# Lecture 11: Frank–Wolfe and Non-Euclidean Descent

Lecture 7 introduced steepest descent by minimizing a local quadratic upper model. Lecture 8 introduced mirror descent by replacing the quadratic penalty by a Bregman divergence. On a constrained set, both viewpoints usually ask us to solve a nonlinear local model over the feasible region:

$$x_{t+1} \in \arg\min_{x \in K} \left\{ \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta_t} \|x - x_t\|^2 \right\},$$

or

$$x_{t+1} \in \arg\min_{x \in K} \left\{ \eta_t \langle g_t, x - x_t \rangle + D_\Phi(x, x_t) \right\}.$$

These projection, proximal, and Bregman-projection steps are already major simplifications of minimizing the original objective $f$, and in many geometries they are cheap enough to be the right primitive. But "cheaper than solving the original problem" is not the same as "cheap per iteration." For some structured feasible sets, a nonlinear projection-like subproblem over $K$ is still expensive, while minimizing a linear function over $K$ is substantially easier.

Frank–Wolfe changes the constraint primitive. It keeps the first-order oracle for the objective, but replaces the projection/proximal oracle for the domain by a linear minimization oracle:

$$s_t \in \arg\min_{s \in K} \langle \nabla f(x_t), s \rangle .$$

General Frank–Wolfe does not require a norm and does not require the feasible set to be symmetric. It only requires a compact convex feasible set and the ability to minimize linear functions over it. The method goes back to Frank and Wolfe [FW56]; the modern projection-free viewpoint and the gap-certificate formulation used below are close to Jaggi [Jag13].

There is one important specialization that connects directly to the non-Euclidean descent language from earlier lectures. When $K$ is a centered norm ball, the linear minimization oracle returns a normalized steepest-descent direction, and the Frank–Wolfe convex-combination update becomes normalized steepest descent with an explicit decoupled weight-decay term.

## 11.1 Linear Minimization Over a Convex Set

We reuse the linear minimization oracle from Definition 7.2. In Lecture 7 the oracle was used mainly on a norm ball to describe steepest-descent directions. Here the same primitive is used on an arbitrary compact convex feasible set $K$. Compactness guarantees that an LMO output exists. Convexity is not needed for the oracle itself, but it is needed for the Frank–Wolfe update to remain feasible, because the update is a convex combination of the old iterate and the returned atom. Symmetry is not needed: the simplex, spectrahedra, and many polytopes are natural Frank–Wolfe domains but are not centered symmetric bodies.

**Definition 11.1** (Frank–Wolfe update)**.** Let $K \subset E$ be nonempty, compact, and convex. Let $U \subset E$ be open with $K \subset U$, and let $f : U \to \mathbb{R}$ be differentiable. Given $x_t \in K$, choose

$$s_t \in \mathrm{LMO}_K(\nabla f(x_t))$$

and a stepsize $\gamma_t \in [0, 1]$. The Frank–Wolfe update is

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

Because $K$ is convex, $x_{t+1} \in K$ whenever $x_t, s_t \in K$.

*Remark* 11.1 (Ambient differentiability convention). In this lecture, whenever a statement writes $\nabla f(x)$ for $x \in K$, the function $f$ is assumed to be differentiable on an open neighborhood $U \supset K$. Thus the Frank–Wolfe statements below do not rely on any boundary-gradient convention for differentiability only on $K$.

## 11.2 Curvature and Deterministic Frank–Wolfe

**Definition 11.2** (Frank–Wolfe curvature constant)**.** Let $K \subset E$ be nonempty, compact, and convex. Let $U \subset E$ be open with $K \subset U$, and let $f : U \to \mathbb{R}$ be differentiable. The curvature constant of $f$ over $K$ is

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

The curvature constant is the smallest constant that makes the quadratic upper model valid along every feasible chord. Unlike $L\rho^2$, it is intrinsic to the pair $(f, K)$ and does not require choosing a norm.

