Final project of Convex Optimization

Instructor: Bao Chenglong
YMSC, Tsinghua University

1 Large-scale Optimal Transport

Acknowledgement: This material is based on the "Convex optimization" taught by Prof. Zaiwen Wen (Peking University). Notations: $X \in \mathbf{R}^{m\times n}$, $\mathbf{1} = (1, 1, \dots, 1)^\top$.

1. Consider the standard form of LP

$$\min_X \langle C, X\rangle, \quad \text{s.t. } X\mathbf{1} = \mu, X^\top \mathbf{1} = \nu, X \geq 0, \tag{1}$$

   where C is cost matrix, μ and ν are two marginal distributions.

   (a) Solving (1) by calling mosek and gurobi directly in Matlab or python. The package "CVX" is not allowed to use here. Compare the performance between the simplex methods and interior point methods.

   (b) Write down and implement a first-order method (e.g. the accelerated proximal gradient method, the alternating direction method of multipliers).

   (c) Test problems:

   - Generate some random data C, μ and ν with different settings of m, n.
   - Find or construct the data sets in the references:
     * Jörn Schrieber, Dominic Schuhmacher, Carsten Gottschilich, DOTmark – A benchmark for Discrete Optimal Transport.
     * Samuel Gerber, Mauro Maggioni, Multiscale Strategies for Computing Optimal Transport.
     * Option: computing the distribution translation (The data can be downloaded from "Web learning" platform).

2. Read the reference:

   - Gabriel Peyre, Marco Curturri, Computational Optimal Transport, `https://arxiv.org/abs/1803/00567`.
     some slides on optimal transport can be found at `https://optimaltransport.github.io/slides/`
   - Ernest K. Ryu, Yongxin Chen, Wuchen Li, Stanley Osher, Vector and Matrix Optimal Mass Transport: Theory, Algorithm, and Applications, `https://arxiv.org/abs/1712.10279`.

(a) Find one of the most important optimization problem from the above references. Write down the background and formulation clearly.

(b) Write and implement an algorithm for the optimization problem in 2(a) from the chosen reference. Try to reproduce the numerical results in that reference.

(c) Try to write down and implement an algorithm covered in this course for the optimization problem in 2(a). This algorithm should be different from the one in 2(b).

3. Requirement:

   (a) Compare the efficiency (CPU time) and accuracy (checking the optimality condition) of different methods.

   (b) Prepare a report including:

      - detailed answers to each question
      - numerical results and their interpretation

   (c) Pack all of your codes in one file named as "project1-name-ID.zip" and send it to both TA and me via tuyouhuathu@163.com.

   (d) If you get significant help from others on one routine, write down the source of references at the beginning of this routine.

   (e) You can do it either individually or in groups of two and can combine your team project. Please indicate each person's contribution in the report.

4. Some references:

   - Yujia Xie, Xiangfeng Wang, Ruijia Wang, Hongyuan Zha, A Fast Proximal Point Method for Computing Exact Wasserstein Distance, `http://auai.org/uai2019/proceedings/papers/158.pdf`.
   - Matt Jacobs, Flaview Léger, A Fast Approach to Optimal Transport: the Back-and-Forth Method, `https://arxiv.org/pdf/1905.12154.pdf`.

2 Algorithm and analysis for shallow neural network

Acknowledgment: this is based on the " Convex optimization" taught by Prof. Zaiwen Wen (Peking University). Let $f(\theta; x_i)$ be the output of a neural network and $\sigma(x)$ be one of the following element-wise activation function

$$\begin{aligned}
\mathrm{Relu}(x) &= \max(0, x),\\
\mathrm{Sigmoid}(x) &= \frac{1}{1+\exp(-x)},\\
\mathrm{Tanh}(x) &= \frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}.
\end{aligned}$$

For a given dataset $\{(x_i, y_i)\}_{i=1}^N$, consider the supervised learning problem:

$$\min_{\theta} \quad h(\theta) = \sum_{i=1}^{N} L(y_i, f(\theta; x_i)), \tag{2}$$

where L is the ℓ_2 loss function, i.e. $L(x, y) = \frac{1}{2}\|x - y\|_2^2$. In the following questions, the activation and loss function can be any of their combinations or other activation function (e.g. leaky Relu, polynomial).

1. Consider the two-layer feed-forward neural network

$$f(\theta, x) = f_{W,v}(x) = \frac{1}{\sqrt{m}} v^\top \sigma(Wx),$$

where x is the input data, $W \in \mathbf{R}^{m\times d}$, $v \in \mathbf{R}^m$ are weight matrices. The parameters are concatenated into one column vector $\theta = [\vec{(W)}^\top, v^\top]^\top$, where $\vec{(W)}$ transforms the matrix W into a vector by stacking all the columns of W.

