ESCAPING SADDLE POINTS NON-CONVEX OPTIMIZATION

Giorgos Gotzias

GRADIENT DESCENT

Gradient Descent step:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \ \nabla \mathbf{f}(\mathbf{x}_t)$$

- If f is convex, then gradient descent converges to an area around a stationary point (point with $\nabla f(x_*) = 0$)
- 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, ... 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

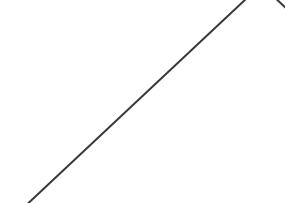
- > The function f is I-gradient Lipschitz, if
 - For each $x_1, x_2, |\nabla f(x_1) \nabla f(x_2)| \le ||x_1 x_2||$
- \succ The function f is ρ -Hessian Lipschitz, if
 - For each $x_1, x_2, \|\nabla^2 f(x_1) \nabla^2 f(x_2)\| \le \rho |x_1 x_2|$
- > The point x is second-order stationary point, if
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

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 $O(l\ f(x_0) f^*)/\epsilon^2)$ 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. $t_{\text{noise}} \leftarrow -t_{\text{thres}} 1$
- 2. For t = 0, 1, 2, ... do
- 3. If $\|\nabla f(x_t)\| \le g_{thres}$ and $t t_{noise} > t_{thres}$ then
- 4. $x'_{t} \leftarrow x_{t}, t_{noise} \leftarrow t$
- 5. $x_{\downarrow} \leftarrow x'_{\downarrow} + \xi_{\downarrow}$
- 6. If $t t_{\text{noise}} = t_{\text{thres}}$ and $f(x_t) f(x'_{\text{tnoise}}) > -f_{\text{thres}}$ then
- 7. **Return** x'_{tnoise}
- 8. $x_{t-1} \eta \nabla f(x_{t-1})$



MAIN IDEAS TO PROVE MAIN THEOREM

• In iteration t, x_t becomes a point such that $\|\nabla f(x_t)\| > g_{thres}$, and for $\eta < 1/I$, the following statement is true:

$$\circ f(x_{t+1}) \le f(x_t) - (\eta/2) \|\nabla f(x_t)\|^2$$

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

$$\circ$$
 $f(x_{t+tthres}) - f(x_t) \le - f_{thres}$

COROLLARY

- \succ If the function is (θ, γ, ζ)-strict saddle, then for each x, at least one of the following stands true:
 - $\bigcirc \qquad |||\nabla f(x_{\downarrow})|| \geq \theta$
 - $0 \qquad \lambda_{\min}(\nabla^{'2}f(x)) \leq -\gamma$
 - \circ The distance of x from the nearest local minimum is at most ζ
- Figure 1. 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*/ ϵ^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:
 - $|\nabla f(x)| ≤ ε και λ_{min} <math> (\nabla^2 f(x)) ≥ √(ρε)$
- However, ε = min(θ, γ^2/ρ), so it is ε ≤ θ and ε ≤ γ^2/ρ
- ightharpoonup Thus, $|\nabla f(x)| \le \theta$ and $\lambda_{\min}(\nabla^2 f(x)) \ge -\gamma$
- Since f is (θ, γ, ζ) -strict saddle, the third condition have to be satisfied for the point x