CMPT 409/981: Optimization for Machine Learning

Lecture 20

Sharan Vaswani

November 24, 2022

Recap

Convex-concave games: $\min_{w \in \mathcal{W}} \max_{v \in \mathcal{V}} f(w, v)$, where $\mathcal{W} \subseteq \mathbb{R}^{d_w}$ and $\mathcal{V} \subseteq \mathbb{R}^{d_v}$ are convex sets and f is convex in w and concave in v. For convex-concave games, (w^*, v^*) is a solution iff for all $w \in \mathcal{W}$, $v \in \mathcal{V}$, $f(w^*, v) \leq f(w^*, v^*) \leq f(w, v^*)$.

To characterize the sub-optimality of $(\hat{w}, \hat{v})$:
Duality Gap$((\hat{w}, \hat{v})) := \max_{v \in \mathcal{V}} f(\hat{w}, v) - \min_{w \in \mathcal{W}} f(w, \hat{v})$.

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

Recap

G-**Lipschitz convex-concave games**: Projected GDA has the guarantee that Duality Gap$((\bar{w}_T, \bar{v}_T)) \leq \frac{4DG}{\sqrt{T}}$ where $\bar{w}_T$ and $\bar{v}_T$ are the average iterates.

Smooth, convex-concave games: Last iterate of GDA will move away from the solution, diverging in the unconstrained setting or hitting the boundary in the constrained setting. For sets with bounded diameter, the average iterates result in an $O\left(1/\sqrt{T}\right)$ decrease on the duality gap.

The GDA update for unconstrained games can be written as $z_{k+1} = z_k - \eta_k F(z_k)$ where $z = \begin{bmatrix} w \\ v \end{bmatrix}$ and $F(z) = \begin{bmatrix} \nabla_w f(w, v) \\ -\nabla_v f(w, v) \end{bmatrix}$. For unconstrained games, $F(z^*) = 0$ where $z^* = \begin{bmatrix} w^* \\ v^* \end{bmatrix}$.

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

Strongly-convex, strongly-concave games: $f(\cdot, v)$ is μ_w strongly-convex and $f(w, \cdot)$ is μ_v strongly-concave. The Nash equilibrium (w^*, v^*) is unique.

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

Claim: If f is strongly-convex strongly-concave with $\mu_w = \mu_v = \mu$, then F is μ strongly-monotone i.e. $\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geq \mu \left\| z_1 - z_2 \right\|^2$.

Proof: By strong-convexity of $f(\cdot, v)$,

$$f(w_2, v_1) \geq f(w_1, v_1) + \langle \nabla_w f(w_1, v_1), w_2 - w_1 \rangle + \frac{\mu}{2} \left\| w_1 - w_2 \right\|^2 \qquad \text{(With } v = v_1\text{)}$$

$$f(w_1, v_2) \geq f(w_2, v_2) + \langle \nabla_w f(w_2, v_2), w_1 - w_2 \rangle + \frac{\mu}{2} \left\| w_1 - w_2 \right\|^2 \qquad \text{(With } v = v_2\text{)}$$

By strong-concavity of $f(w, \cdot)$,

$$-f(w_1, v_2) \geq -f(w_1, v_1) + \langle -\nabla_v f(w_1, v_1), v_2 - v_1 \rangle + \frac{\mu}{2} \left\| v_1 - v_2 \right\|^2 \qquad \text{(With } w = w_1\text{)}$$

$$-f(w_2, v_1) \geq -f(w_2, v_2) + \langle -\nabla_v f(w_2, v_2), v_1 - v_2 \rangle + \frac{\mu}{2} \left\| v_1 - v_2 \right\|^2 \qquad \text{(With } w = w_2\text{)}$$

Adding all the 4 equations,

$$\langle \nabla_w f(w_1, v_1) - \nabla_w f(w_2, v_2), w_1 - w_2 \rangle + \langle \nabla_v f(w_2, v_2) - \nabla_v f(w_1, v_1), v_1 - v_2 \rangle$$
$$\geq \mu \left[\left\| w_1 - w_2 \right\|^2 + \left\| v_1 - v_2 \right\|^2 \right] = \mu \left\| z_1 - z_2 \right\|^2$$

