CMPT 409/981: Optimization for Machine Learning

Lecture 21

Sharan Vaswani

November 21, 2024

Min-Max Optimization

Today's focus will be on problems of the form

$$\min_{w \in \mathcal{W}} \max_{v \in \mathcal{V}} f(w, v) .$$

Example: Two player zero-sum matrix games of the form,

$$\min_{w \in \Delta_A} \max_{v \in \Delta_B} w^\intercal M v ,$$

where A is the set of strategies available to player 1. $\Delta_A = \{w \in [0,1]^{|A|} | \sum_i w_i = 1\}$ is the distribution over these available strategies and $w \in \Delta_A$ is a possible **mixed strategy**.

The matrix $M \in \mathbb{R}^{|A| \times |B|}$ is the **payoff matrix** for player 1 i.e. if player 1 plays strategy i and player 2 plays strategy j, then player 1 is rewarded $-M_{i,j}$ whereas player 2 is rewarded $M_{i,j}$. Both players are trying to maximize their respective payoffs.

Since (reward for player 1) = -(reward for player 2), this is a zero-sum game.
Classic example: Rock-Paper-Scissors

Min-Max Optimization

Example: Generative Adversarial Networks

$$\min_{\theta}\max_{\phi}\left[\mathbb{E}_{x\sim p_{\mathbf{real}}}[\log D_{\phi}(x)] + \mathbb{E}_{z\sim N(0,I_d)}[\log(1-D_{\phi}(G_{\theta}(z)))]\right],$$

where $G_{\theta}(z)$ is the generator parameterized by θ that attempts to generate realistic images from random noise z. $D_{\phi}(x)$ is the discriminator parameterized by ϕ that attempts to discriminate between the real (from p_{real}) and generated (from $G_{\theta}(z)$) images.

Example: Distributionally Robust Optimization

$$\min_{\theta}\max_{P\in\mathcal{P}}\mathbb{E}_{\zeta\sim P}[\ell(\theta,\zeta)],$$

where $\mathcal{P} := \{P|d(P,\hat{P}) \le \rho\}$ is the family of distributions that are "close" (measured by ρ) to the empirical distribution $\hat{P}$ according to a distance metric d (Total variation, KL divergence). We require that the model (parameterized by θ) is robust to distributions close to the empirical distribution from which can obtain samples.

Min-Max Optimization

Let us abstract out these problems and consider the following objective,

$$\min_{w \in \mathcal{W}} \max_{v \in \mathcal{V}} f(w, v)$$

where $\mathcal{W} \subseteq \mathbb{R}^{d_w}$ and $\mathcal{V} \subseteq \mathbb{R}^{d_v}$ are convex sets.

Claim: In general, $\max_{v \in \mathcal{V}} \min_{w \in \mathcal{W}} f(w, v) \leq \min_{w \in \mathcal{W}} \max_{v \in \mathcal{V}} f(w, v)$

Proof: Define $v^* := \arg\max_{v \in \mathcal{V}} \min_{w \in \mathcal{W}} f(w, v)$ and $w^* := \arg\min_{w \in \mathcal{W}} \max_{v \in \mathcal{V}} f(w, v)$.

$$\max_{v \in \mathcal{V}} \min_{w \in \mathcal{W}} f(w, v) = \min_{w \in \mathcal{W}} f(w, v^*) \leq f(w^*, v^*) \leq \max_{v \in \mathcal{V}} f(w^*, v) = \min_{w \in \mathcal{W}} \max_{v \in \mathcal{V}} f(w, v)$$

This referred to as the *max-min inequality* and is true for any f.

Game theoretic interpretation: RHS corresponds to w-player playing first and the v-player reacting, while the LHS corresponds to the v-player playing first and the w-player reacting. Since the v-player aims to maximize f, playing second might be beneficial since they can adapt to the w-player's strategy. Hence, the RHS $\geq$ LHS.

Min-Max Optimization

Convex-Concave Games: $f : \mathcal{W} \times \mathcal{V} \to \mathbb{R}$ is convex-concave iff $f(\cdot, v)$ is a convex function for any $v \in \mathcal{V}$, $f(w, \cdot)$ is a concave function for any $w \in \mathcal{W}$ and $\mathcal{W}, \mathcal{V}$ are convex sets.

