L10 – Week 5
Introduction to Min-max Optimization

CS 295 Optimization for Machine Learning

Ioannis Panageas

GANs

In Generative Adversarial Networks (GANs) one would like to solve

$$\min_{\theta} \max_{w} \mathbb{E}_{x\sim Q}[D_w(x)] - \mathbb{E}_{z\sim F}[D_w(G_\theta(z))]$$

- D_w is the discriminator, G_θ the generator.
- Q is the data distribution, F say Gaussian (noise)
- D_w might (or not) capture the probability to classify data point as true!

- The aforementioned min-max problem is really hard! Many challenges!

GANs (Goodfellow et al.)

In their seminal paper, Goodfellow et al. defined the following min-max problem:

$$\min_{\theta} \max_{w} \mathbb{E}_{x \sim p_{\text{data}}}[\log D_w(x)] + \mathbb{E}_{z \sim p_{\text{noise}}}[\log(1 - D_w(G_\theta(z)))]$$

- D_w is the discriminator, G_θ the generator.
- p_{data} is the data distribution, p_{noise} say Gaussian (noise).
- D_w captures the probability to classify data point as true!
- D is trying to maximize prob to assign correct label to both samples from data and from G.

GANs (Goodfellow et al.)

In their seminal paper, Goodfellow et al. defined the following min-max problem:

$$\min_{\theta} \max_{w} \mathbb{E}_{x \sim p_{\text{data}}}[\log D_w(x)] + \mathbb{E}_{z \sim p_{\text{noise}}}[\log(1 - D_w(G_\theta(z)))]$$

Lemma (Optimality). *For G fixed, the optimal discriminator D has density*

$$D_{w^*}(x) = \frac{p_{data}(x)}{p_{data}(x) + p_G(x)},$$

where p_G is the implicit distribution of the Generator over the data.

GANs (Goodfellow et al.)

In their seminal paper, Goodfellow et al. defined the following min-max problem:

$$\min_{\theta} \max_{w} \mathbb{E}_{x \sim p_{\text{data}}}[\log D_w(x)] + \mathbb{E}_{z \sim p_{\text{noise}}}[\log(1 - D_w(G_\theta(z)))]$$

Lemma (Optimality). *For G fixed, the optimal discriminator D has density*

$$D_{w^*}(x) = \frac{p_{data}(x)}{p_{data}(x) + p_G(x)},$$

where p_G is the implicit distribution of the Generator over the data.

Proof. For fixed G, D is trying to maximize

$$\int_x \log D(x) p_{\text{data}}(x) dx + \int_z \log(1 - D(G(z)) p_{\text{noise}}(z) dz.$$

GANs (Goodfellow et al.)

Proof. For fixed G, D is trying to maximize

$$\int_x \log D(x) p_{\text{data}}(x) dx + \int_z \log(1 - D(G(z)) p_{\text{noise}}(z) dz.$$

The above is nothing but (set $x = G(z)$)

$$\int_x \log D(x) p_{\text{data}}(x) dx + \int_x \log(1 - D(x) p_G(x) dx.$$

GANs (Goodfellow et al.)

Proof. For fixed G, D is trying to maximize

$$\int_x \log D(x) p_{\text{data}}(x) dx + \int_z \log(1 - D(G(z)) p_{\text{noise}}(z) dz.$$

The above is nothing but (set $x = G(z)$)

$$\int_x \log D(x) p_{\text{data}}(x) dx + \int_x \log(1 - D(x) p_G(x) dx.$$

Finally, observe that function

$$f(y) = a \log y + b \log(1 - y)$$

achieves maximum at $\frac{a}{a+b}$.

GANs (Goodfellow et al.)

Proof. For fixed G, D is trying to maximize

$$\int_x \log D(x) p_{\text{data}}(x) dx + \int_z \log(1 - D(G(z)) p_{\text{noise}}(z) dz.$$

The above is nothing but (set $x = G(z)$)

$$\int_x \log D(x) p_{\text{data}}(x) dx + \int_x \log(1 - D(x) p_G(x) dx.$$

F

Define cost function $C(G)$

$$C(G) := \mathbb{E}_{x \sim p_{\text{data}}} \left[\log \frac{p_{\text{data}}}{p_{\text{data}} + p_G} \right] + \mathbb{E}_{x \sim p_G} \left[\log \frac{p_G}{p_{\text{data}} + p_G} \right].$$

achieves maximum at $\frac{\quad}{a+b}$.

GANs (Goodfellow et al.)

Theorem (Global solution)**.** *The global minimum of $C(G)$ is achieved if and only if*

$$p_G = p_{data}.$$

Proof.