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

Rewriting

$$\langle \nabla_w f(w_1, v_1) - \nabla_w f(w_2, v_2), w_1 - w_2 \rangle + \langle \nabla_v f(w_2, v_2) - \nabla_v f(w_1, v_1), v_1 - v_2 \rangle \geq \mu \left\| z_1 - z_2 \right\|^2,$$

$$\left\langle \begin{bmatrix} \nabla_w f(w_1, v_1) - \nabla_w f(w_2, v_2) \\ -\nabla_v f(w_1, v_1) + \nabla_v f(w_2, v_2) \end{bmatrix}, \begin{bmatrix} w_1 - w_2 \\ v_1 - v_2 \end{bmatrix} \right\rangle \geq \mu \left\| z_1 - z_2 \right\|^2$$

$$\left\langle \begin{bmatrix} \nabla_w f(w_1, v_1) \\ -\nabla_v f(w_1, v_1) \end{bmatrix} - \begin{bmatrix} \nabla_w f(w_2, v_2) \\ -\nabla_v f(w_2, v_2) \end{bmatrix}, \begin{bmatrix} w_1 \\ v_1 \end{bmatrix} - \begin{bmatrix} w_2 \\ v_2 \end{bmatrix} \right\rangle \geq \mu \left\| z_1 - z_2 \right\|^2$$

$$\implies \langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geq \mu \left\| z_1 - z_2 \right\|^2$$

Hence, if f is μ strongly-convex and strongly-concave and L-smooth, then the operator F is $2L$-Lipschitz and μ strongly-monotone.

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,

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

Proof: Since GDA can be equivalently written as $z_{k+1} = z_k - \eta F(z_k)$.

$$\begin{aligned}
\|z_{k+1} - z^*\|^2 &= \|z_k - z^* - \eta F(z_k)\|^2 = \|z_k - z^*\|^2 - 2\eta\langle F(z_k), z_k - z^*\rangle + \eta^2 \|F(z_k)\|^2 \\
&= \|z_k - z^*\|^2 - 2\eta\langle F(z_k) - F(z^*), z_k - z^*\rangle + \eta^2 \|F(z_k) - F(z^*)\|^2 \\
&\qquad (F(z^*) = 0 \text{ for unconstrained strongly-convex, strongly-concave games}) \\
&\leq \|z_k - z^*\|^2 - 2\eta\langle F(z_k) - F(z^*), z_k - z^*\rangle + 4L^2\eta^2 \|z_k - z^*\|^2 \\
&\qquad (F \text{ is } 2L\text{-Lipschitz}) \\
&\leq \|z_k - z^*\|^2 - 2\mu\eta \|z_k - z^*\|^2 + 4L^2\eta^2 \|z_k - z^*\|^2 \\
&\qquad (F \text{ is } \mu \text{ strongly-monotone}) \\
&= \|z_k - z^*\|^2 \left(1 - 2\mu\eta + 4L^2\eta^2\right)
\end{aligned}$$

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

Recall that $\|z_{k+1} - z^*\|^2 \leq \|z_k - z^*\|^2 \left(1 - 2\mu\eta + 4L^2\eta^2\right)$. We need to set η such that $\left(1 - 2\mu\eta + 4L^2\eta^2\right) < 1 \implies \eta < \frac{\mu}{2L^2}$. Setting $\eta = \frac{\mu}{4L^2}$

$$\|z_{k+1} - z^*\|^2 \leq \|z_k - z^*\|^2 \left(1 - 2\mu\frac{\mu}{4L^2} + 4L^2\frac{\mu^2 L^2}{16L^4}\right) \implies \|z_{k+1} - z^*\|^2 \leq \left(1 - \frac{\mu^2}{4L^2}\right)\|z_k - z^*\|^2$$

Recursing from $k = 0$ to $T - 1$,

$$\|z_T - z^*\|^2 \leq \left(1 - \frac{\mu^2}{4L^2}\right)^T \|z_0 - z^*\|^2 \leq \exp\left(\frac{-T}{4\kappa^2}\right)\|z_0 - z^*\|^2$$

