6.S979 Topics in Deployable Machine Learning
Lecture: Minimax and Saddle Point Problems

Asu Ozdaglar
MIT

October 3, 2019

Minimax Problem

- We consider a function $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ and the minimax problem:

$$\min_{x \in \mathbb{R}^m} \max_{y \in \mathbb{R}^n} f(x, y).$$

- We are interested in computing a saddle point of the function $f(x, y)$ where a saddle point is defined as a vector pair (x^*, y^*) that satisfies

$$f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*), \qquad \text{for all } x \in \mathbb{R}^m, \ y \in \mathbb{R}^n.$$

- Throughout this lecture, we will focus on cases where

$$\min_{x \in \mathbb{R}^m} \max_{y \in \mathbb{R}^n} f(x, y) = \max_{y \in \mathbb{R}^n} \min_{x \in \mathbb{R}^m} f(x, y).$$

  - **Minimax theorem** [von Neumann 28]: $f(x, y)$ is convex-concave ($f(\cdot, y)$ is convex for all $y \in \mathbb{R}^n$ and $f(x, \cdot)$ is concave for all $x \in \mathbb{R}^m$) and minimization and maximization is over convex sets $X \subset \mathbb{R}^m$ and $Y \subset \mathbb{R}^n$ that are compact.
  - [Moreau 64] and [Rockafellar 64] extended to noncompact sets under convex-analysis type assumptions. [Bertsekas, Nedic and Ozdaglar 03].

Minimax Problems

These problems arise in a multitude of applications:

- Worst-case design (robust optimization): We view y as a parameter and wish to minimize over x a cost function, assuming the worst possible value of y.
- Duality theory for constrained optimization We consider a constrained optimization problem (referred to as the primal problem):

$$\begin{array}{lll} \text{minimize} & f(x) & \\ \text{subject to} & g_j(x) \leq 0, & j = 1, \ldots, r. \end{array}$$

  We introduce a vector $\mu = (\mu_1, \ldots, \mu_r) \in \mathbb{R}^r$ and the *Lagrangian function*

$$L(x, \mu) = f(x) + \sum_{j=1}^{r} \mu_j g_j(x).$$

  We then consider the dual problem

$$\begin{array}{ll} \text{maximize} & \min_{x \in \mathbb{R}^n} L(x, \mu) \\ \text{subject to} & \mu \geq 0. \end{array}$$

  Thus the dual problem (and the primal problem) can be viewed as *minimax problems*.

Minimax Problems

Zero-sum games: There are two players, first choosing an action out of n possible actions, and the other choosing an action out of m possible actions.

- We assume they use *mixed strategies*: first player chooses a probability distribution $x = (x_1, \ldots, x_n)$ and second chooses $y = (y_1, \ldots, y_m)$.
- If actions i and j are selected, player 1 gives amount a_{ij} to the second player.
- The expected amount to be given by the first player to the second is $\sum_{i,j} a_{ij} x_i y_j$ or $x' A y$, where $[A]_{ij} = a_{ij}$.
- Using a worst case viewpoint, the first player must minimize $\max_y x' A y$ and the second player must maximize $\min_x x' A y$.
- Minmax theorem (a central result in game theory) states that these two optimal values are equal, implying there is an amount that can be meaningfully viewed as the value of the game for its participants.

Minimax Problems

Adversarial ML Find model parameters that minimize a loss function against worst case perturbations of input data within allowable constraints.

- Consider a standard classification problem with probability distribution $\mathcal{P}$ over pairs (w, θ) with w denoting examples and θ denoting labels.
- Selecting model parameters x to minimize exp loss $\mathbb{E}_{(w,\theta)\sim\mathcal{P}}[\ell(w, \theta, x)]$.
- A simple and effective approach for robust training of a model is to consider inputs with adversarial modifications represented as ℓ_∞-perturbed versions of data points w.
- The robust learning problem then amounts to choosing x to solve the following minimax problem:

$$\min_x \mathbb{E}_{(w,\theta)\sim\mathcal{P}}\left[\max_{y\in\mathcal{S}} \ell(w + y, \theta, x)\right],$$

  where $\mathcal{S}$ denotes allowable perturbations.

GAN Training: A zero-sum game between a generator deep NN and a discriminator deep NN.

Computing Saddle Points

Dual algorithms: Particularly relevant for constrained optimization problems. Recall the **dual problem**:

$$\begin{array}{ll} \text{maximize} & q(\mu) \\ \text{subject to} & \mu \geq 0, \end{array}$$

with dual function

$$q(\mu) = \inf_{x \in \mathbb{R}^n} L(x, \mu) = \inf_{x \in \mathbb{R}^n} \{ f(x) + \mu' g(x) \}, \quad \forall\ \mu \geq 0,$$

where $g = (g_1, \ldots, g_r)$.

