Lecture 11: Frank–Wolfe and Non-Euclidean Descent

TTIC 31070 / CMSC 35470 / BUSF 36903 / STAT 31015

Convex Optimization

Prof. Zhiyuan Li

Spring 2026

From Projection to Linear Minimization

Lectures 7–9 update on K: solve a nonlinear local subproblem over K

$$x_{t+1} \in \operatorname*{argmin}_{x \in K} \left\{ \langle \nabla f(x_t), x - x_t \rangle + \tfrac{1}{2\eta_t} \|x - x_t\|^2 \right\} \quad \text{or Bregman analog.}$$

For some K (simplex, spectrahedron, ℓ_1 ball, polytopes), even the projection / proximal step is expensive — but *linear minimization* is cheap:

$$s_t \in \operatorname*{argmin}_{s \in K} \langle \nabla f(x_t), s \rangle.$$

Frank–Wolfe (1956). Replace the projection primitive by an LMO; never leave K.

- No norm needed; no symmetric / centered K.
- Compactness $\Rightarrow$ LMO output exists.
- Convex combination keeps iterates in K.

Frank–Wolfe Update

Linear minimization oracle (LMO). For $K \subseteq E$ compact convex and $g \in E^*$,

$$\mathrm{LMO}_K(g) := \operatorname*{argmin}_{s \in K} \langle g, s \rangle \quad (\neq \emptyset).$$

Definition 11.1 (Frank–Wolfe update). f differentiable on a neighborhood of K. Given $x_t \in K$, choose $s_t \in \mathrm{LMO}_K(\nabla f(x_t))$ and stepsize $\gamma_t \in [0, 1]$:

$$x_{t+1} := (1 - \gamma_t)\, x_t + \gamma_t\, s_t.$$

Geometry. Linearize f at $x_t \to$ pull toward the LMO atom $s_t \to$ take a convex combination. Iterate stays in K by convexity.

Curvature Constant

Definition 11.2 (FW curvature). $K \subseteq E$ compact convex; f diff'ble. The curvature constant of f over K is

$$C_f := \sup_{\substack{x,s\in K\\ \gamma\in(0,1]}} \tfrac{2}{\gamma^2}\big(f(x+\gamma(s-x)) - f(x) - \gamma\langle\nabla f(x), s-x\rangle\big).$$

Lemma 11.1 (Quadratic upper model + norm bound). If $C_f < \infty$, then for all $x, s \in K$, $\gamma \in [0,1]$: $f(x+\gamma(s-x)) \le f(x) + \gamma\langle\nabla f(x), s-x\rangle + \frac{\gamma^2}{2}C_f$.
If f is L-smooth w.r.t. $\|\cdot\|$ and $D = \operatorname{diam}_{\|\cdot\|}(K)$: $C_f \le LD^2$ ($\le 4L\rho^2$ if $K = \rho B$).

Why C_f, not LD^2? C_f is intrinsic to (f, K) — norm-free; LD^2 is a lossy norm-dependent upper bound.

Lemma 11.1: Proof

Step 1. Quadratic upper model. For $\gamma \in (0,1]$, the definition of C_f gives

$$\frac{2}{\gamma^2}\big(f(x+\gamma(s-x)) - f(x) - \gamma\langle \nabla f(x), s-x\rangle\big) \le C_f,$$

which rearranges to the displayed inequality. For $\gamma = 0$ both sides equal $f(x)$.

Step 2. Norm-dependent upper bound. If f is L-smooth w.r.t. $\|\cdot\|$, then for $x, s \in K$ and $\gamma \in [0,1]$,

$$f(x+\gamma(s-x)) \le f(x) + \gamma\langle \nabla f(x), s-x\rangle + \tfrac{L}{2}\gamma^2\|s-x\|^2.$$

Since $\|s-x\| \le D$ for $x, s \in K$, taking the sup in the definition gives $C_f \le LD^2$. For $K = \rho B_{\|\cdot\|}$ we have $D = 2\rho$, hence $C_f \le 4L\rho^2$. $\square$

Frank–Wolfe Gap

Definition 11.3 (FW gap). For $x \in K$,

$$g_{\mathrm{FW}}(x) := \max_{s \in K} \langle \nabla f(x), x - s \rangle = -\min_{s \in K} \langle \nabla f(x), s - x \rangle.$$

Computable from one LMO call. Stationarity certificate: $g_{\mathrm{FW}}(x) = 0 \Leftrightarrow$ first-order optimal.

