How to Escape Saddle Points Efficiently

Chi Jin*, Rong Ge ¶, Praneeth Netrapalli ǂ, Sham M. Kakade ⁑, Michael I. Jordan*

* UC Berkeley

¶ Duke University

ǂ Microsoft Research India

⁑ University of Washington, Seattle

Non-convex optimization

Problem: $\min_x f(x)$ $f(\cdot)$: non-convex function

Applications: Matrix/tensor factorization, Distribution learning, neural networks,…

Gradient descent (GD)

Problem: $$\min_x f(x)$$

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

Stepsize (η)

Gradient ($\nabla f(x_t)$)

GD for smooth non-convex functions

- Smoothness: $\|\nabla f(x) - \nabla f(y)\| \leq \ell \|x - y\|$

- Global optimum may not be achievable in general

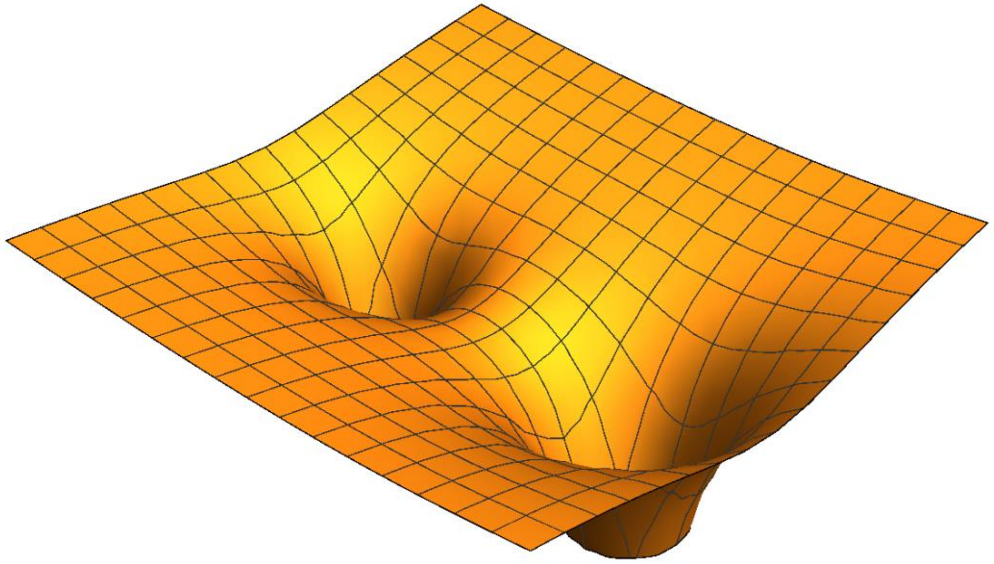
- $\|\nabla f(x_t)\| < \epsilon$ in $t = O\left(\frac{\ell(f(x_0) - f^*)}{\epsilon^2}\right)$ (Nesterov 1998)

ϵ- first order stationary point

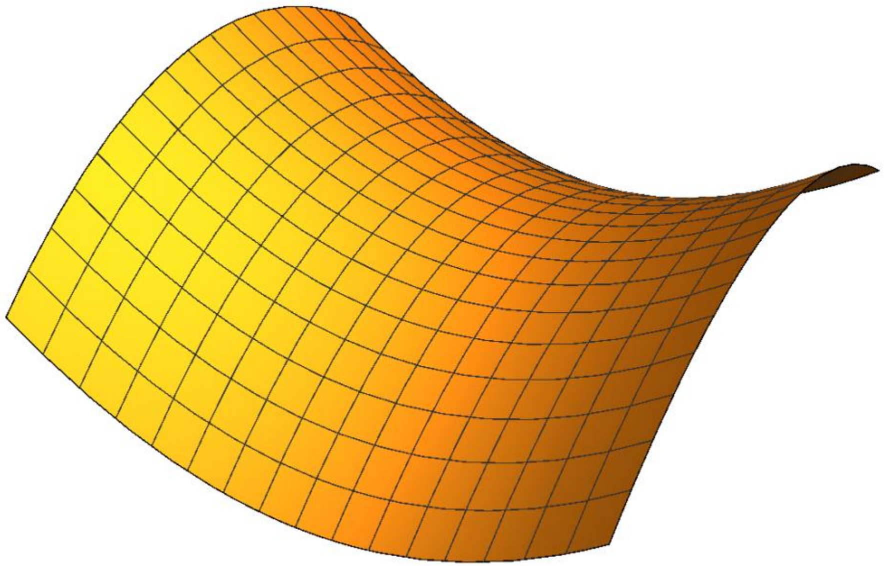
$$f^* \stackrel{\text{def}}{=} \min_x f(x)$$

First-order stationary points

Local minima



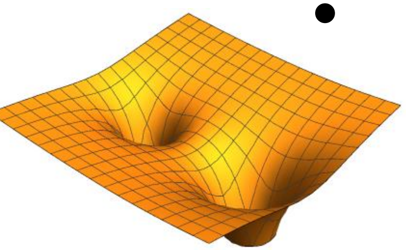
Saddle points/local maxima



First-order stationary points

In many applications such as PCA, matrix completion, dictionary learning etc.

