ELE 539 / COS 512: Homework 4

due on April. 22, 2021 (11:59 PM Blackboard)

1 Noisy Power Method

Consider the noisy version of power method:

Algorithm 1 NOISY POWER METHOD

1: **for** $t = 1, \cdots, T-1$ **do**
2: $\quad \tilde{x}_{t+1} = Ax_t + \zeta_t$.
3: $\quad x_{t+1} = \tilde{x}_{t+1} / \|\tilde{x}_{t+1}\|$.
4: **Output:** $\mathbf{x}_T$.

where noise ζ_t can come from variant sources such as noisy observations of matrix A or inexact matrix vector computation. In this problem, we prove the following theorem:

Theorem 1. *For any p.s.d matrix A with top two eigen-pairs $(\lambda_1, v_1), (\lambda_2, v_2)$, where $\lambda_1 > \lambda_2$, denote $\theta_t := \arccos(|\langle v_1, x_t \rangle|)$. Suppose that for any t, the noise ζ_t satisfies $\|\zeta_t\| \leq \epsilon(\lambda_1 - \lambda_2)/2$ for some small $\epsilon \leq 1/2$. Also assume the initial condition $\cos\theta_1 \geq \epsilon$. Then, we have $\tan\theta_T \leq 3\epsilon$ for any*

$$T \geq \Omega\left(\frac{\lambda_1}{\lambda_1 - \lambda_2} \log \frac{1}{\epsilon}\right)$$

(a) [2 point] Prove that, if we have $\|\zeta_t\| \leq \min\{\epsilon, \cos\theta_t\} \cdot (\lambda_1 - \lambda_2)/2$, then:

$$\tan\theta_{t+1} \leq \left(1 - \frac{\lambda_1 - \lambda_2}{2(\lambda_1 + \lambda_2)}\right) \tan\theta_t + \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2} \cdot \epsilon.$$

(b) [1 point] Prove that when $\cos\theta_t \geq \epsilon$, we have $\cos\theta_{t+1} \geq \epsilon$.

(c) [2 point] Use above results to prove Theorem 1.

2 Gap-free Results for Lanczos Algorithm

In this question, we prove the following theorem:

Theorem 2. *For any p.s.d matrix A with top eigen-pair (λ_1, v_1), let x_1 be the initialization of Lanczos algorithm, and x_t be the output after t iterations. Then we have $x_t^\top A x_t \geq (1-\epsilon)\lambda_1$ for any*

$$t \geq \Omega\left(\frac{1}{\sqrt{\epsilon}} \log \frac{1}{\epsilon \langle v_1, x_1 \rangle^2}\right)$$

(a) [2 point] Prove that

$$x_t^\top A x_t \geq \left(1 - \frac{\epsilon}{2}\right) \lambda_1 \cdot \max_{p \in \mathcal{P}_{t-1}} \left[1 - \frac{\max_{i^\star \leq i \leq d} p^2(\lambda_i)}{\langle v_1, x_1 \rangle^2 \cdot p^2(\lambda_1)}\right]$$

where $\mathcal{P}_t$ is the set of polynomials with degree at most t, and $i^\star = \min\{i \in [d] \mid \lambda_i \leq (1 - \epsilon/2)\lambda_1\}$.

(b) [2 point] Complete the proof using the properties of the Chebyshev polynomial.

3 GD + Lanczos for Finding Second-order Stationary Point

Consider the following algorithm, which is same as GD + power method algorithm introduced in the lecture except here we replace power method with Lanczos algorithm.

Algorithm 2 GD + LANCZOS

> **for** $t = 1, \cdots$ **do**
>   **if** $\|\nabla f(x_t)\| \geq \epsilon_g$ **then**
> $x_{t+1} \leftarrow x_t - \eta \nabla f(x_t)$.
>   **else**
> $v_t \leftarrow \text{LANCZOS}(I - \eta \nabla^2 f(x_t), \mathcal{T})$.
> $x_{t+1} = \operatorname{argmin}_{x \in \{x_t + \lambda v_t, x_t - \lambda v_t\}} f(x)$.

where $\text{LANCZOS}(A, T)$ is a subroutine that computes the approximate top eigenvector of matrix A by running Lanczos algorithm for T iterations. We prove the following theorem.

Theorem 3. *Assume f is ℓ-gradient Lipschitz, ρ-Hessian Lipschitz. For any $\epsilon_g, \epsilon_H, \delta \geq 0$, with $\eta = 1/\ell$, and appropriate choice of hyperparameters $\mathcal{T}$, λ, with probability $1 - \delta$, with no more than*

$$\tilde{\mathcal{O}}\left((f(x_0) - f(x^\star)) \cdot \left(\frac{\ell}{\epsilon_g^2} + \frac{\rho^2 \sqrt{\ell}}{\epsilon_H^{3.5}}\right)\right)$$

gradient or Hessian-vector queries, at least one of the iterate x_t will be (ϵ_g, ϵ_H)-SOSP in the sense:

$$\|\nabla f(x_t)\| \leq \epsilon_g, \qquad \nabla^2 f(x_t) \succeq -\epsilon_H \cdot I$$

We remark that the randomness in Theorem 3 is over the random initializations for Lanczos subroutines. For GD + power method, one can show that the gradient/Hessian-vector complexity is $\tilde{\mathcal{O}}(\epsilon_g^{-2} + \epsilon_H^{-4})$. Thus GD + Lanczos improves over GD + power method on ϵ_H dependency.

(a) [1 point] In case $\|\nabla f(x_t)\| \geq \epsilon_g$, bound the function decrease $f(x_{t+1}) - f(x_t)$.

(b) [2 point] In case $\|\nabla f(x_t)\| < \epsilon_g$ but $\lambda_{\min}(\nabla^2 f(x_t)) \leq -\epsilon_H$, choose proper $\mathcal{T}, \lambda$, and bound the function decrease $f(x_{t+1}) - f(x_t)$. [Hint: you can use the gap-free result of Lanczos in Question 2.]

(c) [1 point] Combine above results to prove Theorem 3.