Theorem 11.2 (Gap upper-bounds suboptimality). If f is convex on K, then for all $x \in K$:

$$f(x) - \min_{y \in K} f(y) \leq g_{\mathrm{FW}}(x).$$

Proof. Convexity at $x^\star \in \operatorname{argmin}_K f$: $f(x^\star) \geq f(x) + \langle \nabla f(x), x^\star - x \rangle$, so $f(x) - f(x^\star) \leq \langle \nabla f(x), x - x^\star \rangle \leq \max_{s \in K} \langle \nabla f(x), x - s \rangle = g_{\mathrm{FW}}(x)$. $\square$

Deterministic Frank–Wolfe Rate

Theorem 11.3 (Frank–Wolfe, 1956; Jaggi, 2013). f convex, $C_f < \infty$, $x_0 \in K$. Run FW with $\gamma_t = \dfrac{2}{t+2}$. Then for all $T \geq 1$,

$$f(x_T) - \min_{x \in K} f(x) \leq \frac{2C_f}{T+2}.$$

Takeaway. $O(1/T)$ rate without projection or norm. With L-smoothness w.r.t. a norm: $\leq 2LD^2/(T+2)$.

Comparison.

- Projected GD (smooth convex): $O(LD^2/T)$ — same rate, but needs projection.
- For sparse / low-rank optimization: each FW iterate is a sparse / rank-1 atom $\Rightarrow$ structured solution.

Thm 11.3: Proof

Step 1. One-step recursion. Let $\delta_t := f(x_t) - f^\star$. Lemma 11.1 + LMO + Thm 11.2:

$$f(x_{t+1}) \le f(x_t) + \gamma_t \langle \nabla f(x_t), s_t - x_t \rangle + \tfrac{\gamma_t^2}{2} C_f = f(x_t) - \gamma_t g_{\mathrm{FW}}(x_t) + \tfrac{\gamma_t^2}{2} C_f.$$

Since $g_{\mathrm{FW}}(x_t) \ge \delta_t$, $\delta_{t+1} \le (1-\gamma_t)\delta_t + \frac{\gamma_t^2}{2} C_f$.

Step 2. Induction on $t \ge 1$ that $\delta_t \le 2C_f/(t+2)$. Base $t = 1$ ($\gamma_0 = 1$, $x_1 = s_0$, LMO + Lemma 11.1): $\delta_1 \le \frac{1}{2} C_f \le 2C_f/3$.

Step 3. Inductive step. With $\gamma_t = 2/(t+2)$:

$$\delta_{t+1} \le \left(1 - \tfrac{2}{t+2}\right) \tfrac{2C_f}{t+2} + \tfrac{1}{2}\left(\tfrac{2}{t+2}\right)^2 C_f = \tfrac{2C_f(t+1)}{(t+2)^2} \le \tfrac{2C_f}{t+3},$$

using $(t+1)(t+3) \le (t+2)^2$. $\square$

Constant-stepsize Frank–Wolfe

Theorem 11.4 (Constant-step FW). Same setup; $\gamma_t \equiv \gamma \in (0, 1]$. Then for all $T \geq 0$,

$$\delta_T \leq (1-\gamma)^T \delta_0 + \frac{\gamma C_f}{2}\big(1 - (1-\gamma)^T\big) \leq (1-\gamma)^T \delta_0 + \frac{\gamma C_f}{2}.$$

Two regimes.

- *Transient:* $(1-\gamma)^T \delta_0$ — geometric in T, but only if γ large.
- *Error:* $\gamma C_f/2$ — doesn't go to 0 for fixed $\gamma > 0$.

