Optimization for Machine Learning

Lecture 16: Nonconvex Saddle-Point Problems

6.881: MIT

Suvrit Sra

Massachusetts Institute of Technology

April 22, 2021

$$\min_{x \in X} \max_{y \in Y} \phi(x, y)$$

Minmax: no global saddle points

Last time we saw the convex-concave case: easy, but in general?

$$\min_{x \in X} \max_{y \in Y} \phi(x, y)$$

Minmax: no global saddle points

Last time we saw the convex-concave case: easy, but in general?

When $\min_x \max_y \phi(x,y) \neq \max_y \min_x \phi(x,y)$, as is almost always the case with usual nonconvex problems, the sequence of play (min-max vs max-min) crucial.

$$\min_{x \in X} \max_{y \in Y} \phi(x,y)$$

Minmax: no global saddle points

Last time we saw the convex-concave case: easy, but in general?

When $\min_x \max_y \phi(x,y) \neq \max_y \min_x \phi(x,y)$, as is almost always the case with usual nonconvex problems, the sequence of play (min-max vs max-min) crucial.

Challenges: Define notions of optimality and study behavior of gradient based algorithms

$$\min_{x \in X} \max_{y \in Y} \phi(x,y)$$

Local saddle points

$$\min_{x \in X} \max_{y \in Y} \phi(x, y)$$

Local saddle points

Local saddle / local Nash equilibrium

$$\phi(x^*, y) \leq \phi(x^*, y^*) \leq \phi(x, y^*)$$

$$x \in X \cap B_\epsilon(x^*), \quad y \in Y \cap B_\epsilon(y^*)$$

Local saddle points

Local saddle / local Nash equilibrium

$$\phi(x^*, y) \leq \phi(x^*, y^*) \leq \phi(x, y^*)$$

$$x \in X \cap B_\epsilon(x^*), \quad y \in Y \cap B_\epsilon(y^*)$$

Thus, fixing x*, point y* is a *local maximizer* of $\phi(x^*, y)$, while fixing y*, point x* is a *local minimizer* of $\phi(x, y^*)$

Local saddle points

Local saddle / local Nash equilibrium

$$\phi(x^*, y) \leq \phi(x^*, y^*) \leq \phi(x, y^*)$$

$$x \in X \cap B_\epsilon(x^*), \quad y \in Y \cap B_\epsilon(y^*)$$

Thus, fixing x*, point y* is a *local maximizer* of $\phi(x^*, y)$, while fixing y*, point x* is a *local minimizer* of $\phi(x, y^*)$

Optimality conditions for local Nash equilibria

Local saddle points

Local saddle / local Nash equilibrium

$$\phi(x^*, y) \leq \phi(x^*, y^*) \leq \phi(x, y^*)$$

$$x \in X \cap B_\epsilon(x^*), \quad y \in Y \cap B_\epsilon(y^*)$$

Thus, fixing x*, point y* is a *local maximizer* of $\phi(x^*, y)$, while fixing y*, point x* is a *local minimizer* of $\phi(x, y^*)$

Optimality conditions for local Nash equilibria

First-order necessary conditions (ignoring boundary of sets X and Y):

$$\nabla_x \phi(x, y) = 0, \nabla_y \phi(x, y) = 0$$

Local saddle points

Local saddle / local Nash equilibrium

$$\phi(x^*, y) \leq \phi(x^*, y^*) \leq \phi(x, y^*)$$

$$x \in X \cap B_\epsilon(x^*), \quad y \in Y \cap B_\epsilon(y^*)$$

Thus, fixing x*, point y* is a *local maximizer* of $\phi(x^*, y)$, while fixing y*, point x* is a *local minimizer* of $\phi(x, y^*)$

Optimality conditions for local Nash equilibria

First-order necessary conditions (ignoring boundary of sets X and Y):

$$\nabla_x \phi(x, y) = 0, \nabla_y \phi(x, y) = 0$$

Similarly, **second order necessary**: $\nabla^2_{xx} \phi(x, y) \succeq 0, \nabla^2_{yy} \phi(x, y) \preceq 0$

Second order sufficient (assuming $\nabla \phi = 0$) : $\nabla^2_{xx} \phi(x, y) \succ 0, \nabla^2_{yy} \phi(x, y) \prec 0$

Local saddle points

Local saddle / local Nash equilibrium

$$\phi(x^*, y) \leq \phi(x^*, y^*) \leq \phi(x, y^*)$$

$$x \in X \cap B_\epsilon(x^*), \quad y \in Y \cap B_\epsilon(y^*)$$

Thus, fixing x*, point y* is a *local maximizer* of $\phi(x^*, y)$, while fixing

Question: Any troubles with this local notion?

Optimality conditions for local Nash equilibria

First-order necessary conditions (ignoring boundary of sets X and Y):

$$\nabla_x \phi(x, y) = 0, \nabla_y \phi(x, y) = 0$$

Similarly, **second order necessary**: $\nabla^2_{xx} \phi(x, y) \succeq 0, \nabla^2_{yy} \phi(x, y) \preceq 0$

Second order sufficient (assuming $\nabla \phi = 0$) : $\nabla^2_{xx} \phi(x, y) \succ 0, \nabla^2_{yy} \phi(x, y) \prec 0$

Local saddle points

> Local saddle also not truly suitable because of the sequential nature of min-max (Stackelberg games) rather than the simultaneous game.

Local saddle points

> Local saddle also not truly suitable because of the sequential nature of min-max (Stackelberg games) rather than the simultaneous game.

More crucially, even local Nash equilibrium may fail to exist

Local saddle points

> Local saddle also not truly suitable because of the sequential nature of min-max (Stackelberg games) rather than the simultaneous game.