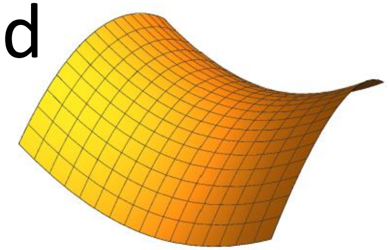
Local minima



- Either all local minima are global minima
- Or all local minima as good as global minima

Saddle points

- Very poor compared to global minima
- Several such points



First-order stationary points

In many applications such as PCA, matrix completion, dictionary learning etc.

Bottomline: Local minima much more desirable than saddle points

However, gradient descent can indeed converge to saddle points.

Can gradient descent escape saddle points?

- By adding noise -- best known results $\mathrm{poly}(d)$ (Ge et al. 2015)

Question: How to escape saddle points efficiently?

Second-order stationary points

- Smoothness: $\|\nabla f(x) - \nabla f(y)\| \leq \ell \|x - y\|$
- Hessian Lipschitz: $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq \rho \|x - y\|$
- x an ϵ-second order stationary point if (Nesterov and Polyak 2006)

$$\|\nabla f(x)\| \leq \epsilon \qquad \text{and} \qquad \lambda_{\min}\left(\nabla^2 f(x)\right) \geq -\sqrt{\rho\epsilon}$$

Our result

Perturbed gradient descent finds ϵ-second order stationary point

in $t = \tilde{O}\left(\frac{\ell(f(x_0)-f^*)}{\epsilon^2}\right)$

- Second order stationary point instead of first order stationary point
- In essentially the same amount of time as gradient descent finds first order stationary point

Perturbed gradient descent

1. **For** $t = 0,1,\cdots$ **do**
2. **if** perturbation_condition_holds **then**
3. $x_t \leftarrow x_t + \xi_t$ where $\xi_t \sim Unif\left(B_0(\epsilon/\ell)\right)$
4. $x_{t+1} \leftarrow x_t - \eta\nabla f(x_t)$

1. $\nabla f(x_t)$ is small
2. No perturbation in last several iterations

Recall second order stationary point

$$\|\nabla f(x)\| \le \epsilon$$
$$\lambda_{\min}\left(\nabla^2 f(x)\right) \ge -\sqrt{\rho\epsilon}$$

Proof idea

- Case I: $\|\nabla f(x_t)\| > \epsilon$

  Smoothness

  Stepsize $\eta = \frac{1}{\ell}$

  $$\Rightarrow f(x_{t+1}) \le f(x_t) - \frac{1}{2\ell}\|\nabla f(x_t)\|^2 \le f(x_t) - \frac{1}{2\ell}\epsilon^2$$

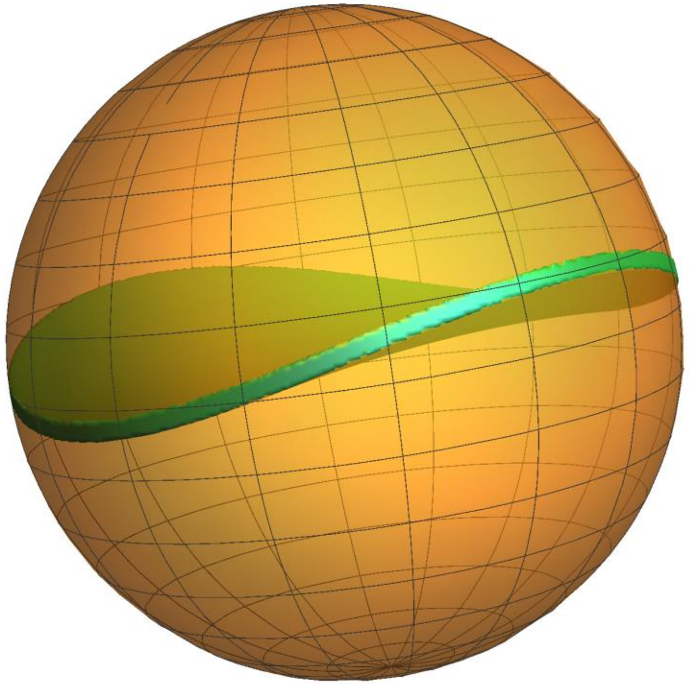
- Case II: $\|\nabla f(x_t)\| \le \epsilon$ and $\lambda_{\min}\left(\nabla^2 f(x_t)\right) < -\sqrt{\rho\epsilon}$

  $x_t \sim$ saddle point

How do we escape from here?