With $\gamma_T = \log T/T$: transient $\leq 1/T$ and error $\leq C_f \log T/(2T)$, so

$$f(x_T) - f^\star \leq \frac{f(x_0) - f^\star}{T} + \frac{C_f \log T}{2T}.$$

Thm 11.4: Proof

Step 1. Same one-step recursion as Thm 11.3. With $\gamma_t \equiv \gamma$,

$$\delta_{t+1} \le (1-\gamma)\delta_t + \tfrac{\gamma^2}{2} C_f.$$

Step 2. Unroll the affine recursion. Iterating,

$$\delta_T \le (1-\gamma)^T \delta_0 + \tfrac{\gamma^2}{2} C_f \sum_{k=0}^{T-1} (1-\gamma)^k = (1-\gamma)^T \delta_0 + \tfrac{\gamma C_f}{2}\big(1-(1-\gamma)^T\big).$$

The looser form $\le (1-\gamma)^T \delta_0 + \gamma C_f/2$ drops the $(1-(1-\gamma)^T)$ factor.

Step 3. Horizon-only choice. With $\gamma_T = \log T / T$ $(T \ge 2)$,

$$(1-\gamma_T)^T \le \exp(-\gamma_T T) = \tfrac{1}{T}.$$

Substituting into the transient-plus-error bound proves $f(x_T) - f^\star \le \frac{\delta_0}{T} + \frac{C_f \log T}{2T}$. $\square$

Normalized Steepest Descent with Weight Decay

Recall. Steepest descent w.r.t. norm $\|\cdot\|$ uses $v_t \in \mathrm{LMO}_B(\nabla f(x_t))$ where $B = \{v : \|v\| \le 1\}$.

Definition 11.4 (NSD with decoupled weight decay). Norm $\|\cdot\|$, weight decay $\lambda > 0$, learning rate $\eta_t \in [0, 1/\lambda]$. Choose $v_t \in \mathrm{LMO}_B(\nabla f(x_t))$ and update

$$\boxed{x_{t+1} = (1 - \lambda\eta_t)\, x_t + \eta_t\, v_t.}$$

Takeaway.

- $(1 - \lambda\eta_t)x_t$: shrink the iterate toward 0 (*weight decay*).
- $\eta_t v_t$: take a normalized steepest-descent step.
- Decoupled: λ shrinks x_t, not $\nabla f(x_t)$ (different from ℓ_2 regularization!).

NSD with Weight Decay = Frank–Wolfe on a Norm Ball

Proposition 11.6. Let $B = \{\|v\| \le 1\}$, $K = \frac{1}{\lambda}B$. Then NSD with decoupled weight decay is exactly Frank–Wolfe on K, with

$$\gamma_t = \lambda\eta_t, \qquad s_t = \tfrac{1}{\lambda}v_t \in K.$$

Conversely, FW on $K = \rho B$ with stepsize γ_t is NSD-WD with $\lambda = 1/\rho$, $\eta_t = \rho\gamma_t$.

Proof. $\mathrm{LMO}_K = \frac{1}{\lambda}\,\mathrm{LMO}_B$, so $s_t = v_t/\lambda$. The FW update with $\gamma_t = \lambda\eta_t$ is $(1-\gamma_t)x_t + \gamma_t s_t = (1-\lambda\eta_t)x_t + \eta_t v_t$. □

Consequence (Thm 11.4 with $C_f \le 4L/\lambda^2$):

$$f(x_T) - \min_{\|x\|\le 1/\lambda} f(x) \le (1-\lambda\eta)^T\delta_0 + \tfrac{2L\eta}{\lambda}.$$

With $\eta_T = \log T/(\lambda T)$: $\delta_0/T + 2L\log T/(\lambda^2 T)$.

Why Weight Decay Matters: Radius Comparison

NSD with weight decay (FW on $K = B/\lambda$):

$$f(x_T) - \min_{\|x\| \le 1/\lambda} f(x) \le \frac{2L/\lambda^2}{T+2}.$$

Comparator radius $= 1/\lambda$ (chosen by us).