More crucially, even local Nash equilibrium may fail to exist

Example

Local saddle points

> Local saddle also not truly suitable because of the sequential nature of min-max (Stackelberg games) rather than the simultaneous game.

More crucially, even local Nash equilibrium may fail to exist

Example

Let $\phi(x, y) = \sin(x + y)$. We have $\nabla\phi(x, y) = [\cos(x + y), \cos(x + y)]$. A local Nash equilibrium must satisfy stationarity, whereby we obtain $x + y = (k + \frac{1}{2})\pi,\ k \in \mathbb{Z}$.

Local saddle points

> Local saddle also not truly suitable because of the sequential nature of min-max (Stackelberg games) rather than the simultaneous game.

More crucially, even local Nash equilibrium may fail to exist

Example

Let $\phi(x, y) = \sin(x + y)$. We have $\nabla\phi(x, y) = [\cos(x + y), \cos(x + y)]$. A local Nash equilibrium must satisfy stationarity, whereby we obtain $x + y = (k+\frac{1}{2})\pi,\ k \in \mathbb{Z}$.

We can verify that for odd k, $\nabla^2_{xx}\phi(x, y) = \nabla^2_{yy}\phi(x, y) = 1$, while for even k, $\nabla^2_{xx}\phi(x, y) = \nabla^2_{yy}\phi(x, y) = -1$, thus ***violating SO-necessity***

min-max and max-min points

$$\min_{x \in X} \max_{y \in Y} \phi(x, y)$$

Minmax in Machine Learning

When $\min_x \max_y \phi(x, y) \neq \max_y \min_x \phi(x, y)$, as is almost always the case with usual nonconvex problems, **the sequence of play (min-max vs max-min) crucial.**

Minmax in Machine Learning

When $\min_x \max_y \phi(x, y) \neq \max_y \min_x \phi(x, y)$, as is almost always the case with usual nonconvex problems, **the sequence of play (min-max vs max-min) crucial.**

GANS: x models the generator, y the discriminator

Minmax in Machine Learning

When $\min_x \max_y \phi(x, y) \neq \max_y \min_x \phi(x, y)$, as is almost always the case with usual nonconvex problems, **the sequence of play (min-max vs max-min) crucial.**

GANS: x models the generator, y the discriminator

Adversarial training: x the params of a robust classifier, y the adv. attacks

Minmax in Machine Learning

When $\min_x \max_y \phi(x,y) \neq \max_y \min_x \phi(x,y)$, as is almost always the case with usual nonconvex problems, **the sequence of play (min-max vs max-min) crucial.**

GANS: x models the generator, y the discriminator

Adversarial training: x the params of a robust classifier, y the adv. attacks

Let us seek notions of optimality sensitive to order of play

Minmax for Nonconvex-Nonconcave

Main reference for this lecture

What is Local Optimality in Nonconvex-Nonconcave Minimax Optimization?

Chi Jin
University of California, Berkeley
chijin@cs.berkeley.edu

Praneeth Netrapalli
Microsoft Research, India
praneeth@microsoft.com

Michael I. Jordan
University of California, Berkeley
jordan@cs.berkeley.edu

Minmax for Nonconvex-Nonconcave

Main reference for this lecture

What is Local Optimality in Nonconvex-Nonconcave Minimax Optimization?

Chi Jin
University of California, Berkeley
chijin@cs.berkeley.edu

Praneeth Netrapalli
Microsoft Research, India
praneeth@microsoft.com

Michael I. Jordan
University of California, Berkeley
jordan@cs.berkeley.edu

Additional reference

Optimality and Stability in Non-convex Smooth Games

Guojun Zhang, Pascal Poupart and Yaoliang Yu

University of Waterloo
Vector Institute
{guojun.zhang, ppoupart, yaoliang.yu}@uwaterloo.ca

Min-max and max-min points

$$f(x) = \sup_{y \in Y} \phi(x, y), \quad g(y) = \inf_{x \in X} \phi(x, y)$$

Min-max and max-min points

$$f(x) = \sup_{y \in Y} \phi(x, y), \quad g(y) = \inf_{x \in X} \phi(x, y)$$

Global min-max point (x*,y*)

$$x^* \in \operatorname*{argmin}_{x \in X} f(x), \quad y^* \in \operatorname*{argmax}_{y \in Y} \phi(x^*, y)$$

$$\phi(x^*, y) \leq \phi(x^*, y^*) = f(x^*) \leq f(x), \quad \forall x \in X, y \in Y$$

Min-max and max-min points

$$f(x) = \sup_{y \in Y} \phi(x, y), \quad g(y) = \inf_{x \in X} \phi(x, y)$$

Global min-max point (x*,y*)

$$x^* \in \operatorname*{argmin}_{x \in X} f(x), \quad y^* \in \operatorname*{argmax}_{y \in Y} \phi(x^*, y)$$

$$\phi(x^*, y) \leq \phi(x^*, y^*) = f(x^*) \leq f(x), \quad \forall x \in X, y \in Y$$

Min-max and max-min points

$$f(x) = \sup_{y \in Y} \phi(x, y), \quad g(y) = \inf_{x \in X} \phi(x, y)$$

Global min-max point (x*,y*)

$$x^* \in \operatorname*{argmin}_{x \in X} f(x), \quad y^* \in \operatorname*{argmax}_{y \in Y} \phi(x^*, y)$$

$$\phi(x^*, y) \leq \phi(x^*, y^*) = f(x^*) \leq f(x), \quad \forall x \in X, y \in Y$$

Global max-min point (x*,y*)

$$y^* \in \operatorname*{argmax}_{y \in Y} g(y), \quad x^* \in \operatorname*{argmin}_{x \in X} \phi(x, y^*)$$

$$\phi(x, y^*) \geq \phi(x^*, y^*) = g(y^*) \geq g(x), \quad \forall x \in X, y \in Y$$

