AIE6003: Optimization for Machine Learning

Xiaopeng Li
CUHK(SZ)
SAI

May 5, 2026

Outline

Final Review: Duality

Final Review: Low-rank & Sparse Optimization

Final Review: Minimax Optimization

Lagrangian Duality

Primal: $\min_x f(x)$ s.t. $h_i(x) \leq 0$, $e_j(x) = 0$.

Lagrangian:

$$L(x, \lambda, \nu) = f(x) + \sum_i \lambda_i h_i(x) + \sum_j \nu_j e_j(x).$$

Dual function: $g(\lambda, \nu) = \inf_x L(x, \lambda, \nu)$.

- Always concave in (λ, ν), regardless of f, h_i, e_j.
- Weak duality: $g(\lambda, \nu) \leq p^*$ for any $\lambda \geq 0$.

Dual problem: $\max_{\lambda \geq 0,\, \nu} g(\lambda, \nu)$, value d^*.

Strong duality: $d^* = p^*$ holds for convex problems under Slater's condition: a strictly feasible point exists for the inequality constraints.

KKT Conditions

Under convexity & Slater, x^* is optimal iff there exist (λ^*, ν^*) such that:

- Stationarity: $0 \in \partial f(x^*) + \sum_i \lambda_i^* \nabla h_i(x^*) + \sum_j \nu_j^* \nabla e_j(x^*)$.
- Primal feasibility: $h_i(x^*) \leq 0$, $e_j(x^*) = 0$.
- Dual feasibility: $\lambda_i^* \geq 0$.
- Complementary slackness: $\lambda_i^* h_i(x^*) = 0$ for all i.

Common mistakes:

- Forgetting sign conditions on multipliers.
- Forgetting complementary slackness.
- Using KKT without verifying Slater (only necessary conditions).

Normal Cones

Definition: For closed convex C and $x \in C$,

$$N_C(x) = \{g : \langle g, y - x \rangle \leq 0,\ \forall\, y \in C\}.$$

Polyhedral $C = \{x : a_i^\top x \leq b_i,\ i \in I\}$**:**

$$N_C(x) = \left\{ \sum_{i \in I(x)} \lambda_i a_i : \lambda_i \geq 0 \right\}, \quad I(x) = \{i : a_i^\top x = b_i\}.$$

i.e., nonnegative combinations of active-constraint gradients.

Examples:

- $C = \mathbb{R}^n$: $N_C(x) = \{0\}$.
- $C = \{x \geq 0\}$: $\{g : g \leq 0,\ g_i x_i = 0\}$.
- C is probability simplex: $\{\alpha \mathbf{1} + \sum_{j : x_j = 0} \beta_j e_j : \alpha \in \mathbb{R}, \beta_j \leq 0\}$

For $\min f(x)$ s.t. $x \in C$, KKT is packed into $-\nabla f(x^*) \in N_C(x^*)$.

Sample Final Q1(a)

Consider the optimization problem

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & f(x) := \frac{1}{2}\|Ax-b\|^2 + \frac{\lambda}{2}\|x\|^2. \\ \text{s.t.} \quad & \mathbf{1}^\top x = 1, \quad x \geq 0, \quad \ell^\top x \leq B, \quad g^\top x \leq r, \end{aligned}$$

Write the inequality constraints as a polyhedron:

$$\Omega = \{x : -x \leq 0,\ \ell^\top x - B \leq 0,\ g^\top x - r \leq 0\}.$$

For a polyhedron $\Omega = \{x : Cx \leq d\}$, we have learned

$$N_\Omega(x) = \{C^\top \mu : \mu \geq 0,\ \mu_j(C_j x - d_j) = 0\}.$$

Therefore, identifying what C is,

$$N_\Omega(x) = \left\{ -s + \alpha\ell + \beta g \;\middle|\; \begin{array}{lll} s \geq 0, & \alpha,\beta \geq 0, & s_i x_i = 0, \\ \alpha(\ell^\top x - B) = 0, & & \beta(g^\top x - r) = 0. \end{array} \right\},$$

Sample Final Q1(a)

The equality constraint is $\mathbf{1}^\top x = 1$.

