

# CMPT 409/981: Optimization for Machine Learning

## Lecture 22

---

Sharan Vaswani

November 26, 2024

- **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 **Nash equilibrium** 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  $\eta$ , (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 convex-concave games:** Projected GDA has the guarantee that  $\text{Duality Gap}((\bar{w}_T, \bar{v}_T)) \leq \frac{4DG}{\sqrt{T}}$  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  $O(1/\sqrt{T})$  decrease on the duality gap.
- **Strongly-convex, strongly-concave games:**  $f(\cdot, v)$  is  $\mu_w$  strongly-convex and  $f(w, \cdot)$  is  $\mu_v$  strongly-concave. The Nash equilibrium  $(w^*, v^*)$  is unique. GDA converges to  $(w^*, v^*)$  at an  $O(\kappa^2 \log(\frac{1}{\epsilon}))$  rate.

# 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  $\mathbb{R}^{d_w+d_v} \rightarrow \mathbb{R}^{d_w+d_v}$  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  $v \in \mathcal{V}$ ,

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

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

$$\implies \left\langle \begin{bmatrix} \nabla_w f(w^*, v^*) \\ -\nabla_v f(w^*, v^*) \end{bmatrix}, \begin{bmatrix} w \\ v \end{bmatrix} - \begin{bmatrix} w^* \\ v^* \end{bmatrix} \right\rangle \geq 0 \implies \langle F(z^*), z - z^* \rangle \geq 0 \quad (1)$$

**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:**

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

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

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

$$\langle \tilde{z}_{k+1} - z_{k+1}, z - z_{k+1} \rangle \leq 0 \implies \langle -\tilde{z}_{k+1}, z_{k+1} - z \rangle \leq \langle -z_{k+1}, z_{k+1} - z \rangle \quad (3)$$

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

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

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

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

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

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

$$\leq \left\langle \frac{z_k - \tilde{z}_{k+1}}{\eta}, z_{k+1/2} - z_{k+1} \right\rangle + \left\langle \frac{z_k - z_{k+1}}{\eta}, z_{k+1} - \tilde{z} \right\rangle \quad (\text{Using eq. (3) for the second term})$$

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

$$\leq \left\langle \frac{z_k - z_{k+1/2}}{\eta}, z_{k+1/2} - z_{k+1} \right\rangle + \left\langle \frac{\tilde{z}_{k+1/2} - \tilde{z}_{k+1}}{\eta}, z_{k+1/2} - z_{k+1} \right\rangle + \left\langle \frac{z_k - z_{k+1}}{\eta}, z_{k+1} - \tilde{z} \right\rangle \quad (\text{Using eq. (2) for the first term})$$



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

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

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

Let us first simplify term B.

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

## 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

$B \leq \frac{1}{2} \left[ 4L^2\eta^2 \|z_{k+1/2} - z_k\|^2 + \|z_{k+1/2} - z_{k+1}\|^2 \right]$ ,  $A := \langle z_k - z_{k+1/2}, z_{k+1/2} - z_{k+1} \rangle$  and

$C := \langle z_k - z_{k+1}, z_{k+1} - \tilde{z} \rangle$ .

In order to simplify  $A$ ,  $C$ , we will use  $\langle a, b \rangle = \frac{\|a+b\|^2 - \|a\|^2 - \|b\|^2}{2}$ .

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

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

$$2[A + B + C]$$

$$\leq \|z_k - z_{k+1}\|^2 - \|z_k - z_{k+1/2}\|^2 - \|z_{k+1/2} - z_{k+1}\|^2 + 4L^2\eta^2 \|z_{k+1/2} - z_k\|^2 + \|z_{k+1/2} - z_{k+1}\|^2 \\ + \|z_k - \tilde{z}\|^2 - \|z_k - z_{k+1}\|^2 - \|z_{k+1} - \tilde{z}\|^2$$

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

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

Putting everything together,

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

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  $T$ ,

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

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

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

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

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

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

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

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

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

- 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].

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

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

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

- Hence, compared to GDA that has an  $O(\kappa^2 \log(1/\epsilon))$  convergence for strongly-convex strongly-concave games, EG has an  $O(\kappa \log(1/\epsilon))$  convergence and matches the rate for smooth, strongly-convex minimization.

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

- 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



-  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.