(Since $1 - x \leq \exp(-x)$ for all x)

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 cycling behaviour like 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]

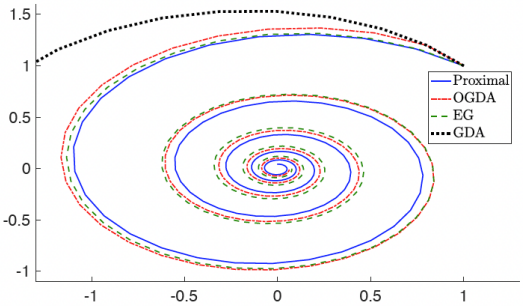

Figure 1: Convergence trajectories of proximal point (PP), extra-gradient (EG), optimistic gradient descent ascent (OGDA), and gradient descent ascent (GDA) for $\min_x \max_y xy$. The proximal point method has the fastest convergence. EG and OGDA approximate the trajectory of PP and both converge to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

Need to implement in Assignment 4!

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 $z = \begin{bmatrix} w \\ v \end{bmatrix}$, $F(z) = \begin{bmatrix} \nabla_w f(w,v) \\ -\nabla_v f(w,v) \end{bmatrix}$ and $\Pi_{\mathcal{Z}}$ is Euclidean projection onto $\mathcal{W} \times \mathcal{V}$.

Using the property of Euclidean projections onto $\mathcal{Z}$, for $z \in \mathcal{Z}$,

$$\langle \tilde{z}_{k+1/2} - z_{k+1/2}, z - z_{k+1/2} \rangle \leq 0 \implies \langle -\tilde{z}_{k+1/2}, z_{k+1/2} - z \rangle \leq \langle -z_{k+1/2}, z_{k+1/2} - z \rangle \quad (1)$$
$$\langle \tilde{z}_{k+1} - z_{k+1}, z - z_{k+1} \rangle \leq 0 \implies \langle -\tilde{z}_{k+1}, z_{k+1} - z \rangle \leq \langle -z_{k+1}, z_{k+1} - z \rangle \quad (2)$$

Extra-Gradient Method

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 $z \in \mathcal{Z}$,

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

We will consider the case when $f(w, v)$ is a smooth game, and using the same derivation as before F is a $2L$-Lipschitz operator i.e. $\|F(z_1) - F(z_2)\| \leq 2L \, \|z_1 - z_2\|$.

Extra-Gradient for smooth, convex-concave games

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 $\bar{v}_T := \sum_{k=1}^{T} v_{k+1/2}/T$,

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

Proof: For $\tilde{w} \in \mathcal{W}$, $\tilde{v} \in \mathcal{V}$,

$$\begin{aligned}
&f(w_{k+1/2}, \tilde{v}) - f(\tilde{w}, v_{k+1/2}) \\
&= f(w_{k+1/2}, \tilde{v}) - f(w_{k+1/2}, v_{k+1/2}) + f(w_{k+1/2}, v_{k+1/2}) - f(\tilde{w}, v_{k+1/2}) \\
&\leq \langle \nabla_v f(w_{k+1/2}, v_{k+1/2}), \tilde{v} - v_{k+1/2} \rangle + \langle \nabla_w f(w_{k+1/2}, v_{k+1/2}), w_{k+1/2} - \tilde{w} \rangle \\
&\qquad\qquad \text{(Convexity of } f(\cdot, v_{k+1/2}) \text{ and concavity of } f(w_{k+1/2}, \cdot)) \\
&= \left\langle \begin{bmatrix} \nabla_w f(w_{k+1/2}, v_{k+1/2}) \\ -\nabla_v f(w_{k+1/2}, v_{k+1/2}) \end{bmatrix}, \begin{bmatrix} w_{k+1/2} - \tilde{w} \\ v_{k+1/2} - \tilde{v} \end{bmatrix} \right\rangle \\
\implies & f(w_{k+1/2}, \tilde{v}) - f(\tilde{w}, v_{k+1/2}) \leq \langle F(z_{k+1/2}), z_{k+1/2} - \tilde{z} \rangle \qquad (4)
\end{aligned}$$

We will bound the $\langle F(z_{k+1/2}), z_{k+1/2} - \tilde{z} \rangle$ term in order to get a handle on $f(w_{k+1/2}, \tilde{v}) - f(\tilde{w}, v_{k+1/2})$ and hence the duality gap.