Let $\mathcal{A} = \{x : Dx = b\}$. For $x \in \Omega \cap \mathcal{A}$ with $\operatorname{relint}(\Omega) \cap \mathcal{A} \neq \emptyset$,

$$N_{\Omega \cap \mathcal{A}}(x) = N_\Omega(x) + \{D^\top \lambda : \lambda \in \mathbb{R}^m\}.$$

For a convex differentiable problem,

$$x^* \text{ optimal} \quad \Longleftrightarrow \quad 0 \in \nabla f(x^*) + N_\Omega(x^*) + \{\nu \mathbf{1} : \nu \in \mathbb{R}\}.$$

Since $\nabla f(x) = A^\top(Ax - b) + \lambda x$, there exist $s^* \geq 0$, $\alpha^* \geq 0$, $\beta^* \geq 0$, and $\nu^* \in \mathbb{R}$ such that

$$A^\top(Ax^* - b) + \lambda x^* + \nu^* \mathbf{1} + \alpha^* \ell + \beta^* g - s^* = 0.$$

Together with

$$\mathbf{1}^\top x^* = 1, \quad x^* \geq 0, \quad \ell^\top x^* \leq B, \quad g^\top x^* \leq r,$$

$$s_i^* x_i^* = 0, \quad \alpha^*(\ell^\top x^* - B) = 0, \quad \beta^*(g^\top x^* - r) = 0.$$

Slater holds, so KKT is necessary and sufficient.

Fenchel Conjugate

Definition: for $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$,

$$f^*(p) := \sup_x \{\langle p, x\rangle - f(x)\}.$$

Properties:

- f^* is always closed and convex.
- Biconjugate: if f is closed convex, $f^{**} = f$.
- Fenchel–Young: $f(x) + f^*(p) \geq \langle x, p\rangle$, with equality iff $p \in \partial f(x)$.

Useful examples:

- quadratic function $f(x) = \frac{1}{2}x^\top Q x$, $Q \succ 0$: $f^*(p) = \frac{1}{2}p^\top Q^{-1} p$.
- support function $f = \delta_C$: $f^*(p) = \sup_{x \in C} \langle p, x\rangle$.
- indicator function $f = \|\cdot\|$: $f^* = \delta_{\{p : \|p\|_* \leq 1\}}$.
- indicator function $f = \|\cdot\|_1$: $f^* = \delta_{[-1,1]^n}$.

Smoothness/Strong-Convexity Duality

Theorem: f is closed and μ-strongly convex iff f^* is finite-valued, differentiable, and $1/\mu$-smooth.

Applications:

- When primal f is SC, dual $g(\lambda)$ is smooth in λ.
- Smoothness of g allows to apply GD on λ.
- Connecting primal regularization and dual algorithm convergence.

Useful Properties: $\nabla f^*(p) = \arg\max_x \{\langle p, x\rangle - f(x)\}$.

- Computing ∇f^* at p is solving an unconstrained subproblem.
- For $f(x) = \frac{1}{2}x^\top Q x - c^\top x$: $\nabla f^*(p) = Q^{-1}(p + c)$.

Dual Function via the Conjugate

Equality-constrained problem: $\min_x f(x)$ s.t. $Ax = b$.

$$g(\lambda) = \inf_x \{f(x) + \lambda^\top (Ax - b)\} = -f^*(-A^\top \lambda) - b^\top \lambda.$$

Dual gradient: if f is SC,

$$\nabla g(\lambda) = Ax(\lambda) - b, \quad x(\lambda) = \nabla f^*(-A^\top \lambda) = \arg\min_x L(x, \lambda).$$

The dual gradient is the primal residual at the inner minimizer.

Inequality constraints $Ax \leq b$**:** same formula plus the projection $\lambda \geq 0$.

Smoothness constant of $-g$**:** when f is μ-SC,

$$L_g = \frac{\|A\|_{\text{op}}^2}{\mu},$$

since $-g = f^* \circ (-A^\top) + b^\top(\cdot)$ and f^* is $1/\mu$-smooth.