GANs (Goodfellow et al.)

Theorem (Global solution). *The global minimum of $C(G)$ is achieved if and only if*

$$p_G = p_{data}.$$

Proof. Observe that for $p_{\text{data}} = p_G$ we get that $C(G) = -\log 4$.

Quick recap $\mathrm{KL}(p||q) = \mathbb{E}_{x\sim p}\left[\log \frac{p(x)}{q(x)}\right]$ is non-negative!

GANs (Goodfellow et al.)

Theorem (Global solution). *The global minimum of $C(G)$ is achieved if and only if*

$$p_G = p_{data}.$$

Proof. Observe that for $p_{\text{data}} = p_G$ we get that $C(G) = -\log 4$.

Quick recap $\mathrm{KL}(p||q) = \mathbb{E}_{x\sim p}\left[\log \frac{p(x)}{q(x)}\right]$ is non-negative!

Finally observe that

$$C(G) = -\log 4 + \mathrm{KL}\left(p_{\text{data}}\Big|\Big|\frac{p_{\text{data}} + p_G}{2}\right) + \mathrm{KL}\left(p_G\Big|\Big|\frac{p_{\text{data}} + p_G}{2}\right).$$

Min-max Optimization

GANs motivate the study of min-max optimization (in general harder than minimization), i.e., for some continuous function f we want to solve

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

Remarks

- Domains are typically compact.
- In general the above problem might not have a solution.
- There are guarantees when domains are compact and f is convex-concave.

Minimax Theorem

Theorem (Minimax by John von Neumann). *Let $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ be compact convex sets. If f is a continuous function that is convex-concave it holds*

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x,y) = \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} f(x,y)$$

Remarks

- Many applications, especially in Game Theory.
- If $f = x^T A y$, and the domains are Δ_n, Δ_m it captures classic zero sum games
- The above is the value of the game.
- Note that It is always true (min-max inequality):

$$\boxed{\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(x,y) \geq \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} f(x,y)}$$

Minimax Theorem

Theorem (Minimax by John von Neumann). *Let $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ be compact convex sets. If f is a continuous function that is convex-concave it holds*

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y) = \max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} f(x,y)$$

Remarks

- Many applications, especially in Game Theory.
- If $f = x^T Ay$, and the domains are Δ_n, Δ_m it captures classic zero sum games
- The above is the value of the game.
- Note that It is always true (min-max inequality):

$$\boxed{\inf_{x\in\mathcal{X}} \sup_{y\in\mathcal{Y}} f(x,y) \geq \sup_{y\in\mathcal{Y}} \mathrm{ir}}$$

Define $g(z) \triangleq \inf_{w\in W} f(z,w)$.

$$\forall w, \forall z, g(z) \leq f(z,w)$$

$$\Longrightarrow \forall w, \sup_z g(z) \leq \sup_z f(z,w)$$

$$\Longrightarrow \sup_z g(z) \leq \inf_w \sup_z f(z,w)$$

$$\Longrightarrow \sup_z \inf_w f(z,w) \leq \inf_w \sup_z f(z,w)$$

Minimax Theorem

Theorem (Minimax by John von Neumann). *Let $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ be compact convex sets. If f is a continuous function that is convex-concave it holds*

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y) = \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} f(x, y)$$

Proof. Let's use no-regret learning for both "players"!

Online Gradient Descent (Recap)

Definition (Online Gradient Descent). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex function, differentiable and L-Lipschitz in some compact convex set $\mathcal{X}$ of diameter D. Online GD is defined:*

Initialize at some x_0.
For t:=1 to T do

1. Choose x_t and observe $\ell_t(x_t)$.
2. $y_t = x_t - \alpha_t \nabla \ell_t(x_t)$.
3. $x_{t+1} = \Pi_{\mathcal{X}}(y_t)$.

Regret: $\frac{1}{T}\left(\sum_{t=1}^{T} \ell_t(x_t) - \min_x \sum_{t=1}^{T} \ell_t(x)\right)$.

Analysis of Online GD for L-Lipschitz (Recap)

Theorem (Online Gradient Descent). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex function, differentiable and L-Lipschitz in some compact convex set $\mathcal{X}$ of diameter D. It holds*

$$\left(\frac{1}{T}\sum_{t=1}^{T}\ell_t(x_t) - \min_x \sum_{t=1}^{T}\ell_t(x)\right) \leq \frac{3}{2}\frac{LD}{\sqrt{T}},$$

with appropriately choosing $\alpha = \frac{D}{L\sqrt{t}}$.

Remarks:

- If we want error ϵ, we need $T = \Theta\left(\frac{L^2D^2}{\epsilon^2}\right)$ iterations (same as GD for L-Lipschitz).