NSD without weight decay (assume monotone iterates stay in initial sublevel set):

$$f(x_T) - f(x^\star) \le \frac{2LR_{\mathrm{sub}}^2}{T+2}, \qquad R_{\mathrm{sub}} := \sup\{\|x - x^\star\| : f(x) \le f(x_0)\}.$$

Both look like $O(LR^2/T)$. But the radius means different things:

- With WD: $R = 1/\lambda$ is a *chosen* feasible radius; comparator is the constrained min.
- Without WD: $R = R_{\mathrm{sub}}$ is the radius of the *initial sublevel set* around $x^\star$.

R_{sub} can be arbitrarily large, even when a useful centered ball exists.

Example: Sublevel Radius Can Be Huge

Example 11.1. Asymmetric quadratic on $\mathbb{R}$: for $0 < \varepsilon \leq 1$,

$$f_\varepsilon(x) = \begin{cases} \frac{1}{2}(x-1)^2, & x \leq 1, \\ \frac{\varepsilon}{2}(x-1)^2, & x \geq 1. \end{cases}$$

1-smooth; $x^\star = 1$.

Starting at $x_0 = 0$: $f_\varepsilon(0) = 1/2$. The sublevel set is

$$\{x : f_\varepsilon(x) \leq 1/2\} = [0,\, 1 + 1/\sqrt{\varepsilon}],$$

so $R_{\mathrm{sub}} = 1/\sqrt{\varepsilon}$ — arbitrarily large as $\varepsilon \to 0$.

But the centered domain $K = [-1, 1]$ already contains $x^\star = 1$ and has radius 1.

Takeaway. The two statements measure different geometric quantities; you cannot just compare the constants.

LMO Atoms on Common Norm Balls

Norm ball B	**LMO atom** $\mathrm{LMO}_B(g)$
ℓ_2	$-g/\|g\|_2$
ℓ_∞	$-\operatorname{sign}(g)$ entrywise
ℓ_1	$-\operatorname{sign}(g_{i^\star})e_{i^\star}$, where $i^\star \in \operatorname{argmax}_i \lvert g_i\rvert$
spectral norm ball	$-UV^\top$ from $G = U\Sigma V^\top$
nuclear norm ball	top rank-one atom $-uv^\top$
spectrahedron	$-vv^\top$, where v is a top eigenvector of G

Optimizer connection. Each row gives the LMO direction. Frank–Wolfe on that norm ball is equivalently a normalized update with decoupled weight decay.

ℓ_∞ ball + FW/LMO = sign update + WD (Lion-style). Spectral ball + FW/LMO = orthogonalized momentum update + WD (Muon-style).

Naive Stochastic Frank–Wolfe

Definition 11.5 (Naive single-sample stochastic FW). Filtration $\mathcal{F}_t$; x_t is $\mathcal{F}_t$-measurable; unbiased estimator $\widehat{g}_t$:

$$\mathbb{E}[\widehat{g}_t \mid \mathcal{F}_t] = \nabla f(x_t), \qquad s_t \in \mathrm{LMO}_K(\widehat{g}_t), \qquad x_{t+1} = (1-\gamma_t)x_t + \gamma_t s_t.$$

Question. Does the deterministic $O(C_f/T)$ rate carry over with σ-bounded noise?

Answer. No. Naive stochastic FW has a *non-vanishing error*, regardless of how small the learning rate is and how large the number of steps T is.

Stochastic FW: Non-vanishing Error

Theorem 11.7 (Non-vanishing error). f convex, $C_f < \infty$, $D = \operatorname{diam}_{\|\cdot\|}(K)$, $\mathbb{E}[\|\widehat{g}_t - \nabla f(x_t)\|_* \mid \mathcal{F}_t] \le \sigma$. Run naive stochastic FW with $\gamma_t = 2/(t+2)$. Then

$$\boxed{\mathbb{E}[f(x_T) - \min_K f] \le \frac{2C_f}{T+2} + D\sigma.}$$

The $D\sigma$ error does not vanish as $T \to \infty$.