Dual Gradient Method

Update: with stepsize $1/L_g$,

$$\lambda_{k+1} = \lambda_k + L_g^{-1} \nabla g(\lambda_k).$$

For inequality dual variables, project: $[\cdot]_+$.

Rate: for f μ-SC, A with full-row rank,

$$g(\lambda^*) - g(\lambda_T) \leq \frac{L_g \|\lambda_0 - \lambda^*\|^2}{2T}, \quad L_g = \frac{\|A\|_{\text{op}}^2}{\mu}.$$

Why this works?

- $h := -g$ is convex and L_g-smooth $\Rightarrow$ standard GD on h.
- One-step relation $h(\lambda_{k+1}) - h^* \leq \frac{L_g}{2}(\|\lambda_k - \lambda^*\|^2 - \|\lambda_{k+1} - \lambda^*\|^2)$.
- Telescope + monotone descent of $h(\lambda_k) \Rightarrow 1/T$ rate.

Same proof template as convex-smooth GD.

Sample Final Q1(c)

Setup: $\min f(x)$ s.t. $Ex = h$ with f λ-SC; dual $g(u)$ is

$$g(u) = -f^*(-E^\top u) - u^\top h.$$

Technique:

- f λ-SC $\implies$ f^* is $1/\lambda$-smooth.
- Chain rule: $\nabla g(u) = E\nabla f^*(-E^\top u) - h$.
- Lipschitz: $L_g = \|E\|_{\text{op}}^2 / \lambda$.

Apply convex-smooth GD on $H = -g$ with stepsize $1/L_g$.

Final rate:

$$g(u^*) - g(u_T) \leq \frac{L_g \|u_0 - u^*\|^2}{2T}.$$

Augmented Lagrangian (ALM)

Augmented Lagrangian:

$$L_\rho(x, \lambda) = f(x) + \lambda^\top (Ax - b) + \frac{\rho}{2} \|Ax - b\|^2 .$$

Iteration:

- $x^{k+1} = \arg\min_x L_\rho(x, \lambda^k)$.
- $\lambda^{k+1} = \lambda^k + \rho(Ax^{k+1} - b)$.

Fact: ALM is just proximal-point method on $-g$ with stepsize ρ:

$$\lambda^{k+1} = \arg\max_\lambda \left\{ g(\lambda) - \frac{1}{2\rho} \left\| \lambda - \lambda^k \right\|^2 \right\}.$$

Rate: $g^* - g(\lambda^K) \leq \|\lambda^0 - \lambda^*\|^2 / (2\rho K)$.

- Larger $\rho \implies$ faster dual progress, harder primal subproblem.
- Convergence does not require primal SC.

ADMM

Two-block problem: $\min_{x,z} f(x) + h(z)$ s.t. $Ax + Bz = c$.

Scaled ADMM: with scaled dual $w = \lambda/\rho$,

- $x^{k+1} = \arg\min_x f(x) + \frac{\rho}{2} \left\| Ax + Bz^k - c + w^k \right\|^2$.
- $z^{k+1} = \arg\min_z h(z) + \frac{\rho}{2} \left\| Ax^{k+1} + Bz - c + w^k \right\|^2$.
- $w^{k+1} = w^k + (Ax^{k+1} + Bz^{k+1} - c)$.

Classical z-updates using proximal operator:

- $h = \tau \|\cdot\|_1 \implies$ entrywise soft-thresholding $\mathcal{S}_{\tau/\rho}$.
- $h = \tau \|\cdot\|_* \implies$ singular-value thresholding $\mathrm{SVT}_{\tau/\rho}$.
- $h = \delta_C \implies$ projection Π_C.

Splitting tricks: introduce auxiliary variables to decompose non-smooth terms, then apply ADMM.

Sample Final Q1(b)

Setup: $\min \frac{1}{2} \|Ax - b\|^2 + \frac{\lambda}{2} \|x\|^2 + \tau \|Dx\|_1$.