- The dual objective function is concave (even when primal is nonconvex), but often nondifferentiable.
- Much of large-scale optimization (algorithms and theory) revolves around using "gradients" (to compare the value of a cost function a a given point with its values in neighboring points). This analysis breaks down when the cost function is nondifferentiable.
- Fortunately, for the case of convex cost functions, there is a convenient substitute: subgradients.

Subgradients

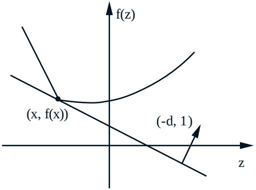
- For a convex and differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, the linearization of f at a vector x underestimates f at all points, i.e.,

$$f(z) \geq f(x) + \nabla f(x)'(z - x), \qquad \forall\, z \in \mathbb{R}^n.$$

- For a differentiable function, this linearization is unique at any given $x \in \mathbb{R}^n$.
- A convex and nondifferentiable f may have multiple linearizations at some points.
- For such functions, a subgradient provides a linearization of f that underestimates f globally at all points.

Subgradients

- For a convex function $f : \mathbb{R}^n \to \mathbb{R}$, a vector d is said to be a subgradient of f at x if
$$f(z) \geq f(x) + d'(z - x), \qquad \forall\ z \in \mathbb{R}^n.$$
- The set of subgradients of f at x is called the subdifferential of f at x and is denoted by $\partial f(x)$.
- When f is differentiable at x, we have $\partial f(x) = \{\nabla f(x)\}$.



- For a concave function $h : \mathbb{R}^n \to \mathbb{R}$, a vector d is said to be a subgradient of h at x if
$$h(z) \leq h(x) + d'(z - x), \qquad \forall\ z \in \mathbb{R}^n.$$

Characterization of the Subdifferential

Danskin's Theorem: Consider the function $f(x) = \max_{y \in Y} \phi(x, y)$, where $\phi : \mathbb{R}^{n+m} \to \mathbb{R}$ is continuous, Y is compact, and $\phi(\cdot, y)$ is convex for each $y \in Y$. Then f is convex and

$$\partial f(x) = \text{Convex Hull}\{\partial_x \phi(x, y) \mid y : \text{attains the max}\}.$$

- If there exists a unique $\bar{y}$ that attains the maximum in $\max_{y \in Y} \phi(x, y)$ and $\phi(\cdot, \bar{y})$ is differentiable at x, then f is differentiable at x, and

$$\nabla f(x) = \nabla_x \phi(x, \bar{y}).$$

- *Intuition:* Since the gradients are local objects, and the function $f(x)$ is locally the same as $\phi(x, \bar{y})$, their gradients will be the same [Madry et al 19].

Computing Subgradients of the Dual Function

- The dual function $q(\mu) = \inf_{x \in X}\{f(x) + \mu' g(x)\}$ is concave.
- Let x_μ be a vector such that

$$f(x_\mu) + \mu' g(x_\mu) = \inf_{x \in X}\{f(x) + \mu' g(x)\} = q(\mu).$$

- Then, the vector $g(x_\mu)$ is a subgradient of q at μ.
- To see this note that for all $\xi \in \mathbb{R}^r$

$$\begin{aligned} q(\xi) &= \inf_{x \in X}\{f(x) + \xi' g(x)\} \\ &\leq f(x_\mu) + \xi' g(x_\mu) \\ &= f(x_\mu) + \mu' g(x_\mu) + (\xi - \mu)' g(x_\mu) \\ &= q(\mu) + (\xi - \mu)' g(x_\mu). \end{aligned}$$

Good News: A subgradient is obtained practically for free as a by-product of the evaluation of the dual function.

Subgradient Method

- Consider maximization of $q(\mu)$ over $\mu \geq 0$.
- Subgradient method:

$$\mu_{k+1} = [\mu_k + \alpha_k g_k]^+,$$

  where g_k is the subgradient $g(x_{\mu_k})$, $[\cdot]^+$ denotes projection on the nonnegative orthant, and α_k is a positive scalar stepsize.
- [Polyak 1969], [Ermoliev 1969], [Shor 1985].
- Unlike gradients, a subgradient may not be a direction of ascent.

