How to Escape Saddle Points Efficiently

Chi Jin Rong Ge Praneeth Netrapalli Sham M. Kakade
Michael I. Jordon

University of California Berkeley

Duke University

Microsoft Research India

University of Washington

ICML, 2017
Presenter: Ritambhara Singh

Outline

1. Introduction
   - Motivation
   - Background
   - State-of-the-art

2. Proposed Approach
   - Perturbed Gradient Descent
   - Proof Sketch

Outline

1. Introduction
   - Motivation
   - Background
   - State-of-the-art
2. Proposed Approach
   - Perturbed Gradient Descent
   - Proof Sketch

Motivation

- Theoretical analysis of perturbed gradient descent algorithm (show it is almost "dimension free")
- Perturbed gradient descent can escape saddle points for free
- Novel characterization of geometry around saddle points

Motivation

- Theoretical analysis of perturbed gradient descent algorithm (show it is almost "dimension free")
- Perturbed gradient descent can escape saddle points for free
- Novel characterization of geometry around saddle points

Algorithm 1 Perturbed Gradient Descent (Meta-algorithm)

for $t = 0, 1, \ldots$ **do**
  if perturbation condition holds **then**
    $\mathbf{x}_t \leftarrow \mathbf{x}_t + \xi_t, \quad \xi_t$ uniformly $\sim \mathbb{B}_0(r)$
  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$

Outline

Gradient Descent: Convex Problem

Gradient Descent:

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

Gradient Descent: Convex Problem

Gradient Descent:

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

Definition

A differentiable function f(.) is l-smooth (or l-gradient Lipschitz):

$$\forall x_1, x_2, ||\nabla f(x_1) - \nabla f(x_2)|| \le l||x_1 - x_2|| \tag{2}$$

Gradient Descent: Convex Problem

Gradient Descent:

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

Definition

A differentiable function f(.) is l-smooth (or l-gradient Lipschitz):

$$\forall x_1, x_2, ||\nabla f(x_1) - \nabla f(x_2)|| \leq l||x_1 - x_2|| \tag{2}$$

Definition

A twice-differentiable function f(.) is α-convex if $\forall x, \lambda_{min}(\nabla^2(f(x)) > \alpha$

Gradient Descent: Convex Problem

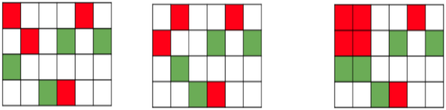
Theorem

Assume above holds for f. For any $\epsilon > 0$, if we run a gradient descent with step $\eta = \frac{1}{l}$, iterate x_t will be ϵ-close to $x*$ in iterations:

$$\frac{2l}{\alpha} log \frac{||x_0 - x*||}{\epsilon} \tag{3}$$

Gradient Descent: Non-Convex Problem

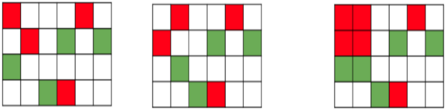
Permutation Symmetry



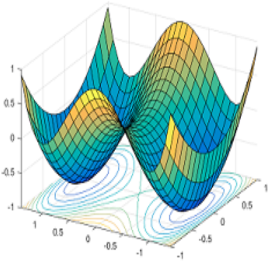
Optimal Solution Equivalent Solution Not optimal

Gradient Descent: Non-Convex Problem

Permutation Symmetry



Optimal Solution   Equivalent Solution   Not optimal



Gradient Descent: Non-Convex Problem

Definition

For a differentiable function f(.), we say that x is a **first order stationary point** if $||\nabla f(x)|| = 0$; also it is ϵ-**first order stationary point** if $||\nabla f(x)|| \leq \epsilon$.

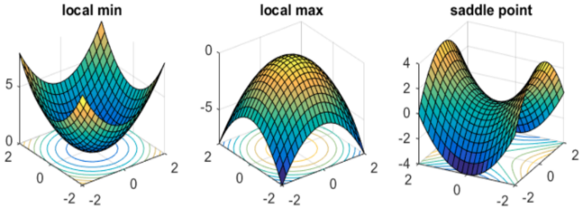
Gradient Descent: Non-Convex Problem

Theorem

Assume f(.) is l-smooth. Then, for any $\epsilon\, 0$*, if we run gradient descent with step size* $\eta = \frac{1}{l}$ *and termination condition* $||\nabla f(x)|| \leq \epsilon$*, the output will be a* ϵ-**first order stationary point***, and the algorithm will terminate within following number of iterations*

$$\frac{l(f(x_0) - f*)}{\epsilon^2} \tag{4}$$

Types of Critical Points



Second Order Properties

Definition

A twice-differentiable function f(.) is ρ-Hessian Lipshitz if:

$$\forall x_1, x_2, ||\nabla^2 f(x_1) - \nabla^2 f(x_2)|| \leq \rho ||x_1 - x_2|| \tag{5}$$