**Lemma 11.1** (Curvature inequality and norm-dependent upper bound)**.** *Let $K, U, f$ be as in Definition 11.2 and assume $C_f < +\infty$. Then for all $x, s \in K$ and all $\gamma \in [0, 1]$,*

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

*If, in addition, a norm $\|\cdot\|$ is chosen on $E$, $f$ is $L$-smooth with respect to this norm on $K$, and*

$$D := \mathrm{diam}_{\|\cdot\|}(K) := \sup_{x,y \in K} \|x - y\|,$$

*then*

$$C_f \le LD^2.$$

*In particular, if $K = \rho B_{\|\cdot\|}$ is the centered norm ball of radius $\rho$, then*

$$C_f \le 4L\rho^2.$$

*Proof of Lemma 11.1.* For $\gamma \in (0,1]$, the first inequality is exactly the definition of $C_f$, rearranged. For $\gamma = 0$, both sides equal $f(x)$.
Now assume $f$ is $L$-smooth with respect to $\|\cdot\|$ on $K$. For every $x, s \in K$ and $\gamma \in [0,1]$, the smoothness inequality gives

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

Since $\|s - x\| \le D$, taking the supremum in Definition 11.2 gives $C_f \le LD^2$. If $K = \rho B_{\|\cdot\|}$, then $D = 2\rho$. □

**Definition 11.3** (Frank–Wolfe gap)**.** Let $K \subset E$ be nonempty, compact, and convex. Let $U \subset E$ be open with $K \subset U$, and let $f : U \to \mathbb{R}$ be differentiable. The Frank–Wolfe gap at $x \in K$ is

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

The gap is a computable first-order stationarity certificate: it measures how much decrease the best feasible linearized move can promise from the current point.

**Theorem 11.2** (Frank–Wolfe gap upper bounds suboptimality)**.** *Let $K \subset E$ be nonempty, compact, and convex. Let $U \subset E$ be open with $K \subset U$, and let $f : U \to \mathbb{R}$ be differentiable and convex on $K$. Then for every $x \in K$,*

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

*Proof of Theorem 11.2.* Let $x^\star \in \arg\min_{y \in K} f(y)$, which exists because $K$ is compact and $f$ is continuous on $K$. By convexity and differentiability,

$$f(x^\star) \ge f(x) + \langle \nabla f(x), x^\star - x \rangle .$$

Rearranging gives

$$f(x) - f(x^\star) \le \langle \nabla f(x), x - x^\star \rangle .$$

Since $x^\star \in K$,

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

□

**Theorem 11.3** (Deterministic Frank–Wolfe rate)**.** *Let $K \subset E$ be nonempty, compact, and convex. Let $U \subset E$ be open with $K \subset U$, and let $f : U \to \mathbb{R}$ be differentiable and convex on $K$. Assume $C_f < +\infty$. Start from $x_0 \in K$ and run Frank–Wolfe with*

$$\gamma_t := \frac{2}{t+2}, \qquad t = 0, 1, 2, \ldots .$$

*Then for every $T \ge 1$,*

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

*Proof of Theorem 11.3.* Let $x^\star \in \arg\min_{x\in K} f(x)$ and define

$$\delta_t := f(x_t) - f(x^\star).$$

By Lemma 11.1,

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

Because $s_t \in \mathrm{LMO}_K(\nabla f(x_t))$,

$$\langle \nabla f(x_t), s_t - x_t\rangle = -g_{\mathrm{FW}}(x_t).$$

Thus

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

By Theorem 11.2, $g_{\mathrm{FW}}(x_t) \ge \delta_t$. Hence

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

We prove by induction that

$$\delta_t \le \frac{2C_f}{t+2} \qquad \forall t \ge 1.$$

For $t = 1$, $\gamma_0 = 1$, so $x_1 = s_0$. Using the curvature inequality with $x = x_0$, $s = x^\star$, and $\gamma = 1$, and using the LMO property of $s_0$, gives

$$\begin{aligned} f(x_1) &\le f(x_0) + \langle \nabla f(x_0), s_0 - x_0\rangle + \frac{1}{2} C_f \\ &\le f(x_0) + \langle \nabla f(x_0), x^\star - x_0\rangle + \frac{1}{2} C_f \\ &\le f(x^\star) + \frac{1}{2} C_f. \end{aligned}$$