Min-max and max-min points

$$f(x) = \sup_{y \in Y} \phi(x, y), \quad g(y) = \inf_{x \in X} \phi(x, y)$$

Global min-max point (x*,y*)

$$x^* \in \operatorname*{argmin}_{x \in X} f(x), \quad y^* \in \operatorname*{argmax}_{y \in Y} \phi(x^*, y)$$

$$\phi(x^*, y) \leq \phi(x^*, y^*) = f(x^*) \leq f(x), \quad \forall x \in X, y \in Y$$

Global max-min point (x*,y*)

$$y^* \in \operatorname*{argmax}_{y \in Y} g(y), \quad x^* \in \operatorname*{argmin}_{x \in X} \phi(x, y^*)$$

$$\phi(x, y^*) \geq \phi(x^*, y^*) = g(y^*) \geq g(x), \quad \forall x \in X, y \in Y$$

Global min-max points

$$x^* \in \operatorname*{argmin}_{x \in X} f(x), \quad y^* \in \operatorname*{argmax}_{y \in Y} \phi(x^*, y)$$

$$\phi(x^*, y) \leq \phi(x^*, y^*) = f(x^*) \leq f(x), \quad \forall x \in X, y \in Y$$

Global min-max points

$$x^* \in \operatorname*{argmin}_{x \in X} f(x), \quad y^* \in \operatorname*{argmax}_{y \in Y} \phi(x^*, y)$$

$$\phi(x^*, y) \le \phi(x^*, y^*) = f(x^*) \le f(x), \quad \forall x \in X, y \in Y$$

Ordering matters!

First find x^* then find $y^* \mid x^*$

Global min-max points

$$x^* \in \operatorname*{argmin}_{x \in X} f(x), \quad y^* \in \operatorname*{argmax}_{y \in Y} \phi(x^*, y)$$

$$\phi(x^*, y) \le \phi(x^*, y^*) = f(x^*) \le f(x), \quad \forall x \in X, y \in Y$$

Ordering matters!

First find x^* then find $y^* \,|\, x^*$

Aka *Stackelberg game*, first leader plays then follower chooses best action (in contrast to the usual setting, where both *play simultaneously*)

Global min-max points

$$x^* \in \operatorname*{argmin}_{x \in X} f(x), \quad y^* \in \operatorname*{argmax}_{y \in Y} \phi(x^*, y)$$

$$\phi(x^*, y) \le \phi(x^*, y^*) = f(x^*) \le f(x), \quad \forall x \in X, y \in Y$$

Ordering matters!

First find x^* then find $y^* \mid x^*$

Aka *Stackelberg game*, first leader plays then follower chooses best action (in contrast to the usual setting, where both *play simultaneously*)

Saddle points (simultaneous)

$$\phi(x^*, y) \le \phi(x^*, y^*) \le \phi(x, y^*), \quad \forall x \in X, y \in Y$$

$$x^* \in \operatorname*{argmin}_{x \in X} f(x), \quad y^* \in \operatorname*{argmin}_{y \in Y} g(y)$$

Global min-max points

$$x^* \in \operatorname*{argmin}_{x \in X} f(x), \quad y^* \in \operatorname*{argmax}_{y \in Y} \phi(x^*, y)$$

$$\phi(x^*, y) \le \phi(x^*, y^*) = f(x^*) \le f(x), \quad \forall x \in X, y \in Y$$

Ordering matters!

First find x^* then find $y^* \mid x^*$

Aka *Stackelberg game*, first leader plays then follower chooses best action (in contrast to the usual setting, where both *play simultaneously*)

Saddle points (simultaneous)

$$\phi(x^*, y) \le \phi(x^*, y^*) \le \phi(x, y^*), \quad \forall x \in X, y \in Y$$

$$x^* \in \operatorname*{argmin}_{x \in X} f(x), \quad y^* \in \operatorname*{argmin}_{y \in Y} g(y)$$

Global min-max point is ≠ saddle point in general

Examples

Examples

Let $\phi(x, y) = xy$ and $X = Y = \mathbb{R}$. Then, global min-max points are the set $\{0\} \times \mathbb{R}$, while global max-min points are $\mathbb{R} \times \{0\}$.

Examples

Let $\phi(x, y) = xy$ and $X = Y = \mathbb{R}$. Then, global min-max points are the set $\{0\} \times \mathbb{R}$, while global max-min points are $\mathbb{R} \times \{0\}$.

Taking intersection, the unique (global) saddle point is (0,0). Moreover, note that not every $\hat{y} \in \arg\max_{y \in Y} \phi(x^*, y)$ forms a saddle pair with $x^* = 0$.

Examples

Let $\phi(x, y) = xy$ and $X = Y = \mathbb{R}$. Then, global min-max points are the set $\{0\} \times \mathbb{R}$, while global max-min points are $\mathbb{R} \times \{0\}$.

Taking intersection, the unique (global) saddle point is (0,0). Moreover, note that not every $\hat{y} \in \arg\max_{y \in Y} \phi(x^*, y)$ forms a saddle pair with $x^* = 0$.

(Turns out that "last iterate" of GD/MD does not converge to this unique saddle point)

Examples

Let $\phi(x, y) = xy$ and $X = Y = \mathbb{R}$. Then, global min-max points are the set $\{0\} \times \mathbb{R}$, while global max-min points are $\mathbb{R} \times \{0\}$.

Taking intersection, the unique (global) saddle point is (0,0). Moreover, note that not every $\hat{y} \in \arg\max_{y \in Y} \phi(x^*, y)$ forms a saddle pair with $x^* = 0$.

(Turns out that "last iterate" of GD/MD does not converge to this unique saddle point)

Exercise: Let $\phi(x, y) = ax^2 + by^2 + cxy$ with $a < 0,\ b < 0,\ c^2 \geq ab$. Verify that for this function only global min-max points exist, but no global max-min points (and thus also, no saddle points) exist.