Second Order Properties

Definition

A twice-differentiable function f(.) is ρ-Hessian Lipshitz if:

$$\forall x_1, x_2, ||\nabla^2 f(x_1) - \nabla^2 f(x_2)|| \leq \rho ||x_1 - x_2|| \tag{5}$$

Definition

For a ρ–Hessian Lipshitz function f(.), we say that x is a **second order stationary point** if $||\nabla f(x)|| = 0$ and $\lambda_{min}(\nabla^2 f(x)) \geq 0$; also it is ϵ-**second order stationary point** if

$$||\nabla f(x)|| \leq \epsilon; \lambda_{min}(\nabla^2 f(x)) \geq -\sqrt{\rho\epsilon} \tag{6}$$

Escaping Saddle Points

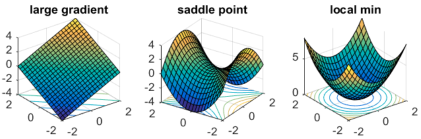
Second order Taylor Expansion:

$$f(y) \approx f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}(y - x)^\top \nabla^2 f(x)(y - x).$$

Strict Saddle Points

A function $f(\boldsymbol{x})$ is strict saddle if all points $\boldsymbol{x}$ satisfy at least one of the following

1. Gradient $\nabla f(\boldsymbol{x})$ is large.
2. Hessian $\nabla^2 f(\boldsymbol{x})$ has a negative eigenvalue that is bounded away from 0.
3. Point $\boldsymbol{x}$ is near a local minimum.



Outline

1. Introduction
   - Motivation
   - Background
   - State-of-the-art
2. Proposed Approach
   - Perturbed Gradient Descent
   - Proof Sketch

State-of-the-art

- First Order Methods
- Second Order Methods
- Compromise between the two

Algorithm	Iterations	Oracle
Ge et al. [2015]	$O(\text{poly}(d/\epsilon))$	Gradient
Levy [2016]	$O(d^3 \cdot \text{poly}(1/\epsilon))$	Gradient
This Work	$O(\log^4(d)/\epsilon^2)$	Gradient
Agarwal et al. [2016]	$O(\log(d)/\epsilon^{7/4})$	Hessian-vector product
Carmon et al. [2016]	$O(\log(d)/\epsilon^{7/4})$	Hessian-vector product
Carmon and Duchi [2016]	$O(\log(d)/\epsilon^2)$	Hessian-vector product
Nesterov and Polyak [2006]	$O(1/\epsilon^{1.5})$	Hessian
Curtis et al. [2014]	$O(1/\epsilon^{1.5})$	Hessian

Outline

1. Introduction
   - Motivation
   - Background
   - State-of-the-art
2. Proposed Approach
   - Perturbed Gradient Descent
   - Proof Sketch

Perturbed Gradient Descent

Algorithm 2 Perturbed Gradient Descent: $\text{PGD}(\mathbf{x}_0, \ell, \rho, \epsilon, c, \delta, \Delta_f)$

$\chi \leftarrow 3\max\{\log(\frac{d\ell\Delta_f}{c\epsilon^2\delta}), 4\}$, $\eta \leftarrow \frac{c}{\ell}$, $r \leftarrow \frac{\sqrt{c}}{\chi^2} \cdot \frac{\epsilon}{\ell}$, $g_{\text{thres}} \leftarrow \frac{\sqrt{c}}{\chi^2} \cdot \epsilon$, $f_{\text{thres}} \leftarrow \frac{c}{\chi^3} \cdot \sqrt{\frac{\epsilon^3}{\rho}}$, $t_{\text{thres}} \leftarrow \frac{\chi}{c^2} \cdot \frac{\ell}{\sqrt{\rho\epsilon}}$

$t_{\text{noise}} \leftarrow -t_{\text{thres}} - 1$

for $t = 0, 1, \ldots$ **do**

  if $\|\nabla f(\mathbf{x}_t)\| \le g_{\text{thres}}$ and $t - t_{\text{noise}} > t_{\text{thres}}$ **then**

    $\tilde{\mathbf{x}}_t \leftarrow \mathbf{x}_t, \quad t_{\text{noise}} \leftarrow t$

    $\mathbf{x}_t \leftarrow \tilde{\mathbf{x}}_t + \xi_t, \qquad \xi_t$ uniformly $\sim \mathbb{B}_0(r)$

  if $t - t_{\text{noise}} = t_{\text{thres}}$ and $f(\mathbf{x}_t) - f(\tilde{\mathbf{x}}_{t_{\text{noise}}}) > -f_{\text{thres}}$ **then**

    return $\tilde{\mathbf{x}}_{t_{\text{noise}}}$

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

Perturbed Gradient Descent

Assumption A1. Function f(.) is both l-smooth and ρ-Hessian Lipschitz.