Thus $\delta_1 \le C_f/2 \le 2C_f/3$.
Assume $t \ge 1$ and $\delta_t \le 2C_f/(t+2)$. Since $\gamma_t = 2/(t+2)$,

$$\begin{aligned} \delta_{t+1} &\le \left(1 - \frac{2}{t+2}\right)\frac{2C_f}{t+2} + \frac{1}{2}\left(\frac{2}{t+2}\right)^2 C_f \\ &= \frac{2C_f(t+1)}{(t+2)^2} \le \frac{2C_f}{t+3}, \end{aligned}$$

where the last inequality is $(t+1)(t+3) \le (t+2)^2$. □

**Theorem 11.4** (Constant-stepsize Frank–Wolfe)**.** *Under the assumptions of Theorem 11.3, run Frank–Wolfe with a constant stepsize*

$$\gamma_t \equiv \gamma \in (0, 1].$$

*Let*

$$\delta_t := f(x_t) - f^\star, \qquad f^\star := \min_{x\in K} f(x).$$

*Then for every $T \geq 0$,*

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

*In particular,*

$$f(x_T) - f^\star \leq (1-\gamma)^T (f(x_0) - f^\star) + \frac{\gamma C_f}{2}.$$

*For a horizon-only constant stepsize, take*

$$\gamma_T := \frac{\log T}{T} \qquad (T \geq 2).$$

*Then*

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

*Proof of Theorem 11.4.* The proof begins with the same one-step recursion as in Theorem 11.3:

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

Unrolling this affine recursion gives

$$\begin{aligned}\delta_T &\leq (1-\gamma)^T \delta_0 + \frac{\gamma^2 C_f}{2} \sum_{k=0}^{T-1} (1-\gamma)^k \\ &= (1-\gamma)^T \delta_0 + \frac{\gamma C_f}{2}\left(1 - (1-\gamma)^T\right).\end{aligned}$$

For the horizon-only choice $\gamma_T = \log T / T$, use

$$(1-\gamma_T)^T \leq \exp(-\gamma_T T) = \frac{1}{T}.$$

Substituting $\gamma_T$ into the transient-plus-error bound proves the displayed estimate. □

**Corollary 11.5** (Norm-smooth Frank–Wolfe rates)**.** *Under the assumptions of Theorem 11.3, choose any norm $\|\cdot\|$ on $E$. If $f$ is $L$-smooth with respect to $\|\cdot\|$ on $K$ and $D = \operatorname{diam}_{\|\cdot\|}(K)$, then decreasing-stepsize Frank–Wolfe satisfies*

$$f(x_T) - \min_{x \in K} f(x) \leq \frac{2LD^2}{T+2} \qquad (T \geq 1).$$

*Constant-stepsize Frank–Wolfe satisfies*

$$f(x_T) - f^\star \leq (1-\gamma)^T (f(x_0) - f^\star) + \frac{\gamma L D^2}{2}.$$

*Proof of Corollary 11.5.* By Lemma 11.1, $C_f \leq LD^2$. Substitute this into Theorems 11.3 and 11.4. □

### 11.3 Normalized Steepest Descent with Weight Decay = Frank–Wolfe

The general Frank–Wolfe theorem does not need a norm. A norm appears only after specializing $K$ to a centered symmetric body. Recall from Proposition 7.2 and Definition 7.3 that the LMO over the unit ball is the normalized steepest-descent direction associated with that norm.

**Definition 11.4** (Normalized steepest descent with decoupled weight decay)**.** Let $\|\cdot\|$ be a norm on $E$, and let

$$B := \{v \in E : \|v\| \le 1\}.$$

Fix a weight-decay parameter $\lambda > 0$. Given $x_t \in E$, choose

$$v_t \in \mathrm{LMO}_B(\nabla f(x_t))$$

and set

$$x_{t+1} = (1 - \lambda\eta_t)x_t + \eta_t v_t, \qquad \eta_t \in [0, 1/\lambda].$$

**Proposition 11.6** (NSD with weight decay is Frank–Wolfe on a norm ball)**.** *Let $\|\cdot\|$ be a norm on $E$, let*

$$B := \{v \in E : \|v\| \le 1\}, \qquad K := \frac{1}{\lambda}B$$

*where $\lambda > 0$. The normalized steepest-descent update with decoupled weight decay in Definition 11.4 is exactly Frank–Wolfe on $K$, with Frank–Wolfe stepsize*

$$\gamma_t = \lambda\eta_t$$

*and atom*

$$s_t = \frac{1}{\lambda}v_t \in K.$$

*Conversely, Frank–Wolfe on $K = \rho B$ with stepsize $\gamma_t$ can be written as normalized steepest descent with weight decay by taking $\lambda = 1/\rho$ and $\eta_t = \rho\gamma_t$.*