Von Neumann-Fan Minimax Theorem: If $\mathcal{W}$ and $\mathcal{V}$ are compact, convex sets, and f is a convex-concave function, then $\max_{v \in \mathcal{V}} \min_{w \in \mathcal{W}} f(w, v) = \min_{w \in \mathcal{W}} \max_{v \in \mathcal{V}} f(w, v)$.

Example: $f(w, v) = \min_{w \in \Delta_A} \max_{v \in \Delta_B} w^\intercal My$ is convex-concave and the simplex Δ is a convex set. Hence it is a convex-concave game.

- Recall that $v^* := \arg\max_{v \in \mathcal{V}} \min_{w \in \mathcal{W}} f(w, v)$ and $w^* := \arg\min_{w \in \mathcal{W}} \max_{v \in \mathcal{V}} f(w, v)$. If f convex-concave and $\mathcal{W}$ and $\mathcal{V}$ are convex sets, then,

$$\max_{v \in \mathcal{V}} \min_{w \in \mathcal{W}} f(w, v) = \min_{w \in \mathcal{W}} f(w, v^*) = f(w^*, v^*) = \max_{v \in \mathcal{V}} f(w^*, v) = \min_{w \in \mathcal{W}} \max_{v \in \mathcal{V}} f(w, v)\,.$$

Hence, (w^*, v^*) is a solution to the game iff for all $w \in \mathcal{W}$, $v \in \mathcal{V}$,

$$f(w^*, v) \leq f(w^*, v^*) \leq f(w, v^*)\,.$$

Example: For rock-paper-scissors, the optimal mixed strategy for each player is to play either rock/paper/scissors with uniform probability.

Min-Max Optimization

Recall that for convex-concave games, (w^*, v^*) is a solution iff for all $w \in \mathcal{W}$, $v \in \mathcal{V}$, $f(w^*, v) \leq f(w^*, v^*) \leq f(w, v^*)$.

Game theoretic interpretation: From the perspective of a game between the w-player and the v-player, since $f(w^*, v^*) = \min_{w \in \mathcal{W}} f(w, v^*)$, if the v-player is playing v^*, it is optimal for the w-player to play w^*. Similarly, if the w-player is playing w^*, it is optimal for the v-player to play v^*. Hence, (w^*, v^*) is the **Nash equilibrium** since neither player has an incentive to move away from their strategy.

• For convex-concave games, the Nash equilibrium is guaranteed to exist, but need not be unique.

Duality Gap: To characterize the sub-optimality of the point $(\hat{w}, \hat{v})$:

$$\text{Duality Gap}((\hat{w}, \hat{v})) := \max_{v \in \mathcal{V}} f(\hat{w}, v) - \min_{w \in \mathcal{W}} f(w, \hat{v}) \,.$$

If $(\hat{w}, \hat{v})$ is a Nash equilibrium, then $\max_{v \in \mathcal{V}} f(\hat{w}, v) = f(\hat{w}, \hat{v}) = \min_{w \in \mathcal{W}} f(w, \hat{v})$ and hence the duality gap is 0. Point $(\hat{w}, \hat{v})$ is an ϵ-Nash equilibrium, if the Duality Gap$((\hat{w}, \hat{v})) \leq \epsilon$.

Questions?

Gradient Descent Ascent

- Gradient Descent Ascent is the simplest algorithm to solve min-max games.

Gradient Descent Ascent: At iteration k, for a step-size η, (simultaneous) projected Gradient Descent Ascent (GDA) has the following update:

$$w_{k+1} = \Pi_{\mathcal{W}}[w_k - \eta_k \nabla_w f(w_k, v_k)] \quad ; \quad v_{k+1} = \Pi_{\mathcal{V}}[v_k + \eta_k \nabla_v f(w_k, v_k)] \,,$$

where $\Pi_{\mathcal{W}}$ and $\Pi_{\mathcal{V}}$ are Euclidean projections onto $\mathcal{W}$ and $\mathcal{V}$ respectively.