Extra-Gradient for smooth, convex-concave games

$$\langle F(z_{k+1/2}), z_{k+1/2} - \tilde{z} \rangle = \left\langle \frac{z_k - \tilde{z}_{k+1}}{\eta}, z_{k+1/2} - \tilde{z} \right\rangle \qquad \text{(Using the update)}$$

$$= \left\langle \frac{z_k - \tilde{z}_{k+1}}{\eta}, z_{k+1/2} - z_{k+1} \right\rangle + \left\langle \frac{z_k - \tilde{z}_{k+1}}{\eta}, z_{k+1} - \tilde{z} \right\rangle \qquad \text{(Add/Subtract } z_{k+1}\text{)}$$

$$\leq \left\langle \frac{z_k - \tilde{z}_{k+1}}{\eta}, z_{k+1/2} - z_{k+1} \right\rangle + \left\langle \frac{z_k - z_{k+1}}{\eta}, z_{k+1} - \tilde{z} \right\rangle$$

(Using Eq. (2) for the second term)

$$= \left\langle \frac{z_k - \tilde{z}_{k+1/2}}{\eta}, z_{k+1/2} - z_{k+1} \right\rangle + \left\langle \frac{\tilde{z}_{k+1/2} - \tilde{z}_{k+1}}{\eta}, z_{k+1/2} - z_{k+1} \right\rangle + \left\langle \frac{z_k - z_{k+1}}{\eta}, z_{k+1} - \tilde{z} \right\rangle$$

(Add/Subtract $\tilde{z}_{k+1/2}$)

$$\leq \left\langle \frac{z_k - z_{k+1/2}}{\eta}, z_{k+1/2} - z_{k+1} \right\rangle + \left\langle \frac{\tilde{z}_{k+1/2} - \tilde{z}_{k+1}}{\eta}, z_{k+1/2} - z_{k+1} \right\rangle + \left\langle \frac{z_k - z_{k+1}}{\eta}, z_{k+1} - \tilde{z} \right\rangle$$

(Using Eq. (1) for the first term)

Extra-Gradient for smooth, convex-concave games

Recall that $\langle F(z_{k+1/2}), z_{k+1/2} - \tilde{z} \rangle \leq$
$\left\langle \frac{z_k - z_{k+\mathbf{1/2}}}{\eta}, z_{k+1/2} - z_{k+1} \right\rangle + \left\langle \frac{\tilde{z}_{k+\mathbf{1/2}} - \tilde{z}_{k+\mathbf{1}}}{\eta}, z_{k+1/2} - z_{k+1} \right\rangle + \left\langle \frac{z_k - z_{k+\mathbf{1}}}{\eta}, z_{k+1} - \tilde{z} \right\rangle.$

$$\implies \eta \langle F(z_{k+1/2}), z_{k+1/2} - \tilde{z} \rangle$$
$$\leq \underbrace{\langle z_k - z_{k+1/2}, z_{k+1/2} - z_{k+1} \rangle}_{:=A} + \underbrace{\langle \tilde{z}_{k+1/2} - \tilde{z}_{k+1}, z_{k+1/2} - z_{k+1} \rangle}_{:=B} + \underbrace{\langle z_k - z_{k+1}, z_{k+1} - \tilde{z} \rangle}_{:=C}$$

Let us first simplify term B.