*Proof of Proposition 11.6.* Because $K = (1/\lambda)B$,

$$\mathrm{LMO}_K(\nabla f(x_t)) = \frac{1}{\lambda}\mathrm{LMO}_B(\nabla f(x_t)).$$

Thus any Frank–Wolfe atom on $K$ can be written as $s_t = (1/\lambda)v_t$ with $v_t \in \mathrm{LMO}_B(\nabla f(x_t))$. Substituting this into the Frank–Wolfe update with $\gamma_t = \lambda\eta_t$ gives

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

The converse is the same identity with $\lambda = 1/\rho$ and $\eta_t = \rho\gamma_t$. □

The constant-stepsize result from Theorem 11.4 now has a direct optimizer reading. Under the norm-smooth estimate $C_f \le 4L/\lambda^2$, the constant-learning-rate update in Definition 11.4 satisfies

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

where

$$\delta_0 := f(x_0) - \min_{\|x\| \le 1/\lambda} f(x).$$

With the horizon-only choice

$$\eta_T = \frac{\log T}{\lambda T},$$

this gives

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

Without weight decay, normalized steepest descent uses

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

This is a different optimization statement. If $f : E \to \mathbb{R}$ is convex and $L$-smooth with respect to $\|\cdot\|$, if $f$ attains its unconstrained minimum at $x^\star$, and if the initial sublevel radius

$$R_{\mathrm{sub}} := \sup \{\|x - x^\star\| : f(x) \le f(x_0)\}$$

is finite, then this is the natural radius scale for the no-weight-decay statement. More precisely, for any run whose iterates remain in the initial sublevel set, for example a monotone version obtained by line search, the usual NSD calculation with $\eta_t = 2R_{\mathrm{sub}}/(t+2)$ gives

$$f(x_T) - f(x^\star) \le \frac{2LR_{\mathrm{sub}}^2}{T+2}.$$

Thus the two rates have the same $1/T$ flavor, but the radius means something different. With weight decay, the radius is the chosen feasible-domain radius $1/\lambda$, and the comparator is the constrained optimum. Without weight decay, the radius is the radius of the relevant sublevel set around an unconstrained minimizer; this radius can be much larger than the norm ball on which one would run Frank–Wolfe.

**Example 11.1** (Initial sublevel radius can be much larger than a centered domain radius)**.** The radius $R_{\mathrm{sub}}$ can be much worse than the radius of a useful centered constrained domain. Let $0 < \varepsilon \le 1$ and define the one-dimensional convex differentiable function

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

It is 1-smooth, and its minimizer is $x^\star = 1$. Starting from $x_0 = 0$, $f_\varepsilon(0) = 1/2$, and therefore

$$\{x : f_\varepsilon(x) \le f_\varepsilon(0)\} = \left[0, 1 + \frac{1}{\sqrt{\varepsilon}}\right].$$

Thus

$$R_{\mathrm{sub}} = \sup \{|x-1| : f_\varepsilon(x) \le f_\varepsilon(0)\} = \frac{1}{\sqrt{\varepsilon}},$$

which can be arbitrarily large. However, the centered domain $K = [-1, 1]$ already contains the minimizer and has radius 1. The point is not that the algorithms must behave this differently on this toy problem; the point is that the two theorems depend on different domains.

The same norm-ball viewpoint also recovers several optimizer directions that appear in modern machine-learning practice. The point here is not to catalog every case, but to note that after adding decoupled weight decay, these normalized directions can be read as Frank–Wolfe updates on the corresponding norm balls. For example, on the $\ell_\infty$ ball,

$$\mathrm{LMO}_{B_\infty}(g) \ni -\mathrm{sign}(g),$$

so the corresponding normalized direction is the sign direction. On the spectral-norm ball, if $G = U\Sigma V^\top$ is a singular-value decomposition, then

$$\arg\min_{\|S\|_{\mathrm{op}} \le 1} \langle G, S \rangle$$

contains $-UV^\top$ when the full polar factor is well defined. This is the matrix analogue of the sign direction: the support function of the spectral-norm ball is the nuclear norm, and the LMO selects a negative subgradient of that dual norm.