Takeaway. The $2C_f/(T+2)$ piece is the deterministic rate; the additive $D\sigma$ piece is a non-vanishing error that grows linearly in σ and in the diameter D, and is independent of T.

Thm 11.7: Proof (1/2)

Let $x^\star \in \operatorname{argmin}_K f$, $\delta_t := f(x_t) - f(x^\star)$.

Step 1. Curvature inequality (Lem 11.1).

$$\delta_{t+1} \le \delta_t + \gamma_t \langle \nabla f(x_t), s_t - x_t \rangle + \tfrac{\gamma_t^2}{2} C_f.$$

Step 2. Insert/subtract $\widehat{g}_t$, use LMO on $\widehat{g}_t$. $\langle \widehat{g}_t, s_t - x_t \rangle \le \langle \widehat{g}_t, x^\star - x_t \rangle$ since $x^\star \in K$. Thus

$$\langle \nabla f(x_t), s_t - x_t \rangle = \langle \widehat{g}_t, s_t - x_t \rangle + \langle \nabla f(x_t) - \widehat{g}_t, s_t - x_t \rangle.$$

Add and subtract $\langle \widehat{g}_t, x^\star - x_t \rangle$:

$$\langle \nabla f(x_t), s_t - x_t \rangle \le \langle \nabla f(x_t), x^\star - x_t \rangle + \langle \widehat{g}_t - \nabla f(x_t), x^\star - s_t \rangle.$$

Step 3. Convexity. $\langle \nabla f(x_t), x^\star - x_t \rangle \le -\delta_t$. Combine with Step 1:

$$\delta_{t+1} \le (1 - \gamma_t)\delta_t + \tfrac{\gamma_t^2}{2} C_f + \gamma_t \langle \widehat{g}_t - \nabla f(x_t), x^\star - s_t \rangle.$$

Thm 11.7: Proof (2/2)

Step 4. Hölder + bounded diameter. $\|x^\star - s_t\| \le D$, so

$$|\langle \widehat{g}_t - \nabla f(x_t), x^\star - s_t\rangle| \le \|\widehat{g}_t - \nabla f(x_t)\|_* \cdot D.$$

Conditional expectation gives $\mathbb{E}[|\cdots| \mid \mathcal{F}_t] \le D\sigma$. (Note: *not* 0, because s_t depends on $\widehat{g}_t$.)

Step 5. Take expectation, recurse. $\mathbb{E}\delta_{t+1} \le (1-\gamma_t)\mathbb{E}\delta_t + \frac{\gamma_t^2}{2}C_f + \gamma_t D\sigma$.

Step 6. Induction with $\gamma_t = 2/(t+2)$. Same argument as Thm 11.3 with the additional constant $D\sigma$:

$$\mathbb{E}\delta_T \le \frac{2C_f}{T+2} + D\sigma.$$

Base case $\mathbb{E}\delta_1 \le \frac{1}{2}C_f + D\sigma \le 2C_f/3 + D\sigma$. $\square$

Constant-step Stochastic FW Error

Corollary 11.8 (Constant-step stochastic error). Same setup as Thm 11.7; $\gamma_t \equiv \gamma \in (0, 1]$. Then

$$\mathbb{E}\delta_T \le (1-\gamma)^T \delta_0 + \left(D\sigma + \tfrac{\gamma C_f}{2}\right)\left(1 - (1-\gamma)^T\right) \le (1-\gamma)^T \delta_0 + D\sigma + \tfrac{\gamma C_f}{2}.$$

Proof. From the Step-5 recursion of Thm 11.7 with $\gamma_t \equiv \gamma$:

$$\mathbb{E}\delta_{t+1} \le (1-\gamma)\mathbb{E}\delta_t + \gamma D\sigma + \tfrac{\gamma^2}{2} C_f.$$

Unrolling this affine recursion gives the displayed bound. □

Takeaway. Compared with deterministic Thm 11.4: the stochastic error adds $D\sigma$ on top of $\gamma C_f/2$. Even at $\gamma \to 0$, the $D\sigma$ persists.