$$\begin{aligned}
B &:= \langle \tilde{z}_{k+1/2} - \tilde{z}_{k+1}, z_{k+1/2} - z_{k+1} \rangle \\
&= \langle \tilde{z}_{k+1/2} - z_k, z_{k+1/2} - z_{k+1} \rangle + \langle z_k - \tilde{z}_{k+1}, z_{k+1/2} - z_{k+1} \rangle && \text{(Add/subtract } z_k\text{)} \\
&= \eta \langle F(z_{k+1/2}) - F(z_k), z_{k+1/2} - z_{k+1} \rangle && \text{(Using the updates)} \\
&\leq \eta \left\| F(z_{k+1/2}) - F(z_k) \right\| \left\| z_{k+1/2} - z_{k+1} \right\| && \text{(Cauchy-Schwarz)} \\
&\leq (2L)\, \eta \left\| z_{k+1/2} - z_k \right\| \left\| z_{k+1/2} - z_{k+1} \right\| && \text{(Since } F \text{ is } 2L\text{-Lipschitz)} \\
\implies B &\leq \frac{1}{2} \left[4L^2 \eta^2 \left\| z_{k+1/2} - z_k \right\|^2 + \left\| z_{k+1/2} - z_{k+1} \right\|^2 \right] && \text{(Young's inequality)}
\end{aligned}$$

Extra-Gradient for smooth, convex-concave games

Recall that $\eta \langle F(z_{k+1/2}), z_{k+1/2} - \tilde{z} \rangle \leq A + B + C$ where $B \leq \frac{1}{2}\left[4L^2\eta^2 \left\|z_{k+1/2} - z_k\right\|^2 + \left\|z_{k+1/2} - z_{k+1}\right\|^2\right]$, $A := \left\langle z_k - z_{k+1/2}, z_{k+1/2} - z_{k+1} \right\rangle$ and $C := \langle z_k - z_{k+1}, z_{k+1} - \tilde{z} \rangle$. In order to simplify A, C, we will use $\langle a, b \rangle = \frac{\|a+b\|^2 - \|a\|^2 - \|b\|^2}{2}$.

$$A = \left\langle \underbrace{z_k - z_{k+1/2}}_{:=a}, \underbrace{z_{k+1/2} - z_{k+1}}_{:=b} \right\rangle = \frac{1}{2}\left[\|z_k - z_{k+1}\|^2 - \left\|z_k - z_{k+1/2}\right\|^2 - \left\|z_{k+1/2} - z_{k+1}\right\|^2\right]$$

$$C = \left\langle \underbrace{z_k - z_{k+1}}_{:=a}, \underbrace{z_{k+1} - \tilde{z}}_{:=b} \right\rangle = \frac{1}{2}\left[\|z_k - \tilde{z}\|^2 - \|z_k - z_{k+1}\|^2 - \|z_{k+1} - \tilde{z}\|^2\right]$$

$$\begin{aligned} &2[A + B + C] \\ &\leq \|z_k - z_{k+1}\|^2 - \left\|z_k - z_{k+1/2}\right\|^2 - \left\|z_{k+1/2} - z_{k+1}\right\|^2 + 4L^2\eta^2 \left\|z_{k+1/2} - z_k\right\|^2 + \left\|z_{k+1/2} - z_{k+1}\right\|^2 \\ &+ \|z_k - \tilde{z}\|^2 - \|z_k - z_{k+1}\|^2 - \|z_{k+1} - \tilde{z}\|^2 \\ &\implies 2[A + B + C] \leq \left\|z_k - z_{k+1/2}\right\|^2 \left(4L^2\eta^2 - 1\right) + \|z_k - \tilde{z}\|^2 - \|z_{k+1} - \tilde{z}\|^2 \end{aligned}$$

Extra-Gradient for smooth, convex-concave games

Putting everything together,

$$\eta \langle F(z_{k+1/2}), z_{k+1/2} - \tilde{z} \rangle \leq \frac{1}{2} \left[\left\| z_k - z_{k+1/2} \right\|^2 \left(4L^2\eta^2 - 1 \right) + \| z_k - \tilde{z} \|^2 - \| z_{k+1} - \tilde{z} \|^2 \right] \tag{5}$$

Setting $\eta = \frac{1}{2L}$,

$$\langle F(z_{k+1/2}), z_{k+1/2} - \tilde{z} \rangle \leq L \left[\| z_k - \tilde{z} \|^2 - \| z_{k+1} - \tilde{z} \|^2 \right]$$