### 11.4 Naive Stochastic Frank–Wolfe

The deterministic proof is stable because the atom is chosen using the true gradient. If the atom is chosen using a fresh noisy gradient, the same proof acquires a noise term that is not a martingale difference. The issue is not stochastic gradients by themselves; the issue is applying a nonlinear LMO before the noise has been averaged or otherwise controlled.

**Definition 11.5** (Naive single-sample stochastic Frank–Wolfe)**.** Let $(\mathcal{F}_t)_{t\ge 0}$ be a filtration and suppose $x_t$ is $\mathcal{F}_t$-measurable. A naive stochastic Frank–Wolfe step draws a stochastic gradient estimator $\widehat{g}_t \in E^*$ satisfying

$$\mathbb{E}[\widehat{g}_t \mid \mathcal{F}_t] = \nabla f(x_t),$$

chooses

$$s_t \in \mathrm{LMO}_K(\widehat{g}_t),$$

and updates

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

**Theorem 11.7** (Naive stochastic Frank–Wolfe has a non-vanishing error)**.** *Let $K \subset E$ be nonempty, compact, and convex. Let $U \subset E$ be open with $K \subset U$, and let $f : U \to \mathbb{R}$ be differentiable and convex on $K$. Assume $C_f < +\infty$. Choose a norm $\|\cdot\|$ on $E$, let $\|\cdot\|_*$ be its dual norm, and set*

$$D := \mathrm{diam}_{\|\cdot\|}(K).$$

*Assume the stochastic gradient estimator satisfies*

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

*for all $t \ge 0$. Run naive stochastic Frank–Wolfe with*

$$\gamma_t = \frac{2}{t+2}.$$

*Then for every $T \ge 1$,*

$$\mathbb{E}\left[f(x_T) - \min_{x\in K} f(x)\right] \le \frac{2C_f}{T+2} + D\sigma.$$

*Proof of Theorem 11.7.* Let $x^\star \in \arg\min_{x\in K} f(x)$ and define $\delta_t := f(x_t) - f(x^\star)$. By Lemma 11.1,

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

Insert and subtract $\widehat{g}_t$:

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

Since $s_t \in \mathrm{LMO}_K(\widehat{g}_t)$ and $x^\star \in K$,

$$\langle \widehat{g}_t, s_t - x_t \rangle \le \langle \widehat{g}_t, x^\star - x_t \rangle .$$

Therefore

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

By convexity,

$$\langle \nabla f(x_t), x^\star - x_t \rangle \le -\delta_t.$$

Thus

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

Using Holder's inequality and $\|x^\star - s_t\| \le D$, taking conditional expectations gives

$$\mathbb{E}[\delta_{t+1} \mid \mathcal{F}_t] \le (1-\gamma_t)\delta_t + \frac{\gamma_t^2}{2} C_f + \gamma_t D\sigma.$$

Taking full expectation,

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

With $\gamma_t = 2/(t+2)$, the same induction as in Theorem 11.3, with the additional constant $D\sigma$, yields

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

For the base case, the displayed recursion with $\gamma_0 = 1$ gives

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

□

**Corollary 11.8** (Constant-step stochastic Frank–Wolfe error)**.** *Under the assumptions of Theorem 11.7, if $\gamma_t \equiv \gamma \in (0,1]$, then*

$$\mathbb{E}\left[f(x_T) - f^\star\right] \le (1-\gamma)^T \left(f(x_0) - f^\star\right) + \left(D\sigma + \frac{\gamma C_f}{2}\right)\left(1 - (1-\gamma)^T\right).$$

*In particular,*

$$\mathbb{E}\left[f(x_T) - f^\star\right] \le (1-\gamma)^T \left(f(x_0) - f^\star\right) + D\sigma + \frac{\gamma C_f}{2}.$$

*Proof of Corollary 11.8.* From the proof of Theorem 11.7, with $\gamma_t \equiv \gamma$,

$$\mathbb{E}[\delta_{t+1}] \leq (1-\gamma)\mathbb{E}[\delta_t] + \gamma D\sigma + \frac{\gamma^2}{2} C_f.$$

Unrolling this recursion gives the claim. □

Stochastic mirror descent behaves differently. In the rates from Theorem 9.8, the learning-rate schedule controls the stochastic error: as the horizon $T$ grows, one can choose a smaller learning rate, or a diminishing learning-rate schedule, so the noise contribution in the rate goes to zero. Naive stochastic Frank–Wolfe fails because the noisy covector is fed into the LMO before the update direction is chosen:

$$s_t \in \mathrm{LMO}_K(\nabla f(x_t) + \xi_t).$$

The resulting atom $s_t$ depends nonlinearly on the same noise $\xi_t$, and the proof leaves a term of the form

$$\langle \xi_t, x^\star - s_t \rangle .$$

This term need not vanish merely because the noise is unbiased or concentrated in an averaged sense. This remains true regardless of how small the learning rate is and how large the number of steps $T$ is: shrinking the steps also shrinks the useful progress. The issue is the order of operations: Frank–Wolfe should denoise the covector before passing it through the LMO.

