

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 :

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

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

Show that

$$L(\boldsymbol{\theta}^{(t+1)}) + \lambda r(\boldsymbol{\theta}^{(t+1)}) - L(\boldsymbol{\theta}) - \lambda r(\boldsymbol{\theta})$$

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

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

Show that

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

(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

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

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

- (d) Now consider $\mathbf{g}^{(t+1)}$ being a stochastic subgradient at $\boldsymbol{\theta}^{(t+1)}$ satisfying $\mathbb{E} \mathbf{g}^{(t+1)} \in \partial L(\boldsymbol{\theta}^{(t+1)})$, 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

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

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 algorithm with $\lambda = 0$ and $\lambda = 0.001$ and report the best result of each algorithm on the test set. Note that both GD and AGD require a full gradient evaluation in each iteration, while SGD can run with the same complexity of evaluating a full gradient.