G-**Lipschitz functions**: Define $z = \begin{bmatrix} w \\ v \end{bmatrix}$. The function $f : \mathcal{W} \times \mathcal{V} \to \mathbb{R}$ is G-Lipschitz iff,

$$|f(z_1) - f(z_2)| \leq G \, \|z_1 - z_2\|$$

Similar to convex minimization, this implies bounded gradients, i.e. for all $w \in \mathcal{W}$, $v \in \mathcal{V}$,

$$\|\nabla_w f(w, v)\| \leq G \quad ; \quad \|\nabla_v f(w, v)\| \leq G$$

We will also assume that sets $\mathcal{W}$ and $\mathcal{V}$ have diameter D i.e. for all $w_1, w_2 \in \mathcal{W}$, $\|w_1 - w_2\|^2 \leq D^2$. Similarly, for all $v_1, v_2 \in \mathcal{V}$, $\|v_1 - v_2\|^2 \leq D^2$.

Gradient Descent Ascent for Lipschitz, convex-concave games

Claim: For G-Lipschitz convex-concave games where $\mathcal{W}$ and $\mathcal{V}$ have diameter D, projected GDA with $\eta_k = \frac{D}{\sqrt{2}G\sqrt{k}}$ results in the following bound for $\bar{w}_T := \sum_{k=1}^{T} w_k / T$ and $\bar{v}_T := \sum_{k=1}^{T} v_k / T$

$$\text{Duality Gap}((\bar{w}_T, \bar{v}_T)) \leq \frac{4DG}{\sqrt{T}}$$

Proof: For some fixed $\tilde{w} \in \mathcal{W}$, using the projected gradient descent update for w,

$$\begin{aligned}
\|w_{k+1} - \tilde{w}\|^2 &= \|\Pi_{\mathcal{W}}[w_k - \eta \nabla_w f(w_k, v_k)] - \Pi_{\mathcal{W}}[\tilde{w}]\|^2 && \text{(Since } \tilde{w} \in \mathcal{W}) \\
&\leq \|w_k - \eta \nabla_w f(w_k, v_k) - \tilde{w}\|^2 \\
& && \text{(since projections are non-expansive)} \\
&= \|w_k - \tilde{w}\|^2 - 2\eta_k \langle \nabla_w f(w_k, v_k), w_k - \tilde{w} \rangle + \eta_k^2 \|\nabla_w f(w_k, v_k)\|^2 \\
&\leq \|w_k - \tilde{w}\|^2 - 2\eta_k [f(w_k, v_k) - f(\tilde{w}, v_k)] + \eta_k^2 G^2 \\
& && \text{(Since } f(\cdot, v_k) \text{ is convex and } f \text{ is } G\text{-Lipschitz)}
\end{aligned}$$

$$\implies [f(w_k, v_k) - f(\tilde{w}, v_k)] \leq \frac{\|w_k - \tilde{w}\|^2 - \|w_{k+1} - \tilde{w}\|^2}{2\eta_k} + \frac{\eta_k}{2} G^2 \tag{1}$$

Gradient Descent Ascent for Lipschitz, convex-concave games

Similarly, using the projected gradient ascent update w.r.t $\tilde{v} \in \mathcal{V}$,

$$\begin{aligned}\|v_{k+1} - \tilde{v}\|^2 &\leq \|v_k - \tilde{v}\|^2 + 2\eta_k \langle \nabla_v f(w_k, v_k), v_k - \tilde{v} \rangle + \eta_k^2 \|\nabla_v f(w_k, v_k)\|^2 \\ &\leq \|v_k - \tilde{v}\|^2 + 2\eta_k [f(w_k, v_k) - f(w_k, \tilde{v})] + \eta_k^2 G^2 \\ &\qquad\qquad\qquad\qquad \text{(Since } f(w_k, \cdot) \text{ is concave and } f \text{ is } G\text{-Lipschitz)}\end{aligned}$$

$$\implies [f(w_k, \tilde{v}) - f(w_k, v_k)] \leq \frac{\|v_k - \tilde{v}\|^2 - \|v_{k+1} - \tilde{v}\|^2}{2\eta_k} + \frac{\eta_k}{2} G^2 \tag{2}$$

