Brief Introduction to Saddle Point Escaping Problem

YeLab Group Seminar

March 30, 2019

Overview

- Nonconvex Optimization
- Saddle Point
- Escape Saddle Point
- Some Recent Works
- References

Overview

- Nonconvex Optimization
- Saddle Point
- Escape Saddle Point
- Some Recent Works
- References

Nonconvex Optimization

- Goal: minimize a nonconvex function $f(\mathbf{x})$

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}),$$

where $f(\mathbf{x})$ is nonconvex and $\mathcal{X}$ is the feasible set of $\mathbf{x}$.

- $f(\mathbf{x})$ can have a finite-sum structure

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}),$$

where m is the number of samples.

Nonconvex Optimization

- Example: Nonconvex function $f([x_1, x_2]) = \frac{x_1^2}{x_1^2+1} + \frac{x_2^2}{x_2^2+1}$.

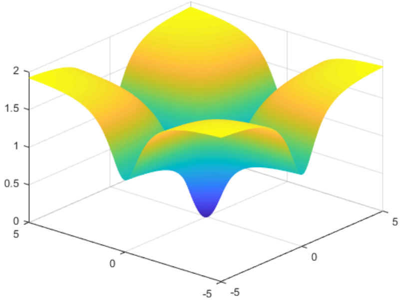

Figure 1: Nonconvex function example.

Nonconvex Optimization

- First Order Methods
  - Gradient Descent (GD)

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

  - Stochastic Gradient Descent (SGD)

$$\mathbf{x}_t = \mathbf{x}_{t-1} - \gamma_t \nabla f_{i_t}(\mathbf{x}_{t-1}),$$

    where i_t is uniformly and independently sampled from $1, \ldots, m$.
  - Other methods only involving first order information.
- Can first-order method lead to global convergence?

Overview

- Nonconvex Optimization
- Saddle Point
- Escape Saddle Point
- Some Recent Works
- References

Saddle Point

- Critical Point (first-order stationary point): $\nabla f(\mathbf{x}) = 0$.
  - Local minimum.
  - Global minimum: also a local minimum.
  - Saddle point: critical point but not local minimum.
- $\nabla f(\mathbf{x}) = 0$ and $\lambda_{\min}[\nabla^2 f(\mathbf{x})] \begin{cases} > 0, & \text{local minimum} \\ = 0, & \text{local minimum or saddle point} \\ < 0, & \text{strict saddle point.} \end{cases}$

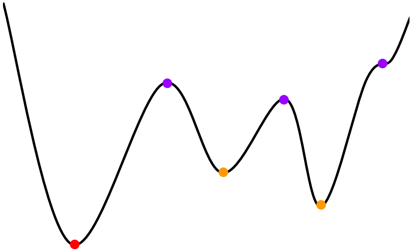

Figure 2: Illustration of critical points.

Saddle Point

- More saddle points examples.
  - $f(x) = x^3$.
  - $f'(0) = 0$.
  - $f''(0) = 0$.
  - Non-strict saddle point.

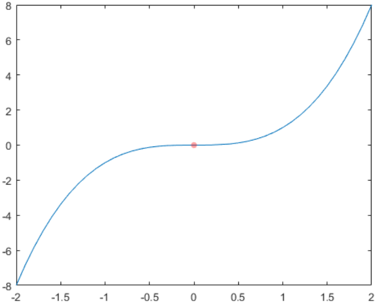

Figure 3: $f(x) = x^3$

Saddle Point

- More saddle points examples.
  - $f([x_1, x_2]) = x_1^3 - 3x_1x_2^2$.
  - $\nabla f([0,0]) = [0,0]$.
  - $\nabla^2 f([0,0]) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \Rightarrow \lambda_{\min}[\nabla^2 f([0,0])] = 0$.
  - Non-strict saddle point.

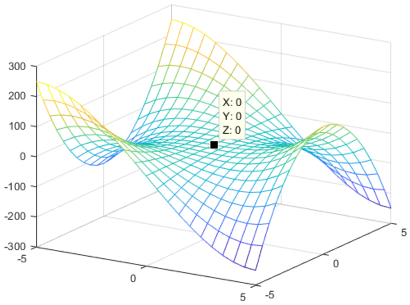

Figure 4: $y = x_1^3 - 3x_1x_2^2$

Saddle Point

- More saddle points examples.
  - $f([x_1, x_2]) = x_1^2 - x_2^2$.
  - $\nabla f([0,0]) = [0,0]$.
  - $\nabla^2 f([0,0]) = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix} \Rightarrow \lambda_{\min}[\nabla^2 f([0,0])] = -2$.
  - Strict saddle point.

