DS 801: Advanced Optimization for Data Science **KAIST, Fall 2024**

Lecture #13: Algorithms for Finding Second-Order Stationary Points April 8, 2024

Lecturer: Dabeen Lee

1 Outline

In this lecture, we study

- the cubic regularization method,
- perturbed gradient descent,
- problems where a second-order stationary point is sufficient.

2 More Algorithms for Finding Second-Order Stationary Points

Recall that the Hessian of f is γ-Lipshitz continuous if

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \le \gamma \|x - y\|_2$$

where $\|\nabla^2 f(x) - \nabla^2 f(y)\|_2$ denotes the spectral norm of $\nabla^2 f(x) - \nabla^2 f(y)$ and the spectral norm of a matrix is its largest singular value. Remember that we defined an (ϵ, δ)-SOSP as a point x such that

$$\|\nabla f(x)\|_2 \le \epsilon \quad \text{and} \quad \nabla^2 f(x) \succeq -\delta I.$$

We learned an algorithm for finding an (ϵ, δ)-SOSP after a bounded number of iterations for a funcion f whose Hessian is γ-Lipschitz continuous. The algorithm applies gradient descent when the gradient norm is high, and if the Hessian has a sufficiently negative eigenvalue, then we take a descent step toward the associated eigenvector. The algorithm works well in practice, but as the algorithm relies on the power method, it sometimes suffers from noise accumulation. In this section, we provide two more algorithms for computing second-order stationary points.

2.1 Cubic Regularization

In this section, we discuss an algorithm referred to as the **cubic regularization method** due to Nesterov and Polyak [NP06]. Given a funcion f whose Hessian is γ-Lipschitz continuous, we define an **ϵ-SOSP** as a point x such that

$$\|\nabla f(x)\|_2 \le \epsilon \quad \text{and} \quad \nabla^2 f(x) \succeq -\sqrt{\gamma \epsilon} I.$$

By definition, an ϵ-SOSP is an $(\epsilon, \sqrt{\gamma\epsilon})$-SOSP.

Theorem 13.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function whose Hessian is γ-Lipschitz continuous. Then the cubic regularization method given by Algorithm 1 finds an ϵ-SOSP after at most*

$$\frac{\sqrt{\gamma}(f(x_1) - f(x^*))}{\epsilon^{1.5}}$$

iterations.

Algorithm 1 Cubic Regularization

```
Initialize $x_1 \in \mathbb{R}^d$
for $t = 1, \ldots, T$ do
    Take
        $$x_{t+1} \in \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ \nabla f(x_t)^\top (x - x_t) + \frac{1}{2}(x - x_t)^\top \nabla^2 f(x_t)(x - x_t) + \frac{\gamma}{6}\|x - x_t\|_2^3 \right\}.$$
end for
```

2.2 Perturbed Gradient Descent

So far, we have discussed two algorithms for computing an approximately second-order stationary points. Both algorithms require second-order information from the Hessian, and as a resuit, they are usually more time-consuming than first-order methods. In fact, there exists a variant of gradient descent that helps to escape from saddle points. The algorithm is due to Jin et al. [JGN+17]. The main idea is to add noise if gradient descent ends up with a stationary point. For this reason, the algorithm is referred to as **perturbed gradient descent**. Let us provide a description of the algorithm. Algorithm 2 perturbs the current point if its gradient norm is sufficiently small. Here, perturbation replaces the step of identifying a negative eigenvalue of the Hessian. [JGN+17] proved that Algorithm 2 finds an ϵ-SOSP with high probability.

Algorithm 2 Perturbed Gradient Descent

```
Initialize $x_1 \in \mathbb{R}^d$
for $t = 1, \ldots, T$ do
    if $\|\nabla f(x_t)\|_2 \leq \epsilon$ and no perturbation has been made for the last $\tau$ iterations then
        Perturb the current point: $x_t \leftarrow x_t + \epsilon_t$ where $\epsilon_t \sim \mathcal{N}(0, \sigma^2 I)$ for some $\sigma > 0$
    else
        Apply gradient descent: $x_{t+1} = x_t - \eta \nabla f(x_t)$.
    end if
end for
```

Theorem 13.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be β-smooth function whose Hessian is γ-Lipschitz continuous. Then Algorithm 2 with $\eta = 1/\beta$ and properly choosing parameters τ and σ finds an ϵ-SOSP after*

$$\tilde{O}\left(\frac{\beta(f(x_1) - f(x^*))}{\epsilon^2}\right)$$

iterations with high probability.

3 Applications where a SOSP is Sufficient

In the previous section, we explained that for any $\epsilon > 0$, we can find an ϵ-SOSP x that satisfies

$$\|\nabla f(x)\|_2 \leq \epsilon \quad \text{and} \quad \nabla^2 f(x) \succeq -\sqrt{\gamma\epsilon} I.$$

We say that a point y is a **strict saddle point** if

$$\|\nabla f(y)\|_2 = 0 \quad \text{and} \quad \nabla^2 f(y) \not\succeq 0.$$

Recall that we may have a saddle point with a positive semidefinite Hessian although a saddle point is a local minimum if its Hessian is positive definite. Hence, applying an algorithm for finding an ϵ-SOSP for a sufficiently small ϵ, we may avoid strict saddle points.