Local min-max points

$$x^* \in \operatorname*{argmin}_{x \in X} f(x), \quad y^* \in \operatorname*{argmax}_{y \in Y} \phi(x^*, y)$$

$$\phi(x^*, y) \le \phi(x^*, y^*) = f(x^*) \le f(x), \quad \forall x \in X, y \in Y$$

Ordering matters!

First find x^* then find $y^* \mid x^*$

Turn the global min-max definition into a local one, here's one way

Local min-max points

$$x^* \in \operatorname*{argmin}_{x \in X} f(x), \quad y^* \in \operatorname*{argmax}_{y \in Y} \phi(x^*, y)$$

$$\phi(x^*, y) \le \phi(x^*, y^*) = f(x^*) \le f(x), \quad \forall x \in X, y \in Y$$

Ordering matters!

First find x^* then find $y^* \,|\, x^*$

Turn the global min-max definition into a local one, here's one way

$$\phi(x^*, y) \le \phi(x^*, y^*) \le \max_{y : \|y - y^*\| \le h(\epsilon)} \phi(x, y)$$

$$x \in X \cap B_\epsilon(x^*), \quad y \in Y \cap B_\epsilon(y^*)$$

Local min-max points

$$x^* \in \operatorname*{argmin}_{x \in X} f(x), \quad y^* \in \operatorname*{argmax}_{y \in Y} \phi(x^*, y)$$

$$\phi(x^*, y) \le \phi(x^*, y^*) = f(x^*) \le f(x), \quad \forall x \in X, y \in Y$$

Ordering matters!

First find x^* then find $y^* \mid x^*$

Turn the global min-max definition into a local one, here's one way

$$\phi(x^*, y) \le \phi(x^*, y^*) \le \max_{y : \|y - y^*\| \le h(\epsilon)} \phi(x, y)$$

$$x \in X \cap B_\epsilon(x^*), \quad y \in Y \cap B_\epsilon(y^*)$$

where $h(\epsilon) \to 0$ as $\epsilon \to 0$, and inequality holds for all $\epsilon \in (0, \epsilon_0]$

Local min-max points

$$x^* \in \operatorname*{argmin}_{x \in X} f(x), \quad y^* \in \operatorname*{argmax}_{y \in Y} \phi(x^*, y)$$

$$\phi(x^*, y) \le \phi(x^*, y^*) = f(x^*) \le f(x), \quad \forall x \in X, y \in Y$$

Ordering matters!

First find x^* then find $y^* \,|\, x^*$

Turn the global min-max definition into a local one, here's one way

$$\phi(x^*, y) \le \phi(x^*, y^*) \le \max_{y : \|y - y^*\| \le h(\epsilon)} \phi(x, y)$$

$$x \in X \cap B_\epsilon(x^*), \quad y \in Y \cap B_\epsilon(y^*)$$

where $h(\epsilon) \to 0$ as $\epsilon \to 0$, and inequality holds for all $\epsilon \in (0, \epsilon_0]$

Local min-max points

$$x^* \in \operatorname*{argmin}_{x \in X} f(x), \quad y^* \in \operatorname*{argmax}_{y \in Y} \phi(x^*, y)$$

$$\phi(x^*, y) \le \phi(x^*, y^*) = f(x^*) \le f(x), \quad \forall x \in X, y \in Y$$

Ordering matters!

First find x^* then find $y^* \mid x^*$

Turn the global min-max definition into a local one, here's one way

$$\phi(x^*, y) \le \phi(x^*, y^*) \le \max_{y : \|y - y^*\| \le h(\epsilon)} \phi(x, y)$$

$$x \in X \cap B_\epsilon(x^*), \quad y \in Y \cap B_\epsilon(y^*)$$

where $h(\epsilon) \to 0$ as $\epsilon \to 0$, and inequality holds for all $\epsilon \in (0, \epsilon_0]$

Lemma: For a continuous function ϕ, a point (x^*, y^*) is *local min-max* point if and only if y^* **is a local maximum** of the function $\phi(x^*, y)$ and there exists an ϵ_0 such that x^* **is a local minimum** of $f_\epsilon(x) := \max_{y : \|y - y^*\| \le \epsilon} \phi(x, y)$ for all $\epsilon \in (0, \epsilon_0]$

Local min-max points

$$x^* \in \operatorname*{argmin}_{x \in X} f(x), \quad y^* \in \operatorname*{argmax}_{y \in Y} \phi(x^*, y)$$

$$\phi(x^*, y) \le \phi(x^*, y^*) = f(x^*) \le f(x), \quad \forall x \in X, y \in Y$$

Ordering matters!

First find x^* then find $y^* \mid x^*$

Turn the global min-max definition into a local one, here's one way

$$\phi(x^*, y) \le \phi(x^*, y^*) \le \max_{y : \|y - y^*\| \le h(\epsilon)} \phi(x, y)$$

$$x \in X \cap B_\epsilon(x^*), \quad y \in Y \cap B_\epsilon(y^*)$$

where $h(\epsilon) \to 0$ as $\epsilon \to 0$, and inequality holds for all $\epsilon \in (0, \epsilon_0]$

Lemma: For a continuous function ϕ, a point (x^*, y^*) is *local min-max* point if and only if y^* **is a local maximum** of the function $\phi(x^*, y)$ and there exists an ϵ_0 such that x^* **is a local minimum** of $f_\epsilon(x) := \max_{y : \|y - y^*\| \le \epsilon} \phi(x, y)$ for all $\epsilon \in (0, \epsilon_0]$

Exercise: Prove that any local Nash equilibrium is local min-max but not vice versa.