Technique:

- Splitting: introduce $z = Dx$ to decouple $\|\cdot\|_1$.
- Constraint: $Dx - z = 0$.
- Scaled augmented Lagrangian: $f(x) + \tau \|z\|_1 + \frac{\rho}{2} \|Dx - z + w\|^2$.

ADMM updates:

- $x^{k+1} = (A^\top A + \lambda I + \rho D^\top D)^{-1}(A^\top b + \rho D^\top (z^k - w^k))$.
- $z^{k+1} = \mathcal{S}_{\tau/\rho}(Dx^{k+1} + w^k)$.
- $w^{k+1} = w^k + Dx^{k+1} - z^{k+1}$.

Outline

Final Review: Duality

Final Review: Low-rank & Sparse Optimization

Final Review: Minimax Optimization

Nuclear Norm and Subdifferential

Nuclear norm: $\|X\|_* = \sum_i \sigma_i(X)$, the convex envelope of $\mathrm{rank}(X)$ on the spectral unit ball.

Subdifferential: take compact SVD $X = U\Sigma V^\top$ with $\Sigma \succ 0$, then

$$\partial \|X\|_* = \{UV^\top + W : U^\top W = 0,\ WV = 0,\ \|W\|_{\mathrm{op}} \le 1\}.$$

Interpretation:

- $UV^\top$ is the analogue of $\mathrm{sign}(x)$ for the entrywise ℓ_1 norm.
- The correction W lives in the orthogonal complement of $\mathrm{span}(U) \oplus \mathrm{span}(V)$ and has bounded operator norm.
- At $X = 0$: $\partial \|0\|_* = \{W : \|W\|_{\mathrm{op}} \le 1\}$.

Singular Value Thresholding (SVT)

Goal: solve $\operatorname{prox}_{\mu\|\cdot\|_*}(M) = \arg\min_X \frac{1}{2}\|X - M\|_F^2 + \mu\|X\|_*$.

Closed form: let $M = \overline{U}\,\overline{\Sigma}\,\overline{V}^\top$ with $\overline{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_{\min(m,n)})$.

$$\mathrm{SVT}_\mu(M) = \overline{U}\,\operatorname{diag}(\max\{\sigma_i - \mu, 0\})\,\overline{V}^\top.$$

Optimality: $0 \in X - M + \mu\partial\|X\|_*$, i.e. $M - X \in \mu\partial\|X\|_*$.

Connection to ℓ_1: SVT_μ is the matrix analogue of soft-thresholding, applied at the level of singular values.

Frank-Wolfe (Conditional Gradient)

Setup: $\min_{x \in C} f(x)$, C compact convex, f L-smooth.

Iteration:

- $s_t \in \arg\min_{s \in C} \langle \nabla f(x_t), s \rangle$ (linear minimization oracle, LMO).
- $x_{t+1} = (1 - \eta_t) x_t + \eta_t s_t$, with $\eta_t = \frac{2}{t+2}$.

Advantage:

- LMO often much cheaper than projection (e.g. nuclear-norm ball: top singular pair vs. full SVT).
- Iterate is a convex combination of LMO outputs, so sparse/low-rank structure is preserved.

Rate: using smoothness, FW oracle inequality, and induction with $\eta_t = 2/(t+2)$, we obtain

$$f(x_t) - f(x^*) \leq \frac{2L \operatorname{diam}(C)^2}{t+2}.$$

Sample Final Q2(a)

Setup: $\mathcal{D}_\tau = \{X \in \mathbb{S}^n : X \succeq 0, \operatorname{tr}(X) \le \tau\}$. Solve $S_t \in \arg\min_{S \in \mathcal{D}_\tau} \langle \nabla F(X_t), S \rangle$.

Technique:

- Let $G_t = \nabla F(X_t)$; symmetric (by symmetry of Y, X_t, Ω).
- For any $S \succeq 0$: $\langle G_t, S \rangle \ge \lambda_{\min}(G_t) \operatorname{tr}(S)$.
- If $\lambda_{\min}(G_t) < 0$, push the trace budget into the most negative eigendirection: $S_t = \tau v_t v_t^\top$ for v_t a unit eigenvector of $\lambda_{\min}$.
- If $\lambda_{\min}(G_t) \ge 0$, take $S_t = 0$.