A Non-vanishing Stochastic Error

The proof of Thm 11.7 gives the one-step recursion

$$\mathbb{E}[\delta_{t+1} \mid \mathcal{F}_t] \le (1-\gamma_t)\delta_t + \frac{\gamma_t^2}{2} C_f + \gamma_t D\sigma, \qquad \delta_t = f(x_t) - f(x^\star). \tag{FW}$$

The key scaling. In one step,

$$\underbrace{-\gamma_t \delta_t}_{\text{useful progress}} \quad + \quad \underbrace{\gamma_t D\sigma}_{\text{biased LMO noise}} \quad + \quad \underbrace{\frac{\gamma_t^2}{2} C_f}_{\text{curvature}} .$$

The curvature term is smaller order when $\gamma_t \to 0$; the biased noise term is not.

Let $S_T = \sum_{t<T} \gamma_t$. Normalizing the accumulated noise by the total movement gives

$$\frac{1}{S_T} \sum_{t<T} \gamma_t D\sigma = D\sigma.$$

So even if $\gamma_t \to 0$ and $S_T \to \infty$, the error created by feeding a single noisy gradient (covector) into the LMO does *not* average away.

MD: Constant LR Averages Gradients

Let h be the mirror map and set the dual coordinate $z_t = \nabla h(x_t)$. With a constant learning rate η, MD with $\widehat{g}_t = g_t + \xi_t$ is

$$z_{t+1} = z_t - \eta \widehat{g}_t, \qquad x_{t+1} = \nabla h^*(z_{t+1}).$$

Unroll and divide by the total movement ηT:

$$\frac{z_1 - z_{T+1}}{\eta T} = \frac{1}{T} \sum_{t<T} g_t + \frac{1}{T} \sum_{t<T} \xi_t.$$

Thus many small MD steps behave like averaging noisy gradients (covectors) before the mirror map sends the result back to primal space.

FW does the opposite. It sends the noisy gradient (covector) through a nonlinear atom selector before averaging:

$$s_t \in \mathrm{LMO}_K(g_t + \xi_t), \qquad \mathbb{E}[\mathrm{LMO}_K(g + \xi)] \neq \mathrm{LMO}_K(g) \text{ in general.}$$

Small steps average selected atoms, not the gradients (covectors) themselves.

Constant-LR MD: The Bound

Why does the averaging picture enter the proof? Since x_t and the comparator u are chosen before seeing the current noise,

$$\mathbb{E}[\langle \xi_t, x_t - u \rangle \mid \mathcal{F}_t] = 0.$$

The noise appears only through a predictable linear pairing.

With constant η and $\bar{x}_T = T^{-1} \sum_{t<T} x_t$, a typical MD bound is

$$\mathbb{E}[f(\bar{x}_T) - f(u)] \le \frac{D_h(u, x_1)}{\eta T} + \frac{\eta}{2\alpha}(G^2 + \sigma^2), \qquad \mathbb{E}\|\widehat{g}_t\|_*^2 \le G^2 + \sigma^2.$$

Small LR + many steps. For a fixed horizon, choose one constant $\eta \asymp T^{-1/2}$. Then both terms are $O(T^{-1/2})$. Naive stochastic FW has the extra $D\sigma$ term instead: the noise has already selected the atom before any averaging occurs.

Large-Batch FW: Denoise Before the LMO

Principle. FW cannot rely on many small steps to average gradients (covectors). Average the stochastic gradients *before* the LMO:

$$\bar{g}_t = \frac{1}{B}\sum_{b=1}^{B} g(x_t;\xi_{t,b}), \qquad s_t \in \mathrm{LMO}_K(\bar{g}_t), \qquad x_{t+1} = (1-\gamma_t)x_t + \gamma_t s_t.$$

If

$$\mathbb{E}[\|\bar{g}_t - \nabla f(x_t)\|_* \mid \mathcal{F}_t] \le \sigma_B,$$

then the proof of Thm 11.7 applies with σ replaced by σ_B. The non-vanishing residual term becomes $D\sigma_B$.