Optimality conditions: local min-max

Local min-max

Local saddle point

Optimality conditions: local min-max

Local min-max

First-order necessary: Assuming (x^*,y^*) in the interior of $X \times Y$, we have

$$\nabla_x \phi(x,y) = 0, \nabla_y \phi(x,y) = 0$$

Local saddle point

Optimality conditions: local min-max

Local min-max

First-order necessary: Assuming (x^*,y^*) in the interior of $X \times Y$, we have

$$\nabla_x \phi(x,y) = 0, \nabla_y \phi(x,y) = 0$$

Local saddle point

FO-necessary: Assuming (x^*,y^*) lies in the interior of $X \times Y$, we have

$$\nabla_x \phi(x,y) = 0, \nabla_y \phi(x,y) = 0$$

Optimality conditions: local min-max

Local min-max

First-order necessary: Assuming (x^*,y^*) in the interior of $X \times Y$, we have

$$\nabla_x \phi(x,y) = 0, \nabla_y \phi(x,y) = 0$$

SO-necessary: $\nabla^2_{yy} \phi(x,y) \preceq 0$ and for

$$\nabla^2_{yx} \phi(x,y) \cdot v \in \text{colspan}(\nabla^2_{yy} \phi(x,y))$$

$$v^T[\nabla^2_{xx}\phi - \nabla^2_{xy}\phi(\nabla^2_{yy}\phi)^\dagger \nabla^2_{yx}\phi]v \geqslant 0$$

(psd over a subspace only)

Local saddle point

FO-necessary: Assuming (x^*,y^*) lies in the interior of $X \times Y$, we have

$$\nabla_x \phi(x,y) = 0, \nabla_y \phi(x,y) = 0$$

Optimality conditions: local min-max

Local min-max

First-order necessary: Assuming (x^*,y^*) in the interior of $X \times Y$, we have

$$\nabla_x \phi(x,y) = 0, \nabla_y \phi(x,y) = 0$$

SO-necessary: $\nabla^2_{yy}\phi(x,y) \preceq 0$ and for

$\nabla^2_{yx}\phi(x,y) \cdot v \in \text{colspan}(\nabla^2_{yy}\phi(x,y))$

$v^T[\nabla^2_{xx}\phi - \nabla^2_{xy}\phi(\nabla^2_{yy}\phi)^\dagger \nabla^2_{yx}\phi]v \geqslant 0$

(psd over a subspace only)

Local saddle point

FO-necessary: Assuming (x^*,y^*) lies in the interior of $X \times Y$, we have

$$\nabla_x \phi(x,y) = 0, \nabla_y \phi(x,y) = 0$$

SO-necessary: $\nabla^2_{xx}\phi(x,y) \succeq 0$,

$\nabla^2_{yy}\phi(x,y) \preceq 0$

Optimality conditions: local min-max

Local min-max

First-order necessary: Assuming (x^*,y^*) in the interior of $X \times Y$, we have

$$\nabla_x \phi(x,y) = 0, \nabla_y \phi(x,y) = 0$$

SO-necessary: $\nabla^2_{yy}\phi(x,y) \preceq 0$ and for $\nabla^2_{yx}\phi(x,y) \cdot v \in \text{colspan}(\nabla^2_{yy}\phi(x,y))$

$$v^T[\nabla^2_{xx}\phi - \nabla^2_{xy}\phi(\nabla^2_{yy}\phi)^\dagger \nabla^2_{yx}\phi]v \geqslant 0$$

(psd over a subspace only)

SO-sufficient: (assuming FON):

$$\nabla^2_{yy}\phi(x,y) \prec 0,$$

$$\nabla^2_{xx}\phi - \nabla^2_{xy}\phi(\nabla^2_{yy}\phi)^{-1}\nabla^2_{yx}\phi \succ 0$$

Local saddle point

FO-necessary: Assuming (x^*,y^*) lies in the interior of $X \times Y$, we have

$$\nabla_x \phi(x,y) = 0, \nabla_y \phi(x,y) = 0$$

SO-necessary: $\nabla^2_{xx}\phi(x,y) \succeq 0$, $\nabla^2_{yy}\phi(x,y) \preceq 0$

Optimality conditions: local min-max

Local min-max

First-order necessary: Assuming (x^*,y^*) in the interior of $X \times Y$, we have

$$\nabla_x \phi(x,y) = 0, \nabla_y \phi(x,y) = 0$$

SO-necessary: $\nabla^2_{yy}\phi(x,y) \preceq 0$ and for $\nabla^2_{yx}\phi(x,y) \cdot v \in \text{colspan}(\nabla^2_{yy}\phi(x,y))$

$$v^T[\nabla^2_{xx}\phi - \nabla^2_{xy}\phi(\nabla^2_{yy}\phi)^\dagger \nabla^2_{yx}\phi]v \geqslant 0$$

(psd over a subspace only)

SO-sufficient: (assuming FON):

$$\nabla^2_{yy}\phi(x,y) \prec 0,$$
$$\nabla^2_{xx}\phi - \nabla^2_{xy}\phi(\nabla^2_{yy}\phi)^{-1}\nabla^2_{yx}\phi \succ 0$$

Local saddle point

FO-necessary: Assuming (x^*,y^*) lies in the interior of $X \times Y$, we have

$$\nabla_x \phi(x,y) = 0, \nabla_y \phi(x,y) = 0$$

SO-necessary: $\nabla^2_{xx}\phi(x,y) \succeq 0$, $\nabla^2_{yy}\phi(x,y) \preceq 0$

SO-sufficient (assuming FON):

$$\nabla^2_{xx}\phi(x,y) \succ 0, \nabla^2_{yy}\phi(x,y) \prec 0$$

Local min-max: some problems

Local min-max: some problems

Bad news: Global minimax point could fail to be local minimax (easiest way to construct these by having global minimax points violate local necessary conditions)

Local min-max: some problems

Bad news: Global minimax point could fail to be local minimax (easiest way to construct these by having global minimax points violate local necessary conditions)

Example: Try $\phi(x, y) = (2/10)xy - \cos(y)$ on the region $[-1,1] \times [-2\pi,2\pi]$ and verify that $(0, -\pi)$, $(0,\pi)$ are two global solutions but the gradients are nonzero!