Subgradient Method - Convergence Properties

- Along the subgradient direction g_k, there is a range of stepsizes $(0, \tilde{\alpha})$ such that at every point $\mu_k + \alpha g_k$ for $\alpha \in (0, \tilde{\alpha})$, the distance to the optimal solution set M^* is decreased, i.e.,

$$\mathrm{dist}(\mu_k + \alpha g_k, M^*) < \mathrm{dist}(\mu_k, M^*).$$

- Remarks:
  - With the constant step, the convergence to q^* is within an error that depends on the stepsize and the bound on subgradient norms (at rate $O(1/k)$).
  - Convergence of the sequence $\{\mu_k\}$ to some dual optimal solution μ^* can be established under diminishing stepsize rule.

Computing Saddle Points

Primal-Dual algorithms:

- Let's go back to the general problem:

$$\min_{x \in \mathbb{R}^m} \max_{y \in \mathbb{R}^n} f(x, y).$$

- Assume the function $f(x, y)$ is continuously differentiable in x and y.
- An alternative method for computing the saddle points of $f(x, y)$ is the gradient descent-ascent (GDA) method: For

$$x_{k+1} = x_k - \eta \nabla_x f(x_k, y_k)$$
$$y_{k+1} = y_k + \eta \nabla_y f(x_k, y_k),$$

where $\eta > 0$ is a constant stepsize.

Computing Saddle Points

Primal-Dual algorithms - Some History

- [Samuelson 49] "The gradient method may be considered as a decentralized or computational mechanism for achieving optimum allocation of scarce resources."
- [Arrow, Hurwicz, Uzawa 58] proposed continuous-time versions of these methods for general convex-concave functions and proved global stability results under strict convexity assumptions.
- [Uzawa 58] focused on a discrete-time version and showed convergence to a neighborhood under strong convexity assumptions.
- [Gol'shtein 74] and [Maistroskii 77] provided convergence with diminishing stepsize rules under stability assumptions (weaker than strong convexity).
- [Korpelevich 77] introduced extragradient method which is a gradient method with extrapolation (see also [Nemirovski 04] for convergence rate for the convex-concave case).
- [Nedic and Ozdaglar 09] considered subgradient primal-dual methods and provided convergence rate guarantees.

Convergence Properties of GDA

- Assume $f(x, y)$ is μ_x strongly convex with respect to x and μ_y strongly concave with respect to y. Let $\mu = \min\{\mu_x, \mu_y\}$.
- Let L be the Lipschitz continuity parameter of the operator $F = [\nabla_x f(x, y); -\nabla_y f(x, y)]$.
- Define $r_k = \|x_k - x^*\|^2 + \|y_k - y^*\|^2$.

Proposition

Let $\{x_k, y_k\}$ be the iterates generated by GDA. Then for stepsize $\eta \leq \frac{\mu}{2L^2}$ the following inequality is satisfied:

$$r_{k+1} \leq \left(1 - \frac{1}{4\kappa^2}\right) r_k \tag{1}$$

Proof

- We have:

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2\eta \nabla_x f(x_k, y_k)'(x_k - x^*) + \eta^2 \|\nabla_x f(x_k, y_k)\|^2$$
$$\|y_{k+1} - x^*\|^2 = \|y_k - x^*\|^2 + 2\eta \nabla_y f(x_k, y_k)'(y_k - y^*) + \eta^2 \|\nabla_y f(x_k, y_k)\|^2$$

- Using strong convexity and concavity, we have:

$$-\nabla_x f(x_k, y_k)'(x_k - x^*) \leq f(x^*, y_k) - f(x_k, y_k) - \frac{\mu}{2}\|x_k - x^*\|^2$$
$$\nabla_y f(x_k, y_k)'(y_k - y^*) \leq f(x_k, y_k) - f(x_k, y^*) - \frac{\mu}{2}\|y_k - y^*\|^2$$

- Substituting these inequalities and adding them gives (using $z = [x; y]$)

$$\begin{aligned} \|z_{k+1} - z^*\|^2 \leq & (1 - \eta\mu)\|z_k - z^*\|^2 + 2\eta(f(x^*, y_k) - f(x_k, y^*)) \\ & + \eta^2(\|\nabla_x f(x_k, y_k)\|^2 + \|\nabla_y f(x_k, y_k)\|^2). \end{aligned} \tag{2}$$

- Using Lipschitz continuity, we obtain

$$\eta^2(\|\nabla_x f(x_k, y_k)\|^2 + \|\nabla_y f(x_k, y_k)\|^2) \leq \eta^2 L^2(\|z_k - z^*\|^2)$$

Proof (Continued)

- Using the saddle point property, we have $f(x^*, y_k) - f(x_k, y^*) \leq 0$.
- Substituting these inequalities in Equation (2), this yields

$$\|z_{k+1} - z^*\|^2 \leq (1 - \eta\mu + \eta^2 L^2)\|z_k - z^*\|^2$$

- For $\eta = \frac{\mu}{2L^2}$, we get:

$$\|z_{k+1} - z^*\|^2 \leq (1 - \frac{\mu^2}{4L^2})\|z_k - z^*\|^2$$

which can be written as:

$$\|z_{k+1} - z^*\|^2 \leq (1 - \frac{1}{4\kappa^2})\|z_k - z^*\|^2$$

Issues with GDA

- Consider the following bilinear problem:

$$\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^d} x'y$$

  The solution is $(x^*, y^*) = (0, 0)$.