Adding eq. (1) and eq. (2),

$$f(w_k, \tilde{v}) - f(\tilde{w}, v_k) \leq \frac{\|w_k - w\|^2 - \|w_{k+1} - w\|^2}{2\eta_k} + \frac{\|v_k - v\|^2 - \|v_{k+1} - v\|^2}{2\eta_k} + \eta_k G^2$$

$$\begin{aligned}\sum_{k=1}^{T} [f(w_k, \tilde{v}) - f(\tilde{w}, v_k)] &\leq \sum_{k=1}^{T} \left[\frac{\|w_k - \tilde{w}\|^2 - \|w_{k+1} - \tilde{w}\|^2}{2\eta_k} \right] + \sum_{k=1}^{T} \left[\frac{\|v_k - \tilde{v}\|^2 - \|v_{k+1} - \tilde{v}\|^2}{2\eta_k} \right] \\ &\quad + G^2 \sum_{k=1}^{T} \eta_k\end{aligned}$$

Gradient Descent Ascent for Lipschitz, convex-concave games

Simplifying the first term in the equation from the previous slide,

$$\sum_{k=1}^{T}\left[\frac{\|w_k-\tilde{w}\|^2-\|w_{k+1}-\tilde{w}\|^2}{2\eta_k}\right] \le \sum_{k=2}^{T}\|w_k-\tilde{w}\|^2\left[\frac{1}{\eta_k}-\frac{1}{\eta_{k-1}}\right]+\frac{\|w_1-w^*\|^2}{2\eta_1}$$
$$\le \frac{D^2}{2\eta_T}$$

Bounding the second term in a similar manner and putting everything together,

$$\sum_{k=1}^{T}[f(w_k,\tilde{v})-f(\tilde{w},v_k)] \le \frac{D^2}{\eta_T}+G^2\sum_{k=1}^{T}\eta_k = \frac{D^2\sqrt{T}}{\eta}+G^2\eta\sum_{k=1}^{T}\frac{1}{\sqrt{k}} \qquad (\eta_k = \eta/\sqrt{k})$$

$$\le \frac{D^2\sqrt{T}}{\eta}+2G^2\eta\sqrt{T} \qquad \left(\textstyle\sum_{k=1}^{T} 1/\sqrt{k} \le 2\sqrt{T}\right)$$

$$\implies \frac{1}{T}\left[\sum_{k=1}^{T}[f(w_k,\tilde{v})-f(\tilde{w},v_k)]\right] \le \frac{D^2\sqrt{T}}{\eta}+2G^2\eta\sqrt{T} = \frac{4\,DG}{\sqrt{T}} \qquad \left(\eta = \tfrac{D}{\sqrt{2}G}\right)$$

Gradient Descent Ascent for Lipschitz, convex-concave games

Recall that $\frac{1}{T}\left[\sum_{k=1}^{T}[f(w_k,\tilde{v}) - f(\tilde{w}, v_k)]\right] \leq \frac{4\,DG}{\sqrt{T}}$. Since $f(\cdot,\tilde{v})$ and $-f(\tilde{w},\cdot)$ are convex, using Jensen's inequality and by definition of $\bar{w}_T$ and $\bar{v}_T$,

$$f(\bar{w}_T,\tilde{v}) - f(\tilde{w},\bar{v}_T) \leq \frac{4\,DG}{\sqrt{T}}$$

Since the above statement is true for all $\tilde{v} \in \mathcal{V}$ and $\tilde{w} \in \mathcal{W}$, taking the maximum over $\tilde{v} \in \mathcal{V}$ and the minimum over $\tilde{w} \in \mathcal{W}$,

$$\max_{v\in\mathcal{V}} f(\bar{w}_T, v) - \min_{w\in\mathcal{W}} f(w,\bar{v}_T) \leq \frac{4\,DG}{\sqrt{T}} \implies \text{Duality Gap}((\bar{w}_T,\bar{v}_T)) \leq \frac{4DG}{\sqrt{T}}$$

• Recall that GD attains an $O(1/\sqrt{T})$ rate when minimizing convex, Lipschitz functions, and hence GDA has a similar behaviour when solving convex-concave Lipschitz games.