Local min-max: some problems

Bad news: Global minimax point could fail to be local minimax (easiest way to construct these by having global minimax points violate local necessary conditions)

Example: Try $\phi(x, y) = (2/10)xy - \cos(y)$ on the region $[-1,1] \times [-2\pi,2\pi]$ and verify that $(0, -\pi)$, $(0,\pi)$ are two global solutions but the gradients are nonzero!

Bad news 2: Local minimax points may also fail to exist

Local min-max: some problems

Bad news: Global minimax point could fail to be local minimax (easiest way to construct these by having global minimax points violate local necessary conditions)

Example: Try $\phi(x, y) = (2/10)xy - \cos(y)$ on the region $[-1,1] \times [-2\pi,2\pi]$ and verify that $(0, -\pi)$, $(0,\pi)$ are two global solutions but the gradients are nonzero!

Bad news 2: Local minimax points may also fail to exist

Example: Try $\phi(x, y) = y^2 - 2xy$ on the region $[-1,1] \times [-1,1]$
(observe: not too surprising, in light of *global minimax $\not\subset$ local minimax*)

Local min-max: some problems

Bad news: Global minimax point could fail to be local minimax (easiest way to construct these by having global minimax points violate local necessary conditions)

Example: Try $\phi(x, y) = (2/10)xy - \cos(y)$ on the region $[-1,1] \times [-2\pi,2\pi]$ and verify that $(0, -\pi)$, $(0,\pi)$ are two global solutions but the gradients are nonzero!

Bad news 2: Local minimax points may also fail to exist

Example: Try $\phi(x, y) = y^2 - 2xy$ on the region $[-1,1] \times [-1,1]$
(observe: not too surprising, in light of *global minimax $\not\subset$ local minimax*)

Explore: What could be other alternatives?

Local min-max

Local min-max

Optimality and Stability in Non-Convex-Non-Concave Min-Max Optimization

Guojun Zhang, Pascal Poupart and Yaoliang Yu

Local min-max

Optimality and Stability in Non-Convex-Non-Concave Min-Max Optimization

Guojun Zhang, Pascal Poupart and Yaoliang Yu

- Offers more recent, and more precise notion of local min-max points (called "generalized local points") that includes existing local min-max, local saddles, etc. as special cases.

Local min-max

Optimality and Stability in Non-Convex-Non-Concave Min-Max Optimization

Guojun Zhang, Pascal Poupart and Yaoliang Yu

- Offers more recent, and more precise notion of local min-max points (called "generalized local points") that includes existing local min-max, local saddles, etc. as special cases.

- Their first and second order necessary conditions are more careful

Local min-max

Optimality and Stability in Non-Convex-Non-Concave Min-Max Optimization

Guojun Zhang, Pascal Poupart and Yaoliang Yu

- Offers more recent, and more precise notion of local min-max points (called "generalized local points") that includes existing local min-max, local saddles, etc. as special cases.

- Their first and second order necessary conditions are more careful

- They give a complete picture of quadratic saddle point problems, as well as stability analysis of various algorithms

Algorithms

Gradient Descent-Ascent (GDA)

Recall, Mirror-Descent for Saddle Points from last time. Use it with $\|\cdot\| = \|\cdot\|_2$ and $\omega(z) = \frac{1}{2} z^T z$. Moreover, let us allow different step-sizes for x and y

Gradient Descent-Ascent (GDA)

Recall, Mirror-Descent for Saddle Points from last time. Use it with $\|\cdot\| = \|\cdot\|_2$ and $\omega(z) = \frac{1}{2} z^T z$. Moreover, let us allow different step-sizes for x and y

Algorithm 1 Gradient Descent Ascent (γ-GDA)

Input: $(\mathbf{x}_0, \mathbf{y}_0)$, step size η, ratio γ.

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

$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - (\eta/\gamma)\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)$.

$\mathbf{y}_{t+1} \leftarrow \mathbf{y}_t + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t)$.

(Remark: when η is small, as if we're alternating between 1 step of GD and γ steps of Gradient Ascent; the 'f' above is our 'ϕ')

Gradient Descent-Ascent (GDA)

Recall, Mirror-Descent for Saddle Points from last time. Use it with $\|\cdot\| = \|\cdot\|_2$ and $\omega(z) = \frac{1}{2} z^T z$. Moreover, let us allow different step-sizes for x and y

Algorithm 1 Gradient Descent Ascent (γ-GDA)

Input: $(\mathbf{x}_0, \mathbf{y}_0)$, step size η, ratio γ.

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

$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - (\eta/\gamma)\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)$.

$\mathbf{y}_{t+1} \leftarrow \mathbf{y}_t + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t)$.

(Remark: when η is small, as if we're alternating between 1 step of GD and γ steps of Gradient Ascent; the 'f' above is our 'ϕ')

Theorem. For any fixed γ, for any twice differentiable ϕ, the set of local Nash equilibria is a subset of stable points of γ-GDA, but not all stable points of γ-GDA are LNE.

Gradient Descent-Ascent (GDA)

Recall, Mirror-Descent for Saddle Points from last time. Use it with $\|\cdot\| = \|\cdot\|_2$ and $\omega(z) = \frac{1}{2} z^T z$. Moreover, let us allow different step-sizes for x and y

Algorithm 1 Gradient Descent Ascent (γ-GDA)

Input: $(\mathbf{x}_0, \mathbf{y}_0)$, step size η, ratio γ.