*Remark* 11.2 (Denoising before the LMO). The theorem above suggests a simple algorithmic principle: average or smooth the stochastic covectors before applying the LMO. We record two concrete algorithm boxes without proving general norm-space convergence guarantees for the momentum variant in this lecture.

**Example 11.2** (Large-batch stochastic Frank–Wolfe)**.** This variant replaces the current noisy covector by a mini-batch average before querying the LMO.

**Large-batch stochastic Frank–Wolfe**

**Require:** A compact convex set $K$, an initial point $x_0 \in K$, a batch size $B$, and stepsizes $\gamma_t \in [0,1]$.

1: **for** $t = 0, 1, 2, \ldots$ **do**
2: Draw a batch $\xi_{t,1}, \ldots, \xi_{t,B}$.
3: Form $\bar{g}_t \leftarrow B^{-1} \sum_{b=1}^{B} g(x_t; \xi_{t,b})$.
4: Choose $s_t \in \mathrm{LMO}_K(\bar{g}_t)$.
5: Set $x_{t+1} \leftarrow (1-\gamma_t)x_t + \gamma_t s_t$.
6: **end for**

If one can prove a bound

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

then Theorem 11.7 applies with $\sigma$ replaced by the effective noise level. In Euclidean norm this kind of $B^{-1/2}$ scaling follows from second-moment additivity. For norms beyond the Euclidean norm, however, a rate such as $\sigma_B \asymp B^{-1/2}$ is not automatic; it requires a norm-dependent concentration inequality, such as one coming from Rademacher type 2, martingale type 2, or a 2-smooth norm assumption, or it must be assumed directly.

**Example 11.3** (Momentum stochastic Frank–Wolfe / momentum NSD)**.** This variant keeps a covector memory and queries the LMO using that smoothed covector.

**Momentum stochastic Frank–Wolfe**

**Require:** A compact convex set $K$, an initial point $x_0 \in K$, $m_{-1} = 0$, averaging weights $\beta_t$, and stepsizes $\gamma_t \in [0, 1]$.

1: **for** $t = 0, 1, 2, \dots$ **do**
2: Draw a stochastic gradient $\widehat{g}_t$.
3: Set $m_t \leftarrow \beta_t m_{t-1} + (1 - \beta_t)\widehat{g}_t$.
4: Choose $s_t \in \mathrm{LMO}_K(m_t)$.
5: Set $x_{t+1} \leftarrow (1 - \gamma_t)x_t + \gamma_t s_t$.
6: **end for**

Large batch is unbiased at the current point. Momentum is cheaper, but biased, because it averages gradients from previous points. Smoothness and slow movement are needed to control that bias.

For the Hilbert-space stochastic optimization setting, Algorithm 4.5 and Theorem 4.12 in the conditional-gradient survey of Braun et al. [BCC+22] summarize the momentum stochastic Frank–Wolfe result of Mokhtari, Hassani, and Karbasi [MHK20]: under $L$-smoothness, compact diameter, and bounded second-moment noise, a suitable choice of $\gamma_t$ and momentum weight gives an $O(T^{-1/3})$ expected primal-gap rate. We do not state that result as a theorem here because its proof is a Hilbert-space gradient-tracking argument; extending it cleanly to the general normed-space LMO viewpoint of this lecture would require additional geometry-dependent assumptions.

*Remark* 11.3 (Nesterov-style two-parameter momentum NSD)*.* The LMO viewpoint gives a clean way to write sign-style and orthogonalized momentum updates. Maintain a covector state $m_t$, form a current update covector $c_t$, query the norm-ball LMO at $c_t$, and then apply the resulting normalized direction with decoupled weight decay:

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

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

The special case $\beta_1 = \beta_2$ is the single-filter momentum-NSD update produced directly by Example 11.3. Popular optimizers often use two different coefficients: $\beta_1$ controls the covector used for the current normalized direction, while $\beta_2$ controls the persistent memory state. From the present convex-optimization viewpoint, why this extra degree of freedom helps, and how it should interact with stochastic noise, curvature, and LMO inexactness, remains an open problem.