Gradient Descent Ascent for smooth, convex-concave games

- **Smoothness**: $f : \mathcal{W} \times \mathcal{V} \to \mathbb{R}$ is L-smooth iff

$$\|\nabla_w f(w_1, v_1) - \nabla_w f(w_2, v_2)\| \leq L \|z_1 - z_2\| \quad ; \quad \|\nabla_v f(w_1, v_1) - \nabla_v f(w_2, v_2)\| \leq L \|z_1 - z_2\| \,,$$

where $z_1 = \begin{bmatrix} w_1 \\ v_1 \end{bmatrix}$ and $z_2 = \begin{bmatrix} w_2 \\ v_2 \end{bmatrix}$.

Example: The *bilinear game* $f(w, v) = w\, v$ is $\sqrt{2}$-smooth since $\nabla_w f(w, v) = v$ and $|v_1 - v_2| \leq |v_1 - v_2| + |w_1 - w_2| \leq \sqrt{2} \|z_1 - z_2\|$. A similar reasoning works for $\nabla_v f(w, v)$. Since $f(\cdot, v)$ is linear w.r.t w, it is convex. By symmetry, $f(w, \cdot)$ is linear in v and hence concave.

If $\mathcal{W} = \mathbb{R}$ and $\mathcal{V} = \mathbb{R}$, $\min_{w \in \mathbb{R}} \max_{v \in \mathbb{R}} wv$ is a smooth, convex-concave game whose unique solution is at $(0, 0)$ since $f(0, 0) \leq f(w, 0)$ for all w and $f(0, 0) \geq f(0, v)$ for all v.

Game theoretically, if the v-player deviates from 0 such that $v = \epsilon$, the w-player can choose $-\infty$ to make the objective small. Similarly, if the w-player deviates from 0 such $w = \epsilon$, then the v-player can choose $+\infty$ to make the objective large. Hence, neither play has an incentive to deviate from $(0, 0)$ which corresponds to the Nash equilibrium.

Gradient Descent Ascent for smooth, convex-concave games

Let us consider running GDA for $\min_{w \in \mathbb{R}} \max_{v \in \mathbb{R}} wv$. The update can be given as:

$$w_{k+1} = w_k - \eta_k \nabla_w f(w_k, v_k) = w_k - \eta_k v_k \quad ; \quad v_{k+1} = v_k + \eta_k \nabla_v f(w_k, v_k) = v_k + \eta_k w_k$$

Calculating the distance from the solution $(0, 0)$ after one iteration,

$$(w_{k+1} - 0)^2 + (v_{k+1} - 0)^2 = (w_k - \eta_k v_k)^2 + (v_k + \eta_k w_k)^2 = (1 + \eta_k^2)\,(w_k^2 + v_k^2)$$

• Hence, for any η_k, the last iterate of GDA will move away from the solution, diverging in the unconstrained setting or hitting the boundary in the constrained setting.

• Compare this to GD for smooth, convex minimization where the sub-optimality corresponding to the last iterate decreases at an $O(1/T)$ rate (Lecture 4). However, we can show that the average iterate will converge at an $O(1/\sqrt{T})$ rate.

Gradient Descent Ascent for smooth, convex-concave games

Claim: An L-smooth game $\min_{w \in \mathcal{W}} \max_{v \in \mathcal{V}} f(w, v)$ where $\mathcal{W}$ and $\mathcal{V}$ have diameter D is $G := 2DL + \sqrt{2} \max\{\|\nabla_w f(w_0, v_0)\|, \|\nabla_v f(w_0, v_0)\|\}$-Lipschitz.

Proof: By the definition of L-smoothness, for any (w_1, v_2) and (w_2, v_2),

$\|\nabla_w f(w_1, v_1) - \nabla_w f(w_2, v_2)\| \leq L \|z_1 - z_2\| \leq L \sqrt{\|w_1 - w_2\|^2 + \|v_1 - v_2\|^2} \leq \sqrt{2} DL.$

For any w, v, $\|\nabla_w f(w, v)\| = \|\nabla_w f(w, v) - \nabla_w f(w_0, v_0) + \nabla_w f(w_0, v_0)\| \leq \|\nabla_w f(w, v) - \nabla_w f(w_0, v_0)\| + \|\nabla_w f(w_0, v_0)\| \leq \sqrt{2} DL + \|\nabla_w f(w_0, v_0)\|$.