- The Gradient Descent Ascent (GDA) updates for this problem:

$$\begin{aligned} x_{k+1} &= x_k \ - \ \eta y_k \\ y_{k+1} &= y_k \ + \ \eta x_k \end{aligned}$$

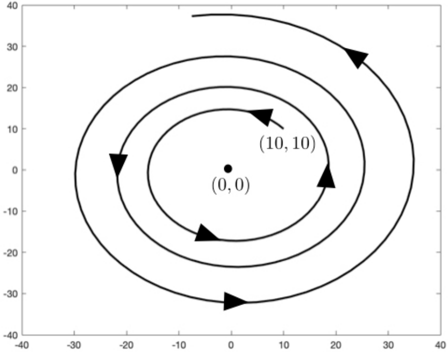
  where η is the stepsize.

GDA

- On running GDA, after k iterations we have:

$$\|x_{k+1}\|^2 + \|y_{k+1}\|^2 = (1+\eta^2)(\|x_k\|^2 + \|y_k\|^2)$$

- GDA diverges as $(1+\eta^2) > 1$



Proximal Point

- The Proximal Point (PP) updates for the same problem:

$$x_{k+1} = x_k \ - \ \eta y_{k+1}$$
$$y_{k+1} = y_k \ + \ \eta x_{k+1}$$

  where η is the stepsize.

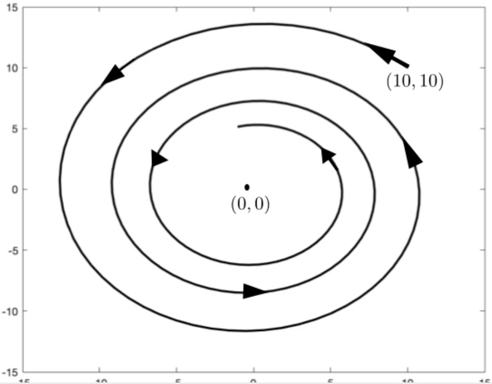
- The difference from GDA is that the gradient at the iterate (x_{k+1}, y_{k+1}) is used for the update instead of the gradient at (x_k, y_k).
- Although for this problem it takes a simple form, the PP method in general involves operator inversion and is not easy to implement.

Proximal Point

- On running PP, after k iterations we have:

$$\|x_{k+1}\|^2 + \|y_{k+1}\|^2 = \frac{1}{1+\eta^2}(\|x_k\|^2 + \|y_k\|^2)$$

- PP converges as $1/(1+\eta^2) < 1$



Proximal Point

- The PP method at each step solves the following:

$$(x_{k+1}, y_{k+1}) = \arg \min_{x \in \mathbb{R}^m} \max_{y \in \mathbb{R}^n} \left\{ f(x,y) + \frac{1}{2\eta}\|x - x_k\|^2 - \frac{1}{2\eta}\|y - y_k\|^2 \right\}.$$

- Using the first order optimality conditions leads to the following update:

$$x_{k+1} = x_k - \eta \nabla_x f(x_{k+1}, y_{k+1}), \qquad y_{k+1} = y_k + \eta \nabla_y f(x_{k+1}, y_{k+1}).$$

Theorem (Convergence of the PP method)

For any $\eta > 0$: Bilinear Case ($f(x,y) = x'By$, B: square and full-rank matrix)

$$\|x_{k+1}\|^2 + \|y_{k+1}\|^2 \leq \left(\frac{1}{1 + \eta^2 \lambda_{\min}(B^\top B)} \right) (\|x_k\|^2 + \|y_k\|^2),$$

Strongly convex-Strongly concave Case

$$\|x_{k+1} - x^*\|^2 + \|y_{k+1} - x^*\|^2 \leq \left(\frac{1}{1 + \eta\mu} \right)^k (\|x_0 - x^*\|^2 + \|y_0 - y^*\|^2),$$

