ESCAPING SADDLE POINTS IN NON-CONVEX OPTIMIZATION

Giorgos Gotzias

GRADIENT DESCENT

Gradient Descent step:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

- If f is convex, then gradient descent converges to an area around a stationary point (point with $\nabla f(x_t) = 0$)
- That point is the minimum for every convex function
- If f isn’t convex, then gradient descent still converges to a stationary point
- However, a stationary point may be a saddle point or a local maximum or local minimum instead of the optimal point.

Gradient Descent

Possible changes to avoid saddle points

- ➢ Intermittent Perturbations: Add a random perturbation when a condition is satisfied to the Gradient Descent step
- ➢ Random Initialization: The initialization of the Gradient Descent Algorithm should be random

- Both methods ensure escaping from saddle points
- Although random initialization may require exponential time

PERTURBED GRADIENT DESCENT

1. For $t = 1, 2, \ldots$ do
2. $x_t \leftarrow x_{t-1} - \eta \nabla f(x_{t-1})$
3. If perturbation condition holds then
4. $x_t \leftarrow x_t + \xi_t$

- ➢ Where ξ_t is chosen uniformly from Ball centered at zero with a small radius
- ➢ The condition ensures that the gradient value is near to zero

SADDLE POINTS

- Local maxima are include in the term “Saddle Points”
- For each Saddle Point, the gradient value is equal to zero, and the minimum eigenvalue of the Hessian Matrix is non-positive
- There is at least one direction in which they are local maxima
- If the minimum eigenvalue of the Hessian Matrix at a specific point x is equal to zero, this point is either a saddle point or a local minimum
- If the minimum eigenvalue of the Hessian Matrix is negative, then that point is a strict saddle point

Some Definitions

- The function f is l-gradient Lipschitz, if
  - For each x_1, x_2, $|\nabla f(x_1) - \nabla f(x_2)| \leq l|x_1 - x_2|$
- The function f is ρ-Hessian Lipschitz, if
  - For each x_1, x_2, $||\nabla^2 f(x_1) - \nabla^2 f(x_2)|| \leq \rho|x_1 - x_2|$
- The point x is second-order stationary point, if
  - $\nabla f(x) = 0$ and $\lambda_{min}(\nabla^2 f(x)) \geq 0$
- The point x is ε-second-order stationary point, if
  - $|\nabla f(x)| \leq \varepsilon$ and $\lambda_{min}(\nabla^2 f(x)) \geq -\sqrt{(\rho\varepsilon)}$

Main Theorem

- If f is a l-gradient and ρ-Hessian Lipschitz function, then the Perturbed Gradient Descent algorithm with $\eta = O(1/l)$ finds an ε-second-order stationary point with high probability after $O(l\, f(x_0) - f^*)/\varepsilon^2)$ iterations (polynomials of logarithmic terms are ignored)
- If all saddle points of this function are strict, then all second-order stationary points are local minima
- The number of iterations is asymptotically equal to the number of iterations of common Gradient Descent for convex functions (ignoring logarithmic terms)

Perturbed Gradient Descent

1. $t_{noise} \leftarrow - t_{thres} - 1$
2. **For** $t = 0, 1, 2, \ldots$ **do**
3. **If** $\|\nabla f(x_t)\| \leq g_{thres}$ and $t - t_{noise} > t_{thres}$ **then**
4. $x'_t \leftarrow x_t$, $t_{noise} \leftarrow t$
5. $x_t \leftarrow x'_t + \xi_t$
6. **If** $t - t_{noise} = t_{thres}$ and $f(x_t) - f(x'_{tnoise}) > -f_{thres}$ **then**
7. **Return** x'_{tnoise}
8. $x_t \leftarrow x_{t-1} - \eta \nabla f(x_{t-1})$

Main Ideas to Prove Main Theorem

- In iteration t, x_t becomes a point such that $\|\nabla f(x_t)\| > g_{thres}$, and for $\eta < 1/l$, the following statement is true:
  - $f(x_{t+1}) \leq f(x_t) - (\eta/2) \|\nabla f(x_t)\|^2$

- If x_t is a point such that $\|\nabla f(x_t)\| \leq g_{thres}$ and $\lambda\min(\nabla^2 f(x)) < -\sqrt{\rho\varepsilon}$, then a perturbation is added and after t_{thres} iterations of common Gradient Descent, it stands with high probability that
  - $f(x_{t+tthres}) - f(x_t) \leq - f_{thres}$

COROLLARY

- If the function is (θ, γ, ζ)-strict saddle, then for each x, at least one of the following stands true:
  - $\|\|\nabla f(x_t)\|\| \geq \theta$
  - $\lambda_{min}(\nabla^2 f(x)) \leq -\gamma$
  - The distance of x from the nearest local minimum is at most ζ
- If the function f is l-gradient Lipschitz, ρ-Hessian Lipschitz and (θ, γ, ζ)-strict saddle,then the Perturbed Gradient Descent algorithm finds a point with a distance from a local minimum at most ζ with high probability after $O(l\ f(x_0) - f^* / \varepsilon^2)$ iterations, if logarithmic terms are ignored

Proof Of Corollary

- According to main theorem, the Perturbed Gradient Descent finds a ε-second-order stationary point x, hence:
  - $|\nabla f(x)| \leq \varepsilon$ και $\lambda_{min}(\nabla^2 f(x)) \geq -\sqrt{(\rho\varepsilon)}$
- However, $\varepsilon = \min(\theta, \gamma^2/\rho)$, so it is $\varepsilon \leq \theta$ and $\varepsilon \leq \gamma^2/\rho$
- Thus, $|\nabla f(x)| \leq \theta$ and $\lambda_{min}(\nabla^2 f(x)) \geq -\gamma$
- Since f is (θ, γ, ζ)-strict saddle, the third condition have to be satisfied for the point x