Minimax Theorem

Theorem (Minimax by John von Neumann). *Let $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ be compact convex sets. If f is a continuous function that is convex-concave it holds*

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y) = \max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} f(x,y)$$

Proof. Let's use no-regret learning for both "players"!

Let $x_1, ..., x_T$ and $y_1, ..., y_T$ be the iterates as advised by some no-regret algorithm and define $\hat{x} = \frac{1}{T}\sum_{i=1}^{T} x_i$ and $\hat{y} = \frac{1}{T}\sum_{i=1}^{T} y_i$ and $T = \Theta(\frac{1}{\epsilon^2})$.

Choose any x, then from the no-regret property for x we get that

$$\frac{1}{T}\sum_t f(x_t, y_t) \leq \frac{1}{T}\sum_t f(x, y_t) + \epsilon$$

Minimax Theorem

Theorem (Minimax by John von Neumann). *Let* $\mathcal{X} \subset \mathbb{R}^n$ *and* $\mathcal{Y} \subset \mathbb{R}^m$ *be compact convex sets. If* f *is a continuous function that is convex-concave it holds*

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y) = \max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} f(x,y)$$

Proof. Let's use no-regret learning for both "players"!

Let $x_1, ..., x_T$ and $y_1, ..., y_T$ be the iterates as advised by some no-regret algorithm and define $\hat{x} = \frac{1}{T}\sum_{i=1}^{T} x_i$ and $\hat{y} = \frac{1}{T}\sum_{i=1}^{T} y_i$ and $T = \Theta(\frac{1}{\epsilon^2})$.

Choose any x, then from the no-regret property for x we get that

$$\begin{aligned}\tfrac{1}{T}\textstyle\sum_t f(x_t, y_t) &\le \tfrac{1}{T}\textstyle\sum_t f(x, y_t) + \epsilon \\ &\le f(x, \hat{y}) + \epsilon \text{ by concavity.}\end{aligned}$$

Minimax Theorem

Proof cont.

Choose any y, then from the no-regret property for y we get that

$$\frac{1}{T}\sum_t f(x_t, y_t) \geq \frac{1}{T}\sum_t f(x_t, y) - \epsilon$$

$$\geq f(\hat{x}, y) - \epsilon \text{ by convexity.}$$

Minimax Theorem

Proof cont.

Choose any y, then from the no-regret property for y we get that

$$\begin{aligned}\tfrac{1}{T}\textstyle\sum_t f(x_t, y_t) &\geq \tfrac{1}{T}\textstyle\sum_t f(x_t, y) - \epsilon \\ &\geq f(\hat{x}, y) - \epsilon \text{ by convexity.}\end{aligned}$$

We conclude that for all x, y we have

$$\boxed{f(\hat{x}, y) - 2\epsilon \leq f(x, \hat{y}).}$$

Minimax Theorem

Proof cont.

Choose any y, then from the no-regret property for y we get that

$$\begin{aligned} \tfrac{1}{T}\textstyle\sum_t f(x_t, y_t) &\geq \tfrac{1}{T}\textstyle\sum_t f(x_t, y) - \epsilon \\ &\geq f(\hat{x}, y) - \epsilon \text{ by convexity.} \end{aligned}$$

We conclude that for all x, y we have

$$\boxed{\max_y f(\hat{x}, y) - 2\epsilon \leq \min_x f(x, \hat{y}).}$$

Finally we get $\max_y \min_x f(x, y) \geq \min_x f(x, \hat{y})$

Minimax Theorem

Proof cont.

Choose any y, then from the no-regret property for y we get that

$$\frac{1}{T}\sum_t f(x_t, y_t) \geq \frac{1}{T}\sum_t f(x_t, y) - \epsilon$$

$$\geq f(\hat{x}, y) - \epsilon \text{ by convexity.}$$

We conclude that for all x, y we have

$$\boxed{\max_y f(\hat{x}, y) - 2\epsilon \leq \min_x f(x, \hat{y}).}$$

Finally we get $\max_y \min_x f(x, y) \geq \min_x f(x, \hat{y})$

$$\geq \max_y f(\hat{x}, y) - 2\epsilon$$

Minimax Theorem

Proof cont.