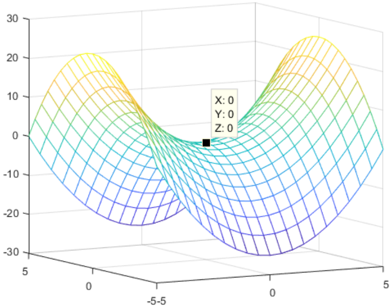

Figure 5: $f([x_1, x_2]) = x_1^2 - x_2^2$

Saddle Point

- In a wide range of practical nonconvex problems, it has been proved that all saddle points are strict. [JJ].
- E.g., PCA, orthogonal tensor decomposition, phase retrieval, dictionary learning, matrix sensing, matrix completion, and ... [GLM16, GJZ17]
- Restrict our discussion to **strict saddle function**.

Saddle Point

- Can first-order method lead to global convergence?
- Not guaranteed.
  - Most previous analysis only targets at finding $\mathbf{x} : \|f(\mathbf{x})\|_2 = 0$ efficiently.

$$T > O(?), \text{ s.t. } \|\nabla f(\mathbf{x}_T)\|_2 \leq \varepsilon. (\text{ deterministic algorithm})$$

  - Not even a local optimum.
  - It might be a strict saddle point.

Overview

- Nonconvex Optimization
- Saddle Point
- Escape Saddle Point
- Some Recent Works
- References

Escape Saddle Point

- Why? If $\mathbf{x}$ is strict saddle, there should exist local minimum $\mathbf{y}$ such that

$$f(\mathbf{y}) \leq f(\mathbf{x}).$$

- Can we design an algorithm to escape strict saddle with theoretical guarantee?

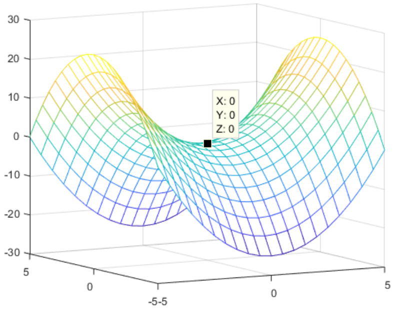

Figure 6: Strict saddle point.

Escape Saddle Point

Assumption 1

Two main assumptions are used in the analysis:

$$
\begin{aligned}
&f(\mathbf{x}) \textit{ is strict saddle function,}\\
&\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \le L\|\mathbf{x}-\mathbf{y}\|_2,\\
&\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\|_2 \le \rho\|\mathbf{x}-\mathbf{y}\|_2.
\end{aligned}
$$

- Goal: design optimization algorithms to efficiently find **local minimum**

$$\mathbf{x} : \|f(\mathbf{x})\|_2 = 0 \text{ and } \lambda_{\min}[\nabla^2 f(\mathbf{x})] \ge 0.$$

- Precisely,

$$T > O(?) : \|f(\mathbf{x}_T)\|_2 \le \varepsilon, \text{ and } \lambda_{\min}[\nabla^2 f(\mathbf{x}_T)] \ge -\varepsilon_H := \sqrt{\rho\varepsilon}$$

(second-order stationary point)

- ε can be arbitrarily small as T gets larger.

Overview

- Nonconvex Optimization
- Saddle Point
- Escape Saddle Point
- Some Recent Works
- References

Some Recent Works

- Full gradient method.
  - Random initialization
  - Random perturbation
- Stochastic gradient method.
  - Beyond first order information
  - SGD variants
  - SGD

Some Recent Works

- Full gradient — Random initialization [LSJR16, PP16].

Theorem 2

GD with a random initialization and sufficiently small constant step size converges to a local minimizer or negative infinity almost surely.

  - Asymptotic result.
  - May take exponential time.