How to prove $\langle G_t, S \rangle \ge \lambda_{\min}(G_t) \operatorname{tr}(S)$? Try to diagonalize G_t!

Burer-Monteiro Factorization

Idea: replace X of expected rank $\leq r$ by a low-rank factorization:

- Symmetric PSD: $X = UU^\top$, $U \in \mathbb{R}^{n\times r}$.
- General: $X = UV^\top$, $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$.

Pros & cons:

- Pro: memory drops from mn to $(m+n)r$.
- Pro: no need for SVT/projection.
- Con: objective in U, V is non-convex.
- Con: representation is non-unique.

Gradient computation: for $\Phi(U,V) = \frac{1}{2}\left\|P_\Omega(UV^\top - Y)\right\|_F^2$:

$$\nabla_U \Phi = P_\Omega(UV^\top - Y)V, \quad \nabla_V \Phi = P_\Omega(UV^\top - Y)^\top U.$$

Sample Final Q2(b)

Setup: $\Phi(U) = \frac{1}{4}\left\|P_\Omega(UU^\top - Y)\right\|_F^2$, $U \in \mathbb{R}^{n\times k}$.

Technique:

- Let $R(U) = P_\Omega(UU^\top - Y)$; symmetric since Y, Ω are.
- Differential: $R(U + \varepsilon H) = R(U) + \varepsilon P_\Omega(HU^\top + UH^\top) + O(\varepsilon^2)$.
- $d\Phi(U)[H] = \frac{1}{2}\left\langle R(U), HU^\top + UH^\top \right\rangle$.
- Symmetry of R: $\left\langle R, HU^\top + UH^\top \right\rangle = 2\left\langle RU, H\right\rangle$.
- Final: $\nabla\Phi(U) = P_\Omega(UU^\top - Y)U$.

LASSO and Soft-Thresholding

LASSO: $\min_x \ \frac{1}{2}\|Ax-b\|^2 + \lambda \|x\|_1$.

Soft-thresholding: prox of $\lambda\|\cdot\|_1$

$$[\mathcal{S}_\lambda(z)]_i = \operatorname{sign}(z_i)\max\{|z_i| - \lambda, 0\}.$$

Subdifferential of $\|\cdot\|_1$:

- $v_i = \operatorname{sign}(x_i)$ if $x_i \neq 0$;
- $v_i \in [-1,1]$ if $x_i = 0$.

Closed form: $\min_\xi \ \frac{a}{2}\xi^2 - c\xi + \lambda|\xi| \implies \xi^* = \frac{1}{a}\mathcal{S}_\lambda(c)$.

This 1D update is the foundation of coordinate descent on LASSO.

Coordinate Descent for LASSO

Setup: fix all coordinates except x_i in $\frac{1}{2}\|Ax-b\|^2+\lambda\|x\|_1$.

Partial residual: $r_{-i} := b - \sum_{j\neq i} A_{:j}x_j$.

1D subproblem:

$$\min_{\xi} \frac{1}{2}\|A_{:i}\xi - r_{-i}\|^2 + \lambda|\xi| \iff \min_{\xi} \frac{\|A_{:i}\|^2}{2}\xi^2 - (A_{:i}^\top r_{-i})\xi + \lambda|\xi|.$$

Coordinate update:

$$x_i^+ = \frac{1}{\|A_{:i}\|^2}\mathcal{S}_\lambda(A_{:i}^\top r_{-i}) = \frac{1}{\|A_{:i}\|^2}\mathcal{S}_\lambda\big(\|A_{:i}\|^2 x_i - A_{:i}^\top e\big),$$

where $e = Ax-b$ is the current residual; $x_i^+ = 0$ if $A_{:i}=0$.

Sample Final Q2(c)

Setup: $H(\theta) = \frac{1}{2}\|B\theta - r\|^2 + \alpha\|\theta\|_1$.