Choose any y, then from the no-regret property for y we get that

$$\begin{aligned}\frac{1}{T}\sum_t f(x_t, y_t) &\geq \frac{1}{T}\sum_t f(x_t, y) - \epsilon \\ &\geq f(\hat{x}, y) - \epsilon \text{ by convexity.}\end{aligned}$$

We conclude that for all x, y we have

$$\boxed{\max_y f(\hat{x}, y) - 2\epsilon \leq \min_x f(x, \hat{y}).}$$

Finally we get $\max_y \min_x f(x, y) \geq \min_x f(x, \hat{y})$

$$\begin{aligned}&\geq \max_y f(\hat{x}, y) - 2\epsilon \\ &\geq \min_x \max_y f(x, y) - 2\epsilon\end{aligned}$$

Minimax Theorem

Proof cont.

Choose an t

Set $\epsilon \to 0$ and we are done!

$$\frac{1}{T}\sum_t f(x_t, y_t) \geq \frac{1}{T}\sum_t f(x_t, y) - \epsilon$$

$$\geq f(\hat{x}, y) - \epsilon \text{ by convexity.}$$

We conclude that for all x, y we have

$$\max_y f(\hat{x}, y) - 2\epsilon \leq \min_x f(x, \hat{y}).$$

Finally we get $\max_y \min_x f(x, y) \geq \min_x f(x, \hat{y})$

$$\geq \max_y f(\hat{x}, y) - 2\epsilon$$

$$\geq \min_x \max_y f(x, y) - 2\epsilon$$

Last iterate convergence?

Convex-concave settings (with compact domains) are easy.
Nevertheless in GANs

- Functions are not necessarily convex-concave.
- Time averaging does not help (Jensen's ineq not applicable).
- Motivation to care about last iterate convergence!

For the rest of the lecture let's focus on

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} x^T A y.$$

Can we guarantee last iterate convergence using GD or MWUA?

Last iterate convergence?

Convex-concave settings (with compact domains) are easy.
Nevertheless in GANs

- Functions are not necessarily convex-concave.
- Time averaging does not help (Jensen's ineq not applicable).
- Motivation to care about last iterate convergence!

For the rest of the lecture let's focus on

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} x^T A y.$$

Can we guarantee last iterate convergence using GD or MWUA?

Not really...

Last iterate convergence

Consider Gradient Descent/Ascent that is

$$\begin{aligned} x_{t+1} &= x_t - \eta \nabla_x f(x_t, y_t), \\ y_{t+1} &= y_t + \eta \nabla_y f(x_t, y_t). \end{aligned}$$

Consider the simplest case $f(x, y) = xy$. GDA boils down to:

$$\begin{aligned} x_{t+1} &= x_t - \eta y_t, \\ y_{t+1} &= y_t + \eta x_t. \end{aligned}$$

Last iterate convergence

Consider Gradient Descent/Ascent that is

$$x_{t+1} = x_t - \eta \nabla_x f(x_t, y_t),$$
$$y_{t+1} = y_t + \eta \nabla_y f(x_t, y_t).$$

Consider the simplest case $f(x, y) = xy$. GDA boils down to:

$$x_{t+1} = x_t - \eta y_t,$$
$$y_{t+1} = y_t + \eta x_t.$$

Claim (Divergence). *It holds that $x_t^2 + y_t^2$ is increasing in t.*

Last iterate convergence

Consider Gradient Descent/Ascent that is

$$
\begin{aligned}
x_{t+1} &= x_t - \eta \nabla_x f(x_t, y_t), \\
y_{t+1} &= y_t + \eta \nabla_y f(x_t, y_t).
\end{aligned}
$$

Consider the simplest case $f(x, y) = xy$. GDA boils down to:

$$
\begin{aligned}
x_{t+1} &= x_t - \eta y_t, \\
y_{t+1} &= y_t + \eta x_t.
\end{aligned}
$$

Claim (Divergence). *It holds that $x_t^2 + y_t^2$ is increasing in t.*

Proof.

$$
x_{t+1}^2 + y_{t+1}^2 = (\eta^2 + 1)(x_t^2 + y_t^2).
$$

Last iterate convergence

Consider MWUA that is

$$x_i^{t+1} = \frac{x_i^t e^{-\eta(Ay^t)_i}}{Z_x},$$

$$y_j^{t+1} = \frac{y_j^t e^{\eta(A^T x^t)_j}}{Z_y}.$$

Theorem (Divergence). *Assume there exists a unique fully mixed Nash* (x^*, y^*) *equilibrium (full support). It holds that the KL divergence between a player strategies the fully mixed Nash goes to infinity, i.e,*

$$\lim_t \mathrm{KL}(x^* || x^t) = \infty \text{ and } \lim_t \mathrm{KL}(y^* || y^t) = \infty.$$

Conclusion

- Introduction to min-max optimization.
  - GANs.
  - Minimax Theorem
  - Last iterate convergence?

- Next lecture we will talk more about min-max optimization and optimism.