## Bibliographic Notes

Frank–Wolfe was introduced by Frank and Wolfe [FW56]. The name conditional-gradient method and the smooth convex-functional formulation are also associated with Demyanov and Rubinov [DR67]. The curvature constant, projection-free viewpoint, and dual-gap certificate used in this lecture follow the modern presentation of Jaggi [Jag13]. The mirror-descent / conditional-gradient duality perspective is developed by Bach [Bac15] and the generalized conditional subgradient

framework of Peña [Peñ19]. For stochastic Frank–Wolfe methods that explicitly reduce gradient noise before applying an LMO, see Hazan and Luo [HL16], Mokhtari, Hassani, and Karbasi [MHK20], and the survey of Braun et al. [BCC+22]. For recent optimizer connections to norm-ball directions, see Bernstein and Newhouse [BN24], the Lion paper [CLH+23], and the Muon notes and scaling report [JJB+24, LSY+25]. For background on Rademacher type, cotype, and probability in Banach spaces, see Ledoux and Talagrand [LT91]; for martingale inequalities in 2-smooth spaces, see Pinelis [Pin94] and the optimization-oriented treatment of Juditsky and Nemirovski [JN08].

### Dependency and Proof Sketch

1. Definition 7.2 supplies the LMO primitive, and Definition 11.1 turns it into the projection-free Frank–Wolfe update. This is the conceptual change from Lectures 7–8.
2. Lemma 11.1 explains why $C_f$ is the intrinsic norm-free curvature quantity for general compact convex $K$, and why $LD^2$ is a norm-dependent upper bound once a norm is chosen.
3. Theorem 11.2 is the direct convexity certificate for Frank–Wolfe.
4. Theorems 11.3 and 11.4 and Corollary 11.5 are all consequences of the same one-step recursion: the LMO direction gives the gap decrease, and curvature pays the quadratic error.
5. Proposition 11.6 is the bridge from projection-free optimization to normalized steepest descent on a centered norm ball. This is where decoupled weight decay appears automatically.
6. The no-weight-decay comparison and Example 11.1 separate the initial-sublevel radius statement from the FW/weight-decay statement. Both have an $O(LR^2/T)$ flavor, but $R$ refers to different sets.
7. Theorem 11.7 and Corollary 11.8 give the stochastic warning. The extra term comes from pairing the current noise with the atom selected using that same noise; it does not telescope.

## References


[Bac15] Francis Bach. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1):115–129, 2015.

[BCC+22] Gábor Braun, Alejandro Carderera, Cyrille W. Combettes, Hamed Hassani, Amin Karbasi, Aryan Mokhtari, and Sebastian Pokutta. Conditional gradient methods, November 2022.

[BN24] Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

[CLH+23] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms. In *Advances in Neural Information Processing Systems*, volume 36, pages 49205–49233, 2023.

[DR67] V. F. Demyanov and A. M. Rubinov. The minimization of a smooth convex functional on a convex set. *SIAM Journal on Control*, 5(2):280–294, 1967.

[FW56] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1–2):95–110, 1956.

[HL16] Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1263–1271. PMLR, 2016.

[Jag13] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 427–435. PMLR, 2013.

[JJB+24] Keller Jordan, Yuchen Jin, Vladislav Boza, Jiacheng You, Franz Cecista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. https://kellerjordan.github.io/posts/muon/, 2024. Blog post.

[JN08] Anatoli Juditsky and Arkadi Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.

[LSY+25] Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Enzhe Xu, Zhiyuan Li, Aonan Liu, Peng Zheng, Yulun Zhang, Tianyu Xie, Zhi Li, and Hongxia Zhou. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025.

[LT91] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, Berlin, Heidelberg, 1991.

[MHK20] Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *Journal of Machine Learning Research*, 21(105):1–49, 2020.

[Peñ19] Javier F. Peña. Generalized conditional subgradient and generalized mirror descent: Duality, convergence, and symmetry. *arXiv preprint arXiv:1903.00459*, 2019.

[Pin94] Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994.