Theorem 3. *Assume that $f(\cdot)$ satisfies A1. Then there exists an absolute constant $c_{\max}$ such that, for any $\delta > 0, \epsilon \leq \frac{\ell^2}{\rho}, \Delta_f \geq f(\mathbf{x}_0) - f^\star$, and constant $c \leq c_{\max}$, PGD$(\mathbf{x}_0, \ell, \rho, \epsilon, c, \delta, \Delta_f)$ will output an ϵ-second-order stationary point, with probability $1-\delta$, and terminate in the following number of iterations:*

$$O\left(\frac{\ell(f(\mathbf{x}_0) - f^\star)}{\epsilon^2} \log^4\left(\frac{d\ell\Delta_f}{\epsilon^2\delta}\right)\right).$$

Outline

Exploiting Large Gradient or Negative Curvature

Second order stationary point has small gradient and here Hessian does not have a significant negative eigenvalue.
If it does not have these properties then:

- Gradient is large: $||\nabla f(x_t)|| \geq g_{thresh}$
- Around saddle point: $||\nabla f(x_t)|| \leq g_{thresh}||$ and $\lambda_{min}(\nabla^2 f(x_t)) \leq -\sqrt{\rho\epsilon}$

Lemma 9 (Gradient). *Assume that $f(\cdot)$ satisfies A1. Then for gradient descent with stepsize $\eta < \frac{1}{\ell}$, we have $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{\eta}{2}\|\nabla f(\mathbf{x}_t)\|^2$.*

Lemma 10 (Saddle). *(informal) Assume that $f(\cdot)$ satisfies A1, If $\mathbf{x}_t$ satisfies $\|\nabla f(\mathbf{x}_t)\| \leq g_{thres}$ and $\lambda_{\min}(\nabla^2 f(\mathbf{x}_t)) \leq -\sqrt{\rho\epsilon}$, then adding one perturbation step followed by t_{thres} steps of gradient descent, we have $f(\mathbf{x}_{t+t_{thres}}) - f(\mathbf{x}_t) \leq -f_{thres}$ with high probability.*

Escaping saddle points quickly

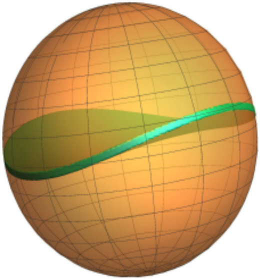
The perturbation ball can be divided into two regions:

- Escaping Region (X_{escape}) : Significant decrease in function value
- Stuck Region ($X_{stuck} = B_{\hat{x}}(r) - X_{escape}$) : Little decrease

Escaping saddle points quickly

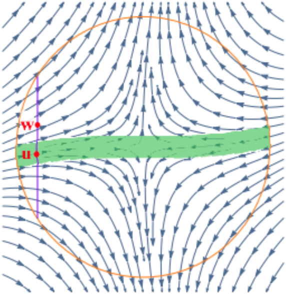
The perturbation ball can be divided into two regions:

- Escaping Region (X_{escape}) : Significant decrease in function value
- Stuck Region ($X_{stuck} = B_{\hat{x}}(r) - X_{escape}$) : Little decrease



Escaping saddle points quickly

Lemma 11. *(informal) Suppose* $\tilde{\mathbf{x}}$ *satisfies the precondition of Lemma 10, and let* $\mathbf{e}_1$ *be the smallest eigendirection of* $\nabla^2 f(\tilde{\mathbf{x}})$. *For any* $\delta \in (0, 1/3]$ *and any two points* $\mathbf{w}, \mathbf{u} \in \mathbb{B}_{\tilde{\mathbf{x}}}(r)$, *if* $\mathbf{w} - \mathbf{u} = \mu r \mathbf{e}_1$ *and* $\mu \geq \delta/(2\sqrt{d})$, *then at least one of* $\mathbf{w}, \mathbf{u}$ *is not in the stuck region* $\mathcal{X}_{stuck}$.



Summary

- Theoretical analysis of perturbed gradient descent algorithm (showed it is almost "dimension free")
- Showed that perturbed gradient descent can escape saddle points for free
- Novel characterization of geometry around saddle points
- Future Direction
  - Can similar techniques be applied to accelerated gradient descent