In this section, we present some applications where a second-order stationary point is a global minimum. To be more precise, we will cover some application settings that have the following two conditions.

1. Every saddle points is strict, which means that there is no saddle point at which the Hessian is positive semidefinite.

2. Every local minimum is a global minimum.

The first condition guarantees that a second-order stationary point is a local minimum. Then it follows from the second condition that a second-order stationary point is a global minimum.

3.1 Computing the Top Eigenvector

Let $A \in \mathbb{R}^{d\times d}$ be a positive semidefinite matrix with eigvenvalues

$$\lambda_1 > \lambda_2 > \cdots > \lambda_d \geq 0$$

and the associated eigvenvectors $v_1, v_2, \ldots, v_d$. Recall that

$$A_1 \in \text{argmin}_{X\in\mathbb{R}^{d\times d}} \{\|A - X\|_F : \ \text{rank}(X) \leq 1\}$$

where $A_1 = \lambda_1 v_1 v_1^\top$. This implies that

$$x = \sqrt{\lambda_1} v_1 \quad \text{and} \quad x = -\sqrt{\lambda_1} v_1$$

are minimizers of

$$\min_{x\in\mathbb{R}^d} \quad \frac{1}{4}\left\|A - xx^\top\right\|_F^2.$$

In fact, we may argue that no other point is a minimizer. Let $f(x)$ denote the objective function

$$f(x) = \frac{1}{4}\left\|A - xx^\top\right\|_F^2 = \frac{1}{4}\sum_{i=1}^d \left(x_i^2 - A_{ii}\right)^2 + \frac{1}{4}\sum_{i=1}^d\sum_{j=1}^d (x_i x_j - A_{ij})^2.$$

Therefore, it follows that

$$\nabla f(x) = \|x\|_2^2 \cdot x - Ax.$$

Moreover,

$$\nabla^2 f(x) = \|x\|_2^2 \cdot I + 2xx^\top - A.$$

Remark 13.3. If x is a stationary point of f, then

$$x \in \{\sqrt{\lambda_i} v_i, -\sqrt{\lambda_i} v_i\}$$

for some $i \in \{1, \ldots, d\}$. This is because, if $\nabla f(x) = 0$, then we have $Ax = \|x\|_2^2 \cdot x$. This implies that x is a scalar multiple of some eigenvector v_i and $\|x\|_2^2 = \lambda_i$. In this case, we have $x = \sqrt{\lambda_i} v_i$ or $x = -\sqrt{\lambda_i} v_i$.

Nevertheless, we know that $\pm\sqrt{\lambda_i}v_i$ are not minimizers of f although they are stationary points. Next we will argue that $\pm\sqrt{\lambda_i}v_i$ for $i \geq 2$ are not second-order stationary points.

Remark 13.4. Let $y \in \{\sqrt{\lambda_i}v_i, -\sqrt{\lambda_i}v_i\}$ for some $i > 2$. Note that

$$\nabla^2 f(y) = \lambda_i \|v_i\|_2^2 \cdot I + 2\lambda_i v_i v_i^\top - A.$$

Then it follows that

$$v_1^\top \nabla^2 f(y) v_1 = \lambda_i \|v_i\|_2^2 \|v_1\|_2^2 + 2\lambda_i (v_i^\top v_1)^2 - v_1^\top A v_1 = \lambda_i - \lambda_1 < 0.$$

Therefore, y is a strict saddle point and thus is not a second-order stationary point.

3.2 Low-Rank Matrix Factorization

Given a positive semidefinite matrix $A \in \mathbb{R}^{d\times d}$ with eigvenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_d \geq 0$ and the associated eigvenvectors $v_1, v_2, \ldots, v_d$, we consider

$$\min_{X\in\mathbb{R}^{d\times k}} \quad \frac{1}{4}\left\|A - XX^\top\right\|_F^2$$

for some $k \geq 1$. We know that

$$A_k \in \operatorname{argmin}_{X\in\mathbb{R}^{d\times d}} \{\|A - X\|_F : \ \operatorname{rank}(X) \leq k\}$$

where $A_k = U_k \Sigma_k U_k^\top$. Hence,

$$X = \Sigma_k^{1/2} U_k \quad \text{and} \quad X = -\Sigma_k^{1/2} U_k$$

are minimizers of the problem. As in the top eigenvector problem we can show that no other matrix is a minimizer. As before, we can argue that the other stationary points are strict saddle points.

References

[JGN+17] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1724–1732. PMLR, 06–11 Aug 2017. 2.2, 2.2

[NP06] Yurii Nesterov and B.T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, page 177–205, 2006. 2.1