Summing from $k = 1$ to T,

$$\sum_{k=1}^{T} \langle F(z_{k+1/2}), z_{k+1/2} - \tilde{z} \rangle \leq L \sum_{k=1}^{T} \left[\| z_k - \tilde{z} \|^2 - \| z_{k+1} - \tilde{z} \|^2 \right] = L \| z_1 - \tilde{z} \|^2 \leq 2D^2 L$$

(Since both $\mathcal{W}$ and $\mathcal{V}$ have diameter D)

Extra-Gradient for smooth, convex-concave games

Recall that $\sum_{k=1}^{T} \langle F(z_{k+1/2}), z_{k+1/2} - \tilde{z} \rangle \leq 2D^2 L$. Using Eq. (4) and dividing by T,

$$\frac{\sum_{k=1}^{T} [f(w_{k+1/2}, \tilde{v}) - f(\tilde{w}, v_{k+1/2})]}{T} \leq \frac{2D^2 L}{T}$$

Since $f(\cdot, \tilde{v})$ and $-f(\tilde{w}, \cdot)$ are convex, using Jensen's inequality and by definition of $\bar{w}_T$ and $\bar{v}_T$,

$$f(\bar{w}_T, \tilde{v}) - f(\tilde{w}, \bar{v}_T) \leq \frac{2D^2 L}{T}$$

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 $\tilde{w} \in \mathcal{W}$,

$$\max_{v \in \mathcal{V}} f(\bar{w}_T, v) - \min_{w \in \mathcal{W}} f(w, \bar{v}_T) \leq \frac{2D^2 L}{T} \implies \text{Duality Gap}((\bar{w}_T, \bar{v}_T)) \leq \frac{2D^2 L}{T}$$

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

Questions?

Extra-Gradient for smooth, strongly-convex strongly-concave games

Claim: For L-smooth, μ strongly-convex strongly-concave games, T iterations of projected EG with $\eta_k = \frac{1}{8L}$ results in the following bound,

$$\left\| \begin{bmatrix} w_T - w^* \\ v_T - v^* \end{bmatrix} \right\|^2 \leq \exp\left(\frac{-T}{8\kappa}\right) \left\| \begin{bmatrix} w_0 - w^* \\ v_0 - v^* \end{bmatrix} \right\|^2 .$$

Proof: Recall that if f is strongly-convex, strongly-concave, F is μ strongly-monotone i.e. $\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geq \mu \left\| z_1 - z_2 \right\|^2$ for all z_1, z_2. Also, recall that by using the definition of the optimality of z^*, we derived that $\langle F(z^*), z - z^* \rangle \geq 0$. Using Eq. (5) with $\tilde{z} = z^*$,

$$\eta \langle F(z_{k+1/2}), z_{k+1/2} - z^* \rangle \leq \frac{1}{2} \left[\left\| z_k - z_{k+1/2} \right\|^2 \left(4L^2\eta^2 - 1\right) + \left\| z_k - z^* \right\|^2 - \left\| z_{k+1} - z^* \right\|^2 \right]$$

Let us simplify the LHS,

$$\begin{aligned} \langle F(z_{k+1/2}), z_{k+1/2} - z^* \rangle &= \langle F(z_{k+1/2}) - F(z^*), z_{k+1/2} - z^* \rangle + \langle F(z^*), z_{k+1/2} - z^* \rangle \\ &\geq \langle F(z_{k+1/2}) - F(z^*), z_{k+1/2} - z^* \rangle \quad (\text{Since } \langle F(z^*), z - z^* \rangle \geq 0) \\ \implies \langle F(z_{k+1/2}), z_{k+1/2} - z^* \rangle &\geq \mu \left\| z_{k+1/2} - z^* \right\|^2 \quad (\text{By } \mu \text{ strong-monotonicity of } F) \end{aligned}$$