Similarly, $\|\nabla_v f(w, v)\| \leq \sqrt{2} DL + \|\nabla_v f(w_0, v_0)\|$, and hence $G = 2DL + \sqrt{2} \max\{\|\nabla_w f(w_0, v_0)\|, \|\nabla_v f(w_0, v_0)\|\}$.

Claim: For L-smooth, convex-concave games, GDA with $\eta_k = \frac{D}{G\sqrt{k}}$ where $G = \left(2DL + \sqrt{2} \max\{\|\nabla_w f(w_0, v_0)\|, \|\nabla_v f(w_0, v_0)\|\}\right)$ results in the following bound for $\bar{w}_T := \sum_{k=1}^{T} w_k / T$ and $\bar{v}_T := \sum_{k=1}^{T} v_k / T$

$$\text{Duality Gap}((\bar{w}_T, \bar{v}_T)) \leq \frac{4\, D\, G}{\sqrt{T}}$$

Proof: Using the result for convex-concave G-Lipschitz games.

Strongly-convex strongly-concave games

Strongly-convex strongly-concave games: $f : \mathcal{W} \times \mathcal{V} \to \mathbb{R}$ is strongly-convex strongly-concave iff $f(\cdot, v)$ is a strongly-convex function for any $v \in \mathcal{V}$, $f(w, \cdot)$ is a strongly-concave function for any $w \in \mathcal{W}$ and the sets $\mathcal{W}, \mathcal{V}$ are convex sets, i.e. for all $w, w_1, w_2 \in \mathcal{W}$ and $v, v_1, v_2 \in \mathcal{V}$,

$$f(w_2, v) \geq f(w_1, v) + \langle \nabla_w f(w_1, v), w_2 - w_1 \rangle + \frac{\mu_w}{2} \|w_1 - w_2\|^2$$

$$-f(w, v_2) \geq -f(w, v_1) + \langle -\nabla_v f(w, v_1), v_2 - v_1 \rangle + \frac{\mu_v}{2} \|v_1 - v_2\|^2$$

If $\mathcal{W} = \mathbb{R}^d$ and $\mathcal{V} = \mathbb{R}^d$ since $w^* := \arg\min_w f(w, v^*)$, $\nabla_w f(w^*, v^*) = 0$. By the strong-convexity of $f(\cdot, v)$ with $v = v^*$, $w_1 = w^*$, $w_2 = w$, $f(w^*, v^*) < f(w, v^*)$ for all w.

Similarly, $v^* := \arg\max_v f(w^*, v)$, $\nabla_v f(w^*, v^*) = 0$. By the strong-concavity of $f(w, \cdot)$ with $w = w^*$, $-f(w^*, v) > -f(w^*, v^*)$. Hence, $f(w^*, v^*) > f(w^*, v)$ for all v.

- Hence, for unconstrained strongly-convex strongly-concave games, (w^*, v^*) is the unique Nash equilibrium and $\nabla_w f(w^*, v^*) = \nabla_v f(w^*, v^*) = 0$.

Gradient Descent Ascent for smooth, strongly-convex strongly-concave games

Claim: For L-smooth, μ strongly-convex strongly-concave games, T iterations of GDA with $\eta_k = \frac{\mu}{4L^2}$ results in the following bound,

$$\left\| \begin{bmatrix} w_T - w^* \\ v_T - v^* \end{bmatrix} \right\|^2 \leq \exp\left(\frac{-T}{4\kappa^2}\right) \left\| \begin{bmatrix} w_0 - w^* \\ v_0 - v^* \end{bmatrix} \right\|^2 .$$

• Hence, for smooth, strongly-convex strongly-concave games with condition number κ, we need to run GDA for $T = O\left(\kappa^2 \log\left(\frac{1}{\epsilon}\right)\right)$ in order to get ϵ-close to the Nash equilibrium. The $O(\kappa^2)$ dependence can not be improved for GDA.