Geometry around saddle points

$S \stackrel{\text{def}}{=}$ set of points around saddle point from where gradient descent does not escape saddle point.

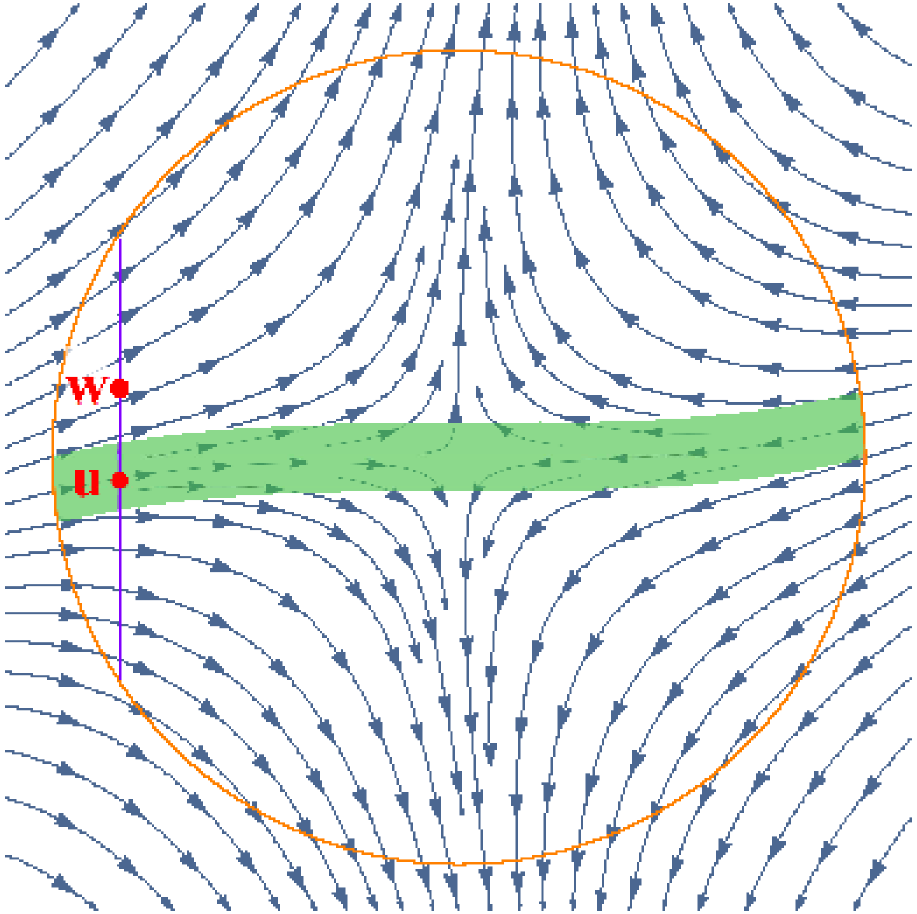


Key technical result

$\mathrm{Vol}(S)$ is small

Geometry around saddle points

$S \overset{\text{def}}{=}$ set of points around saddle point from where gradient descent does not escape saddle point.



Key technical result

$\text{Vol}(S)$ is small

Recap

- Gradient descent converges to first order stationary points
- Perturbed gradient descent converges to second order stationary points
- Depends only logarithmically on dimension
- Key idea: understand structure around saddle points

Further results using local structure

- Strict saddle property: Every saddle point has a strictly negative eigenvalue
  - PCA, CCA, matrix sensing/completion, dictionary learning, orthogonal tensor decomposition etc.
  - Converge to local minima

- Local strong convexity
  - PCA, CCA, matrix factorization
  - Local geometric convergence

Conclusions

- (Gradient descent + a little randomness) can escape saddle points
- In fact, efficiently. Only $\mathrm{polylog}(\mathrm{d})$ dependence.
- Key ingredient: understand geometry around saddle points

<u>Some open directions</u>

- Is randomness in the beginning sufficient?
- Do momentum methods help accelerate for non-convex problems?
- Extensions to the stochastic case