Extra-Gradient for smooth, strongly-convex strongly-concave games

We have $\eta\,\mu\left\|z_{k+1/2}-z^*\right\|^2 \le \frac{1}{2}\left[\left\|z_k-z_{k+1/2}\right\|^2\left(4L^2\eta^2-1\right)+\|z_k-z^*\|^2-\|z_{k+1}-z^*\|^2\right]$.
Further lower-bounding the LHS,

$$\|z_k-z^*\|^2=\left\|z_k-z_{k+1/2}+z_{k+1/2}-z^*\right\|^2\le 2\left\|z_k-z_{k+1/2}\right\|^2+2\left\|z_{k+1/2}-z^*\right\|^2$$
$$\implies 2\left\|z_{k+1/2}-z^*\right\|^2\ge\|z_k-z^*\|^2-2\left\|z_k-z_{k+1/2}\right\|^2$$

Putting everything together,

$$\eta\mu\left[\|z_k-z^*\|^2-2\left\|z_k-z_{k+1/2}\right\|^2\right]\le\left\|z_k-z_{k+1/2}\right\|^2\left(4L^2\eta^2-1\right)+\|z_k-z^*\|^2-\|z_{k+1}-z^*\|^2$$
$$\implies \|z_{k+1}-z^*\|^2\le(1-\mu\eta)\|z_k-z^*\|^2+\left\|z_k-z_{k+1/2}\right\|^2\left(4L^2\eta^2-1+2\mu\eta\right)$$
$$\|z_{k+1}-z^*\|^2\le\left(1-\frac{\mu}{8L}\right)\|z_k-z^*\|^2+\left\|z_k-z_{k+1/2}\right\|^2\underbrace{\left(4L^2\eta^2-1+2\mu\eta\right)}_{<0\text{ for }\eta=\frac{1}{8L}}\quad(\text{Setting }\eta=\tfrac{1}{8L})$$

Extra-Gradient for smooth, strongly-convex strongly-concave games

Recall that $\|z_{k+1} - z^*\|^2 \leq \left(1 - \frac{\mu}{8L}\right) \|z_k - z^*\|^2$. Recursing from $k = 0$ to $T - 1$,

$$\|z_T - z^*\|^2 \leq \left(1 - \frac{\mu}{8L}\right)^T \|z_0 - z^*\|^2 \leq \exp\left(\frac{-T}{8\kappa}\right) \|z_0 - z^*\|^2$$

Hence, compared to GDA that has an $O\left(\kappa^2 \log(1/\epsilon)\right)$ convergence for strongly-convex strongly-concave games, EG has an $O\left(\kappa \log(1/\epsilon)\right)$ convergence.

Questions?

Wrapping Up - What we covered

- Considered optimizing a taxonomy of functions: (i) non-smooth but G-Lipschitz vs L-smooth, (ii) non-convex vs convex vs strongly-convex. Identified solution concepts ((gradient norm and convergence to a stationary point, distance to the minimizer).
- Studied and analyzed the convergence of (projected) gradient descent, Polyak momentum, Nesterov acceleration and the Newton method.
- Studied stochastic gradient descent and analyzed its convergence. Considered ideas to make SGD more robust to the step-size and the concept of variance reduction (E.g. SVRG).
- Considered the online convex optimization setting, and studied the notion of regret. Analyzed the convergence of OGD, FTL and FTRL. Used the online setting to motivate adaptive gradient methods (AdaGrad, Adam, AMSGrad) and analyzed their convergence.
- Considered min-max optimization and identified solution concepts (duality gap and distance to the Nash equilibrium) for convex-concave games. Analyzed the convergence of Gradient Descent Ascent and the Extra-Gradient Method.

Wrapping Up - What we could not cover

- Mirror Descent and its convergence (useful for optimization on the space of probabilities) [Bubeck, Chapter 4]
- Proximal Methods (useful for handling non-smooth regularization terms) [https://www.cs.ubc.ca/~schmidtm/Courses/5XX-S20/S6.pdf]
- (Block) Coordinate Descent (useful for functions that are separable in the coordinates) [https://www.cs.ubc.ca/~schmidtm/Courses/5XX-S20/S8.pdf]

Other important topics in Optimization for ML

- Constrained Optimization
- Global Optimization
- Multi-objective Optimization
- Distributed Optimization

References i

- 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.