• In contrast, for minimizing smooth, strongly-convex functions GD requires $O\left(\kappa \log\left(\frac{1}{\epsilon}\right)\right)$ iterations in order to get ϵ-close to the minimizer.

Questions?

Proximal Point Method

- Recall that the last iterate of GDA diverges on bilinear games of the form $f(w, v) = wv$, and only the averaged iterate converges at an $O(1/\sqrt{T})$ rate. The **proximal point method** and its approximations obtain last-iterate convergence for this class of games.

Proximal Point Method (PPM): At iteration k, PPM has the following update:

$$w_{k+1} = w_k - \eta \nabla_w f(w_{k+1}, v_{k+1}) \; ; \; v_{k+1} = v_k + \eta \nabla_v f(w_{k+1}, v_{k+1})$$

- Has a built in "lookahead" which prevents the diverging behaviour of GDA.
- For bilinear games, attains an $O(\log(1/\epsilon))$ last-iterate convergence to the Nash equilibrium.
- Since computing w_{k+1} relies on computing $\nabla_w f(w_{k+1}, v_{k+1})$, PPM is an *implicit method* and implementing it requires a computationally expensive matrix inversion.

Optimistic GDA and Extra-Gradient Method

Two computationally efficient ways of reproducing the favourable behaviour of PPM:

Extra-Gradient Method (EG): At iteration k, EG has the following update,

$$w_{k+1/2} = w_k - \eta \nabla_w f(w_k, v_k) \; ; \; v_{k+1/2} = v_k + \eta \nabla_v f(w_k, v_k)$$
$$w_{k+1} = w_k - \eta \nabla_w f(w_{k+1/2}, v_{k+1/2}) \; ; \; v_{k+1} = v_k + \eta \nabla_v f(w_{k+1/2}, v_{k+1/2})$$

- The $(w_{k+1/2}, v_{k+1/2})$ iterates approximate the implicit update in PPM.
- Each iteration requires computing two gradients (there are recent "single-call" EG methods).

Optimistic GDA (OGDA): At iteration k, OGDA has the following update,

$$w_{k+1} = w_k - \eta \nabla_w f(w_k, v_k) - \eta \left[\nabla_w f(w_k, v_k) - \nabla_w f(w_{k-1}, v_{k-1})\right]$$
$$v_{k+1} = v_k + \eta \nabla_v f(w_k, v_k) - \eta \left[\nabla_v f(w_{k-1}, v_{k-1}) - \nabla_v f(w_k, v_k)\right]$$

- The second term acts as "negative momentum" preventing the cycling behaviour.
- Compared to EG, each iteration of OGDA requires computing only one gradient.
- For bilinear games, EG and OGDA result in $O(\log(1/\epsilon))$ convergence similar to PPM.
- EG and OGDA have been used to train GANs [DISZ17, GBV$^+$18].

Comparing GDA, PPM, EG, OGDA on a bilinear game [MOP20]

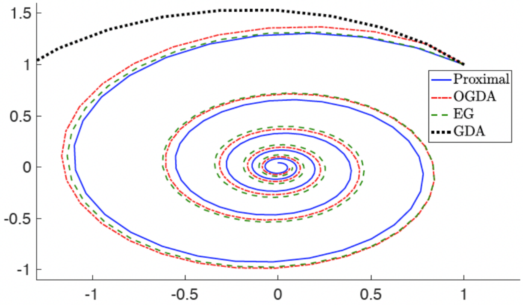

Figure 1: Convergence trajectories of proximal point (PP), extra-gradient (EG), optimistic gradient descent ascent (OGDA), and gradient descent ascent (GDA) for $\min_x \max_y xy$. The proximal point method has the fastest convergence. EG and OGDA approximate the trajectory of PP and both converge to the optimal solution. The GDA method is the only method that diverges.

Need to implement in Assignment 4!

Questions?