Caveat. In Euclidean norm, independent averaging often gives $\sigma_B \asymp B^{-1/2}$. For a general norm, this rate is an additional concentration assumption, not automatic.

Momentum FW: Cheap Denoising

Algorithmic template. Keep a gradient (covector) memory and query the LMO using that memory, with $m_{-1} = 0$:

$$m_t = \beta_t m_{t-1} + (1 - \beta_t)\widehat{g}_t, \qquad s_t \in \mathrm{LMO}_K(m_t),$$

$$x_{t+1} = (1 - \gamma_t)x_t + \gamma_t s_t.$$

What changes? Momentum can reduce variance cheaply, but it is biased: past gradients were evaluated at past points. Smoothness and slow movement are needed to control this bias.

Known result. In the Hilbert/Euclidean stochastic FW setting, the survey of Braun et al. summarizes an $O(T^{-1/3})$ momentum SFW rate under smoothness, compact diameter, and bounded second-moment noise.

Two-Parameter Momentum Optimizers

Practical sign / orthogonalized optimizers often use a Lion-style two-beta momentum template: one gradient (covector) chooses the current LMO direction, while another persistent memory is carried forward.

$$c_t = \beta_1 m_{t-1} + (1-\beta_1)\widehat{g}_t, \qquad m_t = \beta_2 m_{t-1} + (1-\beta_2)\widehat{g}_t,$$

$$v_t \in \mathrm{LMO}_B(c_t), \qquad x_{t+1} = (1-\lambda_t)x_t + \eta_t v_t.$$

- **Lion**: B is the ℓ_∞ ball, so $v_t = -\operatorname{sign}(c_t)$.
- **Muon**: B is the spectral norm ball, so $v_t = -UV^\top$ from $C_t = U\Sigma V^\top$.

This is the optimizer template. The lecture theorem explains why the gradient (covector) must be denoised before the LMO; it does not fully analyze this two-parameter momentum heuristic.

Summary

Frank–Wolfe replaces projection by linear minimization.

- Norm-free, no symmetry, no centered ball.
- Curvature constant C_f is the intrinsic quantity; LD^2 is a norm-dependent upper bound.
- Decreasing-stepsize $\gamma_t = 2/(t+2)$ gives $O(C_f/T)$.

Norm-ball specialization.

- FW on norm ball $\Leftrightarrow$ NSD with decoupled weight decay.
- Atom catalog: $\ell_\infty \to$ sign, spectral$\to$polar, $\ell_1 \to$basis, ...
- Modern optimizers (Lion, Muon) are momentum-FW on a chosen norm ball.

Stochastic warning. MD averages noisy gradients (covectors) in dual coordinates. Naive stochastic FW first selects an atom from the noisy gradient (covector), creating a non-vanishing $D\sigma$ error. Remedy: denoise before the LMO (large batch or momentum).

Next. Lower bounds for first-order methods.

Bibliographic Notes

- **Frank & Wolfe (Naval Res. Logistics Quart. 1956).** Quadratic programming.
- **Demyanov & Rubinov (1967).** Conditional-gradient method, smooth functional.
- **Jaggi (ICML 2013).** Modern projection-free / dual-gap viewpoint, $O(1/T)$ rate.
- **Hazan & Luo (ICML 2016); Mokhtari, Hassani & Karbasi (JMLR 2020).** Variance-reduced / momentum stochastic FW.
- **Bernstein & Newhouse (2024).** Norm-ball / polar duality view of modern optimizers.
- **Lion (Chen et al. 2023); Muon (Jordan et al. 2024; Liu et al. 2025).** Sign / spectral norm-ball directions in deep learning.
- **Type-2 / 2-smooth norms.** Ledoux–Talagrand (1991); Pinelis (1994); Juditsky–Nemirovski (2008).