Some Recent Works

- Full gradient method — Random perturbation.
  - Perturbed GD : $T \geq \widetilde{O}(\varepsilon^{-2})$. [JGN+17]
  - Perturbed Accelerated GD: $T \geq \widetilde{O}(\epsilon^{-1.75})$. [CDHS18, AAZB+17, JNJ17]

Algorithm 1 Perturbed GD

for $t = 1, ..., T$ **do**
    if perturbation condition holds **then**
        $\mathbf{x}_t \leftarrow \mathbf{x}_t + \xi_t$, where ξ_t uniformly $\sim \mathbb{B}_0(r)$.
    end if
    $\mathbf{x}_t = \mathbf{x}_{t-1} - \gamma_t \nabla f(\mathbf{x}_{t-1})$.
end for

Some Recent Works

- Stochastic gradient method - Beyond first order information.
  - Third order smoothness: $T \geq \widetilde{O}(\varepsilon^{-10/3})$. [YXG18]
  - Using Hessian information:
    - ⋆ Cubic regularized Newton: $T \geq \widetilde{O}(\varepsilon^{-3.5})$. [TSJ+18]
    - ⋆ Negative curvature search: $T \geq \widetilde{O}(\varepsilon^{-3.5})$. [AZ18]

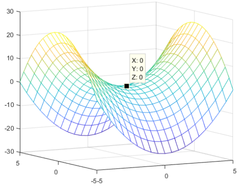

Figure 7: Negative curvature search

Some Recent Works

- Stochastic gradient method - SGD variants.
  - First-order approximates negative curvature search: $T \geq \widetilde{O}(\varepsilon^{-3.5})$. [AZL18]
  - Spider: $T \geq \widetilde{O}(\varepsilon^{-3})$. [FLLZ18]
  - Perturbated SGD: $T \geq \widetilde{O}(d\varepsilon^{-4})$. [JNG+19]

Algorithm 2 Perturbated SGD

1: **for** $t = 1, ..., T$ **do**
2: $\quad \mathbf{x}_t = \mathbf{x}_{t-1} - \gamma_t(g_{t-1} + \xi_t)$, where $\xi_t \sim \mathcal{N}(0, \delta I)$.
3: **end for**

Some Recent Works

- Stochastic gradient method - SGD.
  - SGD: $T \geq \widetilde{O}(\varepsilon^{-3.5})$. [FLZ19]

Algorithm 3 SGD

1: **for** $t = 1, ..., T$ **do**
2: $\quad \mathbf{x}_t = \mathbf{x}_{t-1} - \gamma_t g_{t-1}$.
3: **end for**

Overview

- Nonconvex Optimization
- Saddle Point
- Escape Saddle Point
- Some Recent Works
- References

References I

Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma, *Finding approximate local minima faster than gradient descent*, Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, ACM, 2017, pp. 1195–1199.

Zeyuan Allen-Zhu, *Natasha 2: Faster non-convex optimization than sgd*, Advances in Neural Information Processing Systems, 2018, pp. 2680–2691.

Zeyuan Allen-Zhu and Yuanzhi Li, *Neon2: Finding local minima via first-order oracles*, Advances in Neural Information Processing Systems, 2018, pp. 3720–3730.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford, *Accelerated methods for nonconvex optimization*, SIAM Journal on Optimization **28** (2018), no. 2, 1751–1772.

References II

- Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang, *Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator*, Advances in Neural Information Processing Systems, 2018, pp. 687–697.
- Cong Fang, Zhouchen Lin, and Tong Zhang, *Sharp analysis for nonconvex sgd escaping from saddle points*, arXiv preprint arXiv:1902.00247 (2019).
- Rong Ge, Chi Jin, and Yi Zheng, *No spurious local minima in nonconvex low rank problems: A unified geometric analysis*, Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 1233–1242.
- Rong Ge, Jason D Lee, and Tengyu Ma, *Matrix completion has no spurious local minimum*, Advances in Neural Information Processing Systems, 2016, pp. 2973–2981.

References III

- Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan, *How to escape saddle points efficiently*, Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 1724–1732.
- Chi Jin and Michael Jordan, *How to escape saddle points efficiently*, http://www.offconvex.org/2017/07/19/saddle-efficiency/.
- Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan, *Stochastic gradient descent escapes saddle points efficiently*, arXiv preprint arXiv:1902.04811 (2019).
- Chi Jin, Praneeth Netrapalli, and Michael I Jordan, *Accelerated gradient descent escapes saddle points faster than gradient descent*, arXiv preprint arXiv:1711.10456 (2017).
- Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht, *Gradient descent converges to minimizers*, arXiv preprint arXiv:1602.04915 (2016).

References IV

Ioannis Panageas and Georgios Piliouras, *Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions*, arXiv preprint arXiv:1605.00405 (2016).

Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I Jordan, *Stochastic cubic regularization for fast nonconvex optimization*, Advances in Neural Information Processing Systems, 2018, pp. 2904–2913.

Yaodong Yu, Pan Xu, and Quanquan Gu, *Third-order smoothness helps: Faster stochastic optimization algorithms for finding local minima*, Advances in Neural Information Processing Systems, 2018, pp. 4530–4540.