamma^{(t)}}\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}^{(t+1)}\|^2 \\ &\quad - \frac{1}{2\gamma^{(t)}}(\widetilde{\boldsymbol{\theta}}^{(t+1)} - \widetilde{\boldsymbol{\theta}}^{(t)})^\top(\widetilde{\boldsymbol{\theta}}^{(t+1)} - \widetilde{\boldsymbol{\theta}}^{(t)} + 2\gamma^{(t)}\boldsymbol{g}^{(t+1)}). \end{aligned}$$

Show that

$$(\widetilde{\boldsymbol{\theta}}^{(t+1)} - \widetilde{\boldsymbol{\theta}}^{(t)})^\top(\widetilde{\boldsymbol{\theta}}^{(t+1)} - \widetilde{\boldsymbol{\theta}}^{(t)} + 2\gamma^{(t)}\boldsymbol{g}^{(t+1)}) = \|\boldsymbol{\theta}^{(t+1)} - \widetilde{\boldsymbol{\theta}}^{(t)}\|^2 - \gamma^{(t)2}\|\boldsymbol{g}^{(t+1)}\|^2 \ge -\gamma^{(t)2}\|\boldsymbol{g}^{(t+1)}\|^2.$$

(You only need to show this inequality. All the steps before that are provided for you as part of the overall proof, which you don't need to show.)

(c) Conclude that

$$2\lambda^{(t)} \left(L(\boldsymbol{\theta}^{(t+1)}) + \lambda r(\boldsymbol{\theta}^{(t+1)}) - L(\boldsymbol{\theta}) - \lambda r(\boldsymbol{\theta}) \right) \leq \|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}^{(t)}\|^2 - \|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}^{(t+1)}\|^2 + \gamma^{(t)2}\|\boldsymbol{g}^{(t+1)}\|^2.$$

We can then follow the steps on page 25 of `lec3.pdf` to prove convergence, assuming $\|\boldsymbol{g}^{(t+1)}\|^2 \leq G$ for all t.

(d) Now consider $\boldsymbol{g}^{(t+1)}$ being a stochastic subgradient at $\boldsymbol{\theta}^{(t+1)}$ satisfying $\mathrm{E}\, \boldsymbol{g}^{(t+1)} \in \partial L(\boldsymbol{\theta}^{(t+1)})$, where the expectation is conditioned on $\boldsymbol{\theta}^{(t+1)}$, outline how to show its expected convergence.

2. *Hand-written digits classification.* The MNIST data set is a famous data set for multi-class classification, which can be downloaded here `http://yann.lecun.com/exdb/mnist/`. In this question you will implement various algorithms for multi-class logistic regression with quadratic regularization that solves the following optimization problem

$$\underset{\boldsymbol{\Theta}}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^{n}\left(\log\left(\sum_{c=1}^{k}\exp(\boldsymbol{x}_i^\top\boldsymbol{\theta}_c)\right) - \boldsymbol{x}_i^\top\boldsymbol{\theta}_{y_i}\right) + \lambda\|\boldsymbol{\Theta}\|^2.$$

Here we simply assume that the features are the image pixels themselves (we even ignore the constant 1 here).

(a) Derive the gradient descent (GD), stochastic gradient descent (SGD), and (Nesterov's) accelerated gradient descent (AGD) algorithm for solving it. At iteration t, you can simply denote the step size as $\gamma^{(t)}$ (and similarly for the extrapolation parameter $\delta^{(t)}$ in AGD).

(b) Implement the algorithm in your favorite programming language.

(c) Run the algorithms on the training set of MNIST for two scenarios: weakly convex $\lambda = 0$ and strongly convex $\lambda = 0.1$. Use a constant step size $\gamma^{(t)} = 0.001, 0.01, 0.1, 1$ and report the best result of each algorithm on a figure with horizontal axis the number of full gradient evaluations and vertical axis the prediction error rate on the test set. Note that both GD and AGD require a full gradient evaluation in each iteration, while SGD can run n iterations with the same complexity of evaluating a full gradient.