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

$$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - (\eta/\gamma)\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t).$$

$$\mathbf{y}_{t+1} \leftarrow \mathbf{y}_t + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t).$$

(Remark: when η is small, as if we're alternating between 1 step of GD and γ steps of Gradient Ascent; the 'f' above is our 'ϕ')

Theorem. For any fixed γ, for any twice differentiable ϕ, the set of local Nash equilibria is a subset of stable points of γ-GDA, but not all stable points of γ-GDA are LNE.

Theorem. For any fixed γ, there exists twice differentiable ϕ, s.t. set of local min-max points is not a subset of stable points of γ-GDA. Converse also true.

GD with max-oracle

$$\min_{x \in X} \left[f(x) = \sup_{y \in Y} \phi(x, y) \right]$$

GD with max-oracle

$$\min_{x \in X} \left[f(x) = \sup_{y \in Y} \phi(x, y) \right]$$

Algorithm 2 Gradient Descent with Max-oracle

Input: $\mathbf{x}_0$, step size η.

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

find $\mathbf{y}_t$ so that $f(\mathbf{x}_t, \mathbf{y}_t) \geq \max_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}) - \epsilon$.

$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)$.

Pick t uniformly at random from $\{0, \cdots, T\}$.

return $\bar{\mathbf{x}} \leftarrow \mathbf{x}_t$.

GD with max-oracle

$$\min_{x \in X} \left[f(x) = \sup_{y \in Y} \phi(x, y) \right]$$

Algorithm 2 Gradient Descent with Max-oracle

Input: $\mathbf{x}_0$, step size η.

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

find $\mathbf{y}_t$ so that $f(\mathbf{x}_t, \mathbf{y}_t) \geq \max_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}) - \epsilon$.

$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)$.

Pick t uniformly at random from $\{0, \cdots, T\}$.

return $\bar{\mathbf{x}} \leftarrow \mathbf{x}_t$.

This method can indeed converge to approximate stationary points of f(x)

GD with max-oracle

$$\min_{x \in X} \left[f(x) = \sup_{y \in Y} \phi(x, y) \right]$$

Algorithm 2 Gradient Descent with Max-oracle

Input: $\mathbf{x}_0$, step size η.

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

find $\mathbf{y}_t$ so that $f(\mathbf{x}_t, \mathbf{y}_t) \geq \max_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}) - \epsilon$.

$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)$.

Pick t uniformly at random from $\{0, \cdots, T\}$.

return $\bar{\mathbf{x}} \leftarrow \mathbf{x}_t$.

This method can indeed converge to approximate stationary points of f(x)

Key task: How to define stationarity here, and take advantage of it?

GD-Max: Elementary theory

GD-Max: Elementary theory

Observation: Assume ϕ is L-smooth. Then $f(x)+\frac{L}{2}\|x\|^2$ is convex.

GD-Max: Elementary theory

Observation: Assume ϕ is L-smooth. Then $f(x)+\frac{L}{2}\|x\|^2$ is convex.

For such *weakly convex functions*, stationarity can be measured by norm of the gradient of its *Moreau envelope*

$$f_\lambda(x) := \min_{x'} f(x) + \frac{1}{2\lambda}\|x - x'\|^2$$

GD-Max: Elementary theory

Observation: Assume ϕ is L-smooth. Then $f(x)+\frac{L}{2}\|x\|^2$ is convex.

For such *weakly convex functions*, stationarity can be measured by norm of the gradient of its *Moreau envelope*

$$f_\lambda(x) := \min_{x'} f(x) + \frac{1}{2\lambda}\|x - x'\|^2$$

Theorem. Let f be L-weakly convex. Let λ < 1/L, and let $\bar{x} = \arg\min_x f_\lambda(x)$. Then,

$\|\nabla f_\lambda(x)\| \le \epsilon$ implies $\|x - \bar{x}\| = \lambda\epsilon$, and $\min_{g \in \partial f(x)} \|g\| \le \epsilon$

GD-Max: Elementary theory

Observation: Assume ϕ is L-smooth. Then $f(x)+\frac{L}{2}\|x\|^2$ is convex.

For such *weakly convex functions*, stationarity can be measured by norm of the gradient of its *Moreau envelope*

$$f_\lambda(x) := \min_{x'} f(x) + \frac{1}{2\lambda}\|x - x'\|^2$$

Theorem. Let f be L-weakly convex. Let λ < 1/L, and let $\bar{x} = \arg\min_x f_\lambda(x)$. Then,

$\|\nabla f_\lambda(x)\| \le \epsilon$ implies $\|x - \bar{x}\| = \lambda\epsilon$, and $\min_{g \in \partial f(x)} \|g\| \le \epsilon$

Theorem. Let ϕ be G-Lipschitz and L-smooth, then GD-Max with step-size $\eta = 1/\sqrt{T+1}$ satisfies $\mathbb{E}[\|\nabla f_{2L}(x)\|^2] \le O(G^2/\sqrt{T}) + O(L\epsilon)$

Other methods

- GDA augmented with momentum
- Extragradient (special case of Mirror Prox)
- Optimistic gradient descent (OGD)
- Two time-scale variants of these (GDA we saw already)

Other methods

- GDA augmented with momentum
- Extragradient (special case of Mirror Prox)
- Optimistic gradient descent (OGD)
- Two time-scale variants of these (GDA we saw already)

Recall, vector field for gradient-based methods

$$F(z) = [\nabla_x \phi(x,y), -\nabla_y \phi(x,y)]$$

Other methods

- GDA augmented with momentum
- Extragradient (special case of Mirror Prox)
- Optimistic gradient descent (OGD)
- Two time-scale variants of these (GDA we saw already)