Technique:

- Fix all θ_j, $j \neq i$; set $r_{-i} := r - \sum_{j\neq i} b_j\theta_j = r - B\theta + b_i\theta_i$.
- 1D problem: $\min_\xi \frac{1}{2}\|b_i\xi - r_{-i}\|^2 + \alpha|\xi|$.
- Expand: $\frac{\|b_i\|^2}{2}\xi^2 - (b_i^\top r_{-i})\xi + \alpha|\xi| + \text{const}$.
- Closed form via $\xi^* = \frac{1}{a}\mathcal{S}_\lambda(c)$ from the soft-thresholding.

Update: if $b_i \neq 0$

$$\theta_i^+ = \frac{1}{\|b_i\|^2}\mathcal{S}_\alpha(b_i^\top r_{-i}) = \frac{1}{\|b_i\|^2}\mathcal{S}_\alpha(\|b_i\|^2\theta_i - b_i^\top e),$$

where $e = B\theta - r$ is the current residual. If $b_i = 0$, then $\theta_i^+ = 0$.

Outline

Final Review: Duality

Final Review: Low-rank & Sparse Optimization

Final Review: Minimax Optimization

Saddle Points

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

Saddle point: for (x^*, y^*),

$$f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*), \quad \forall x \in \mathcal{X},\, y \in \mathcal{Y}.$$

Weak duality: $\min_x \max_y f \geq \max_y \min_x f$.

Sion's theorem: if f is convex in x, concave in y, $\mathcal{X}, \mathcal{Y}$ compact convex,

$$\min_x \max_y f = \max_y \min_x f,$$

and a saddle point exists.

Duality gap:

$$\text{gap}(x, y) := \max_{y'} f(x, y') - \min_{x'} f(x', y) \geq 0,$$

zero iff (x, y) is a saddle point. Standard convergence criterion.

Sample Final Q3(a)

Setup: $\chi(x, y) = x^\top B y + a^\top x - c^\top y$ on $\Delta_m \times \Delta_n$. What is gap?

Technique:

- For fixed x, $\max_{y \in \Delta_n} \chi(x, y) = a^\top x + \max_{1 \le j \le n}((B^\top x)_j - c_j)$.
- For fixed y, $\min_{x \in \Delta_m} \chi(x, y) = \min_{1 \le i \le m}((By)_i + a_i) - c^\top y$.
- Gap = (max over y') − (min over x'):

$$\mathrm{gap}(x, y) = a^\top x + c^\top y + \max_j (B^\top x - c)_j - \min_i (By + a)_i.$$

Lemma. For $d \in \mathbb{R}^n$, $\max_{y \in \Delta_n} d^\top y = \max_j d_j$, attained at $y = e_{j^*}$ with $j^* \in \arg\max_j d_j$.

Proof. Let $d_{\max} = \max_j d_j$. For any $y \in \Delta_n$,

$$d^\top y = \sum\nolimits_j d_j y_j \le \sum\nolimits_j d_{\max} y_j = d_{\max} \sum\nolimits_j y_j = d_{\max},$$

using $d_j \le d_{\max}$ and $y_j \ge 0$. Equality at $y = e_{j^*}$: $d^\top e_{j^*} = d_{j^*} = d_{\max}$.

Saddle Operator and Monotonicity

Saddle gradient field: $F(z) = (\nabla_x f(x,y), -\nabla_y f(x,y))$, $z = (x,y)$.

Characterization: z^* is an interior saddle point iff $F(z^*) = 0$.

Monotonicity:

- F monotone: $\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geq 0 \iff f$ is C-C.
- F μ-strongly monotone: $\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geq \mu \|z_1 - z_2\|^2$ if and only if f is SC-SC.

Lipschitzness: $\|F(z_1) - F(z_2)\| \leq \ell \|z_1 - z_2\|$.

The saddle problem becomes equivalent to finding a zero of F, and the proof techniques is similar to SC & Lipschitz GD.

Gradient Descent Ascent (GDA)

Update: $z_{k+1} = z_k - \eta F(z_k)$.