OGDA updates - How prediction takes place

- One way of approximating the Proximal gradient is as follows

$$\nabla_x f(x_{k+1}, y_{k+1})) \approx \nabla_x f(x_k, y_k) + (\nabla_x f(x_k, y_k) - \nabla_x f(x_{k-1}, y_{k-1}))$$
$$\nabla_y f(x_{k+1}, y_{k+1})) \approx \nabla_y f(x_k, y_k) + (\nabla_y f(x_k, y_k) - \nabla_y f(x_{k-1}, y_{k-1}))$$

- This leads to the OGDA update

$$x_{k+1} = x_k - 2\eta \nabla_x f(x_k, y_k) + \eta \nabla_x f(x_{k-1}, y_{k-1})$$
$$y_{k+1} = y_k + 2\eta \nabla_y f(x_k, y_k) - \eta \nabla_y f(x_{k-1}, y_{k-1})$$

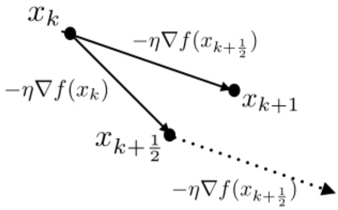
EG updates - How prediction takes place

- The updates of EG

$$x_{k+1/2} = x_k - \eta \nabla_x f(x_k, y_k), \qquad y_{k+1/2} = y_k + \eta \nabla_y f(x_k, y_k).$$

The gradients evaluated at the midpoints $x_{k+1/2}$ and $y_{k+1/2}$ are used to compute the new iterates x_{k+1} and y_{k+1} by performing the updates

$$x_{k+1} = x_k - \eta \nabla_x f(x_{k+1/2}, y_{k+1/2}),$$
$$y_{k+1} = y_k + \eta \nabla_y f(x_{k+1/2}, y_{k+1/2}).$$



EG updates - How prediction takes place

- The update can also be written as:

$$
\begin{aligned}
x_{k+1/2} = x_{k-1/2} - \eta \nabla_x f(x_{k-1/2}, y_{k-1/2}) \\
- \eta \left(\nabla_x f(x_k, y_k) - \nabla_x f(x_{k-1}, y_{k-1}) \right), \\
y_{k+1/2} = y_{k-1/2} + \eta \nabla_y f(x_{k-1/2}, y_{k-1/2}) \\
+ \eta \left(\nabla_y f(x_k, y_k) - \nabla_y f(x_{k-1}, y_{k-1}) \right).
\end{aligned}
$$

- EG tries to predict the gradient using interpolation of the midpoint gradients:

$$
\begin{aligned}
\nabla_x f(x_{k+1/2}, y_{k+1/2})) \approx \nabla_x f(x_{k-1/2}, y_{k-1/2}) \\
+ \left(\nabla_x f(x_k, y_k) - \nabla_x f(x_{k-1}, y_{k-1}) \right) \\
\nabla_y f(x_{k+1/2}, y_{k+1/2})) \approx \nabla_y f(x_{k-1/2}, y_{k-1/2}) \\
+ \left(\nabla_y f(x_k, y_k) - \nabla_y f(x_{k-1}, y_{k-1}) \right)
\end{aligned}
$$

Convergence rates of OGDA and EG

Theorem (Choose the stepsize η appropriately for each algorithm)

Bilinear case ($f(x,y) = x'By$, B: square and full-rank matrix)

$$\|x_{k+1}\|^2 + \|y_{k+1}\|^2 \leq \left(1 - \frac{1}{c\kappa}\right)^k r_0$$

Strongly Convex-Strongly Concave case

$$\|x_{k+1} - x^*\|^2 + \|y_{k+1} - x^*\|^2 \leq \left(1 - \frac{1}{c\kappa}\right)^k r_0,$$

Convex-Concave case

$$|f(\hat{x}_k, \hat{y}_k) - f(x^*, y^*)| \leq \frac{c(\|x_0 - x^*\|^2 + \|y_0 - x^*\|^2)}{k}$$

- A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach [A. Mokhtari, A. Ozdaglar, S.Pattathil 19], arXiv preprint arXiv:1901.08511.
- Convergence rate of $\mathcal{O}(1/k)$ for optimistic gradient and extra-gradient methods in smooth convex-concave saddle point problems [A. Mokhtari, A. Ozdaglar, S.Pattathil 19], arXiv preprint arXiv:1906.01115.

Extensions

- Nonconvex-nonconcave minimax problems - open problem.
- Some progress on special cases:
  - When objective function of one of the players is strongly convex, multi-step gradient descent-ascent converges to an approximate stationary point [Sanjabi, Razaviyayn, Lee 18].
  - There exist some papers which assume nonconvex on both sides, but assume additional conditions that weaken convexity assumptions, and show that inexact proximal methods converge to an approximate stationary point [Lin, Liu, Rafique, and Yang 18].