Recall, vector field for gradient-based methods

$$F(z) = [\nabla_x \phi(x,y), -\nabla_y \phi(x,y)]$$

GDA:

$$z_{t+1} = z_t - \eta F(z_t)$$

Other methods

- GDA augmented with momentum
- Extragradient (special case of Mirror Prox)
- Optimistic gradient descent (OGD)
- Two time-scale variants of these (GDA we saw already)

Recall, vector field for gradient-based methods

$$F(z) = [\nabla_x \phi(x,y), -\nabla_y \phi(x,y)]$$

GDA: $z_{t+1} = z_t - \eta F(z_t)$

GDA+momentum: $z_{t+1} = z_t - \eta F(z_t) + \beta(z_t - z_{t-1})$

Other methods

- GDA augmented with momentum
- Extragradient (special case of Mirror Prox)
- Optimistic gradient descent (OGD)
- Two time-scale variants of these (GDA we saw already)

Recall, vector field for gradient-based methods

$$F(z) = [\nabla_x \phi(x,y), -\nabla_y \phi(x,y)]$$

GDA: $z_{t+1} = z_t - \eta F(z_t)$

GDA+momentum: $z_{t+1} = z_t - \eta F(z_t) + \beta(z_t - z_{t-1})$

Nesterov: $z_{t+1} = z'_t - \eta F(z'_t),\ z'_t = z_t + \beta(z_t - z_{t-1})$

Other methods

- GDA augmented with momentum
- Extragradient (special case of Mirror Prox)
- Optimistic gradient descent (OGD)
- Two time-scale variants of these (GDA we saw already)

Recall, vector field for gradient-based methods

$$F(z) = [\nabla_x \phi(x,y), -\nabla_y \phi(x,y)]$$

GDA: $z_{t+1} = z_t - \eta F(z_t)$

GDA+momentum: $z_{t+1} = z_t - \eta F(z_t) + \beta(z_t - z_{t-1})$

Nesterov: $z_{t+1} = z'_t - \eta F(z'_t),\ z'_t = z_t + \beta(z_t - z_{t-1})$

Extragradient: $z_{t+1} = z_t - \eta F(z_{t+1/2}),\ z_{t+1/2} = z_t - \eta F(z_t)$

Other methods

- GDA augmented with momentum
- Extragradient (special case of Mirror Prox)
- Optimistic gradient descent (OGD)
- Two time-scale variants of these (GDA we saw already)

Recall, vector field for gradient-based methods

$$F(z) = [\nabla_x \phi(x,y), -\nabla_y \phi(x,y)]$$

GDA: $z_{t+1} = z_t - \eta F(z_t)$

GDA+momentum: $z_{t+1} = z_t - \eta F(z_t) + \beta(z_t - z_{t-1})$

Nesterov: $z_{t+1} = z'_t - \eta F(z'_t),\ z'_t = z_t + \beta(z_t - z_{t-1})$

Extragradient: $z_{t+1} = z_t - \eta F(z_{t+1/2}),\ z_{t+1/2} = z_t - \eta F(z_t)$

Optimistic GD: $z_{t+1} = z_t - k\eta F(z_t) + \eta F(z_{t-1})$

Some results on these methods

Some results on these methods

- ☑ Near some local saddles, GDA and momentum methods ***never converge***, even with 2 time-scale modification

Some results on these methods

- ☑ Near some local saddles, GDA and momentum methods ***never converge***, even with 2 time-scale modification

Some results on these methods

- ☑ Near some local saddles, GDA and momentum methods ***never converge***, even with 2 time-scale modification

- ☑ EG and OGD can converge near any local saddle point. Their convergence regions superset of GDA regions; ***strictly more stable***

Some results on these methods

- ☑ Near some local saddles, GDA and momentum methods ***never converge***, even with 2 time-scale modification

- ☑ EG and OGD can converge near any local saddle point. Their convergence regions superset of GDA regions; ***strictly more stable***

Some results on these methods

- ☑ Near some local saddles, GDA and momentum methods ***never converge***, even with 2 time-scale modification

- ☑ EG and OGD can converge near any local saddle point. Their convergence regions superset of GDA regions; ***strictly more stable***

- ☑ Under 2nd order sufficient conditions, ***all these methods converge*** using their 2 time-scale versions

Some results on these methods

- ☑ Near some local saddles, GDA and momentum methods ***never converge***, even with 2 time-scale modification

- ☑ EG and OGD can converge near any local saddle point. Their convergence regions superset of GDA regions; ***strictly more stable***

- ☑ Under 2nd order sufficient conditions, ***all these methods converge*** using their 2 time-scale versions

Some results on these methods

- ☑ Near some local saddles, GDA and momentum methods ***never converge***, even with 2 time-scale modification

- ☑ EG and OGD can converge near any local saddle point. Their convergence regions superset of GDA regions; ***strictly more stable***

- ☑ Under 2nd order sufficient conditions, ***all these methods converge*** using their 2 time-scale versions

- ☑ There exist local min-max points for which no 1 time-scale method would converge to. With 2-time-scale version, only EG and OGD converge to such local min-max points, but cannot take their step sizes to be arbitrarily small.

Some results on these methods

- ☑ Near some local saddles, GDA and momentum methods ***never converge***, even with 2 time-scale modification

- ☑ EG and OGD can converge near any local saddle point. Their convergence regions superset of GDA regions; ***strictly more stable***

- ☑ Under 2nd order sufficient conditions, ***all these methods converge*** using their 2 time-scale versions

- ☑ There exist local min-max points for which no 1 time-scale method would converge to. With 2-time-scale version, only EG and OGD converge to such local min-max points, but cannot take their step sizes to be arbitrarily small.

Note: The last point challenges the conventional idea of using continuous approximation / ODEs to understand these methods, and justifies use of two-time-scale methods in the nonconvex-nonconcave setting.