Strongly monotone + Lipschitz F:

- Optimal stepsize $\eta = \mu/\ell^2$.
- Linear contraction: $\|z_{k+1} - z^*\|^2 \leq (1 - \mu^2/\ell^2)\,\|z_k - z^*\|^2$.
- Same proof skeleton as SC-smooth GD.

Monotone Lipschitz F:

- Average iterate has $\mathrm{gap} = O(1/T)$.
- Last-iterate may not converge.

Bilinear $f(x, y) = x^\top B y$:

- GDA diverges: distance to z^* strictly grows every step.
- Need EG or OGDA to converge.

Sample Final Q3(b)

Setup: bilinear min-max problem with B nonsingular; saddle point z^*.

Technique:

- Center: $\xi_t = x_t - x^*$, $v_t = y_t - y^*$.
- Centered GDA: $\xi_{t+1} = \xi_t - \eta B v_t$, $v_{t+1} = v_t + \eta B^\top \xi_t$.
- Compute $\|\xi_{t+1}\|^2 + \|v_{t+1}\|^2$: cross terms $-2\eta \langle \xi_t, B v_t \rangle$ and $+2\eta \left\langle v_t, B^\top \xi_t \right\rangle$ cancel exactly.

Result:

$$\|\xi_{t+1}\|^2 + \|v_{t+1}\|^2 = \|\xi_t\|^2 + \|v_t\|^2 + \eta^2 \|B v_t\|^2 + \eta^2 \left\|B^\top \xi_t\right\|^2.$$

B nonsingular, so the right-hand side is strictly larger unless $z_t = z^*$.

Beyond GDA: EG, OGDA

Extragradient (EG):

- $z_{k+1/2} = z_k - \eta F(z_k)$.
- $z_{k+1} = z_k - \eta F(z_{k+1/2})$.
- Two gradient calls.

Optimistic GDA (OGDA):

- $z_{k+1} = z_k - 2\eta F(z_k) + \eta F(z_{k-1})$.
- One-step memory; uses $F(z_{k-1})$ to extrapolate.

EG/OGDA vs. GDA:

- EG/OGDA recover linear convergence on SC-SC games.
- Both converge on bilinear games where GDA spirals out, at the price of either two gradient calls (EG) or one-step memory (OGDA).

Sample Final Q3(c)

Setup: OGDA $z_{t+1} = z_t - 2\eta F(z_t) + \eta F(z_{t-1})$ on the bilinear game.

Technique:

- SVD: $B = U\Sigma V^\top$. Set $\bar{\xi}_t = U^\top \xi_t$, $\bar{v}_t = V^\top v_t$, decoupled.
- Each j: 2-term recurrence in $(\bar{\xi}_{j,t}, \bar{v}_{j,t})$.
- Complex variable $w_{j,t} = \bar{\xi}_{j,t} + i\bar{v}_{j,t}$, $c_j = \eta\sigma_j$:

$$w_{j,t+1} = (1 + 2ic_j)w_{j,t} - ic_j w_{j,t-1}.$$

Stability analysis: characteristic equation $r^2 - (1 + 2ic_j)r + ic_j = 0$.

- For $0 < c_j \le 1/2$: both roots have $|r_\pm|^2 \le \frac{1 \pm \sqrt{1-4c_j^2}}{2} < 1$.
- For $c_j > 1/2$: $|r_+|^2 < 1 \iff c_j < 1/\sqrt{3}$.

Final answer: OGDA contracts on every mode iff $c_j := \eta\sigma_j < 1/\sqrt{3}$ for the largest mode $\sigma_{\max} = \|B\|_{\mathrm{op}}$.

Key Takeaways

Master the workhorses: KKT via normal cones, conjugates, smoothness/SC duality, ADMM splitting, FW oracle, soft-thresholding.

For dual methods: f is μ-SC $\implies$ g smooth $\implies$ standard convex-smooth GD machinery applies.

For matrix structure: nuclear norm and its subdifferential drive SVT, matrix completion, and the FW LMO.

For saddle problems: monotonicity replaces convexity; the bilinear example motivates EG/OGDA.

Classical proof patterns: Expand, bound, then telescope.

Good Luck on the Final!