(a) Compute the gradient (or subgradient) of $h(\theta)$ with respect to θ and estimate its Lipschitz constant.

(b) Compute the Hessian of $h(\theta)$ with respect to θ if it is available and estimate the upper and lower bound of its eigenvalues.

(c) Will the function $h(\theta)$ be strongly convex in a small neighborhood of the global optimal solution? Either try numerical experiments on a few small examples or try to establish certain theoretical results.

(d) Suppose that **the gradient descent method** with certain line search schemes or certain stepsize strategies is applied to solve (2). Write down the method and the corresponding convergence results from a few literature or the classic textbooks on nonlinear progamming such as

- Numerical Optimization, Jorge Nocedal and Stephen Wright, Springer
- Optimization Theory and Methods, Wenyu Sun, Ya-Xiang Yuan

Is it possible that this method converges to a global optimal solution of (2)? Either try numerical experiments on a few examples or try to establish certain theoretical results.

(e) Suppose that the **stochastic gradient method** is applied to solve (2). Write down the method and the corresponding convergence results from a few literature. Is it possible that this method converges to a global optimal solution of (2)? Either try numerical experiments on a few examples or try to establish certain theoretical results.

(f) Suppose that **KFAC** method is applied to solve (2).

- Optimizing Neural Networks with Kronecker-factored Approximate Curvature, James Martens, Roger Grosse. `https://arxiv.org/abs/1503.05671`.

Write down the method and the corresponding convergence results from a few literature. Is it possible that this method converges to a global optimal solution of (2)? Either try numerical experiments on a few examples or try to establish certain theoretical results.

2. Requirement:

   (a) Test problems:

   - random examples created by yourself.
   - MNIST or CIFAR-10.
   - You can consider the experiment setup in the following paper: Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks, Arora Sanjeev, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang, `https://arxiv.org/abs/1901.08584`.

   (b) Requirements on numerical experiments:

   - Specify the weight initialization schemes chosen in your experiments.
   - Write all the prerequisites and the usage of your codes in detail.

   (c) Prepare a report including:

   - detailed answers to each question
   - numerical results and their interpretation if there are numerical experiments.

   (d) Pack all of your codes in one file named as "project2-name-ID.zip" and send it to both TA and me via: tuyouhuathu@163.com.

   (e) If you get significant help from others on one routine, write down the source of references at the beginning of this routine.

   (f) You can do it either individually or in groups of two and can combine your team project. Please indicate each person's contribution in the report.

3 Literature review

We provide two topics that are not covered in the lectures, and the literature review should involve in-depth summaries.

1. Dive in to the Nesterov's acceleration.

   - Weijie Su, Stephen Boyd, Emmanuel Candès. A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights, `https://arxiv.org/pdf/1503.01243.pdf`.
   - Brendan O'Donoghue, Emmanuel Candès. Adaptive Restart for Accelerated Gradient Schemes, `https://arxiv.org/abs/1204.3982`.
   - Sebastien Bubeck, Yin Tat Lee, Mohit Singh, A geometric alternative to Nesterov's Accelerated Gradient Descent, `https://arxiv.org/abs/1506.08187`.

2. Extensions to the Non-convex optimization.

   (a) Mingyi Hong, Zhi-quan Luo, Meisam Razaviyayn, Convergence Analysis of Alternating Direction Method of Multipliers for a Family of Nonconvex Problems, https://arxiv.org/abs/1410.1390.

   (b) Saeed Ghadimi, Guanghui Lan, Accelerated Gradient Methods for Nonconvex Nonlinear and Stochastic Programming, https://arxiv.org/abs/1310.3787.

   (c) Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Stephen Boyd, Peter W. Glynn, On the Convergence of Mirror Descent Beyond Stochastic Programming, https://arxiv.org/abs/1706.05681.

3. Requirements:

   - Choose Part 1 or one paper from the paper list in Part 2.
   - Please submit a short report (no more than 1 page) (Due. 15 May) stating the papers you plan to survey. Describe why they are important or interesting, and provide some appropriate references.
   - Finish the report individually and summarize your findings. Send the PDF file to both the TA and me.
   - The length of the report is up to 6 pages with unlimited appendix.
   - Option: you are encouraged the algorithm to other settings: connection between optimization algorithms and ordinary differential equations, asynchronization, ADMM on manifolds, etc.