Extra-Gradient Method

- In order to analyze the convergence of projected EG, we write in the following equivalent way,

$$z_{k+1/2} = \Pi_{\mathcal{Z}}[\tilde{z}_{k+1/2}] \quad ; \quad \tilde{z}_{k+1/2} = z_k - \eta F(z_k)$$
$$z_{k+1} = \Pi_{\mathcal{Z}}[\tilde{z}_{k+1}] \quad ; \quad \tilde{z}_{k+1} = z_k - \eta F(z_{k+1/2})$$

where $z = \begin{bmatrix} w \\ v \end{bmatrix}$, $F(z) = \begin{bmatrix} \nabla_w f(w,v) \\ -\nabla_v f(w,v) \end{bmatrix}$ is an *operator* from $\mathbb{R}^{d_w+d_v} \to \mathbb{R}^{d_w+d_v}$ and $\Pi_{\mathcal{Z}}$ is Euclidean projection onto $\mathcal{W} \times \mathcal{V}$.

- If $z^* = \begin{bmatrix} w^* \\ v^* \end{bmatrix}$ is the solution, then using the definition of optimality, for all $w \in \mathcal{W}$ and $v \in \mathcal{V}$,

$$\langle \nabla_w f(w^*, v), w - w^* \rangle \geq 0 \,;\, \langle -\nabla_v f(w, v^*), v - v^* \rangle \geq 0$$

Setting $v = v^*$ in the first equation, and $w = w^*$ in the second equation, then for all $z \in \mathcal{Z}$,

$$\implies \left\langle \begin{bmatrix} \nabla_w f(w^*, v^*) \\ -\nabla_v f(w^*, v^*) \end{bmatrix}, \begin{bmatrix} w \\ v \end{bmatrix} - \begin{bmatrix} w^* \\ v^* \end{bmatrix} \right\rangle \geq 0 \implies \langle F(z^*), z - z^* \rangle \geq 0 \tag{3}$$

Extra-Gradient Method

Claim: If f is L-smooth, then the operator F is $2L$-Lipschitz i.e. $\|F(z_1) - F(z_2)\| \leq 2L \|z_1 - z_2\|$.

Proof:

$$\begin{aligned}
\|F(z_1) - F(z_2)\| &= \left\| \begin{bmatrix} \nabla_w f(w_1, v_1) - \nabla_w f(w_2, v_2) \\ \nabla_v f(w_2, v_2) - \nabla_v f(w_1, v_1) \end{bmatrix} \right\| \\
&\leq \|\nabla_w f(w_1, v_1) - \nabla_w f(w_2, v_2)\| + \|\nabla_v f(w_1, v_1) - \nabla_v f(w_2, v_2)\| \\
&\leq L \|z_1 - z_2\| + L \|z_1 - z_2\| \qquad \text{(By definition of } L\text{-smoothness)} \\
\|F(z_1) - F(z_2)\| &\leq 2L \|z_1 - z_2\|
\end{aligned}$$

Extra-Gradient Method

In the next class, we will prove the following claim:

Claim: For L-smooth, convex-concave games where $\mathcal{W}$ and $\mathcal{V}$ have diameter D, EG with $\eta_k = \frac{1}{2L}$ results in the following bound for $\bar{w}_T := \sum_{k=1}^{T} w_{k+1/2} / T$ and $\bar{v}_T := \sum_{k=1}^{T} v_{k+1/2} / T$,

$$\text{Duality Gap}((\bar{w}_T, \bar{v}_T)) \leq \frac{2D^2 L}{T}$$

- Hence, compared to GDA that has an $O\left(1/\sqrt{T}\right)$ convergence, the average iterate for EG has an $O\left(1/T\right)$ convergence for the duality gap.
- The last iterate for EG has a slower $\Theta\left(1/\sqrt{T}\right)$ convergence for the duality gap [GPDO20].

- Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng, *Training gans with optimism*, arXiv preprint arXiv:1711.00141 (2017).
- Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien, *A variational inequality perspective on generative adversarial networks*, arXiv preprint arXiv:1802.10551 (2018).
- Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar, *Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems*, Conference on Learning Theory, PMLR, 2020, pp. 1758–1784.
- Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil, *A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach*, International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 1497–1507.