Lecture 12: Oracle Complexity Lower Bounds

Up to this point, most of the course has proved upper bounds. Once an algorithm is fixed — projected gradient, mirror descent, adaptive mirror descent, or Frank–Wolfe — we prove that its error is at most a certain rate. Those theorems answer: what can this algorithm guarantee? Lecture 12 asks the complementary question: are these rates improvable by a different first-order method, or are they imposed by the information available from the oracle?

The general form of an oracle-complexity lower bound is a statement about a *function class*. Fix a class $\mathcal{F}$ of objectives, an oracle model (e.g. subgradient or first-order), and a feasible region. A lower bound says: for every algorithm A, there exists $f \in \mathcal{F}$ on which A makes error at least $\Delta(T)$ after T oracle calls,

$$\inf_{A} \sup_{f \in \mathcal{F}} \big[f(x_{T+1}) - f^\star \big] \geq \Delta(T).$$

The class is essential: without one, the statement is vacuous, since for any single f the algorithm "output $\arg\min f$" has zero error. Lower bounds are inherently statements about worst-case complexity over a family.

Conceptually, a lower bound is a *certificate* one level above the certificates of optimality we have already used in this course. Convex duality gives a dual-feasible point that certifies primal near-optimality; KKT multipliers certify primal optimality; the Frank–Wolfe gap certifies suboptimality of an iterate. A matching pair $\Omega(\Delta(T))$ and $O(\Delta(T))$ is then weak/strong duality at the algorithm level: a tight characterization of the class. Without a lower bound, an $O(1/T)$ rate could be either optimal or off by a factor of T, and we cannot tell which.

Definition 12.1 (Subgradient oracle)**.** Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. A subgradient oracle for f is a rule that, on input $x \in \mathbb{R}^d$, returns the pair $(f(x), g)$ where $g \in \partial f(x)$.

Definition 12.2 (Deterministic first-order method)**.** A deterministic first-order method is specified by a single rule A which maps every finite list of previous oracle replies to a query point in $\mathbb{R}^d$. Define the first query by

$$x_1 := A(\varnothing).$$

After replies

$$\big((f(x_1), g_1), \ldots, (f(x_t), g_t)\big),$$

the next query is

$$x_{t+1} := A\big((f(x_1), g_1), \ldots, (f(x_t), g_t)\big).$$

Thus a run of length T makes exactly T oracle calls and outputs x_{T+1}.

Definition 12.3 (Zero-initialized linear-span method)**.** A deterministic first-order method is called a *zero-initialized linear-span method* if

$$x_1 = A(\varnothing) = 0$$

and, after t oracle replies, the next iterate satisfies

$$x_{t+1} \in \operatorname{span}\{g_1, \dots, g_t\}, \qquad t = 1, 2, \dots,$$

where g_s denotes the subgradient returned by the oracle at round $s \in \{1, \dots, t\}$. This is the zero-initialized version of the linear span assumption from the oracle lower-bound literature; see [Nes04, Assumption 2.1.4].

The linear-span assumption is especially natural under ℓ_2 constraints. Suppose the feasible region is a Euclidean ball (or any set symmetric in the ℓ_2 sense) and a candidate iterate x_{t+1} had a nonzero component orthogonal to $\operatorname{span}\{g_1, \dots, g_t\}$. That orthogonal component is invisible to the oracle so far: it has not been certified by any returned subgradient and the oracle could not have "intended" for the algorithm to move there. Yet under $\|\cdot\|_2$ it strictly increases $\|x_{t+1}\|_2$, so projecting it away improves the iterate without changing the oracle picture. In ℓ_2 geometry, the linear-span restriction is therefore not a real restriction — it is the geometrically efficient choice that any sensible algorithm makes anyway. Under non-ℓ_2 norms this argument fails, which is why the lifting reduction below has to do real work to remove the assumption in general.

Definition 12.4 (Zero-initialized span-respecting transcript for a fixed oracle)**.** Let $h : \mathbb{R}^m \to \mathbb{R}$ be convex, and let O_h be a deterministic subgradient oracle for h:

$$\mathsf{O}_h(y) = (h(y), g(y)), \qquad g(y) \in \partial h(y).$$

A length-T transcript

$$(y_1, g_1), \dots, (y_T, g_T), y_{T+1}$$

is called a *zero-initialized span-respecting transcript* for (h, O_h) if $y_1 = 0$,

$$\mathsf{O}_h(y_t) = (h(y_t), g_t) \qquad \text{for } t = 1, \dots, T,$$

and

$$y_{t+1} \in \operatorname{span}\{g_1, \dots, g_t\} \qquad \text{for } t = 1, \dots, T.$$

The span of the empty set is interpreted as $\{0\}$.

Definition 12.5 (G-Lipschitz convex function)**.** Let $G > 0$. A convex function $f : \mathbb{R}^d \to \mathbb{R}$ is called G-Lipschitz with respect to $\|\cdot\|_2$ if

$$\forall x, y \in \mathbb{R}^d, \qquad |f(x) - f(y)| \le G\|x - y\|_2.$$

Lemma 12.1 (The max-coordinate function is Lipschitz)**.** *Let $d \in \mathbb{N}$ and $G > 0$. Define $f : \mathbb{R}^d \to \mathbb{R}$ by*

$$\forall x \in \mathbb{R}^d, \qquad f(x) := G \max_{1 \le i \le d} x_i.$$

Then f is convex and G-Lipschitz with respect to $\|\cdot\|_2$. Moreover, for every $x \in \mathbb{R}^d$ and every index $j \in \arg\max_{1\le i\le d} x_i$, the vector Ge_j belongs to $\partial f(x)$.

Proof of Lemma 12.1. The function f is the pointwise maximum of the affine functions $x \mapsto Gx_i$, so it is convex. Fix $x \in \mathbb{R}^d$ and let $j \in \arg\max_{1\le i\le d} x_i$. Then, for every $y \in \mathbb{R}^d$,

$$f(y) = G \max_{1\le i\le d} y_i \ge Gy_j = Gx_j + G(y_j - x_j) = f(x) + \langle Ge_j, y - x\rangle .$$

Hence $Ge_j \in \partial f(x)$.
To prove Lipschitz continuity, fix $x, y \in \mathbb{R}^d$. Choose $i^\star \in \arg\max_i x_i$. Then

$$f(x) - f(y) = Gx_{i^\star} - G \max_i y_i \le G(x_{i^\star} - y_{i^\star}) \le G \max_{1\le i\le d} |x_i - y_i| .$$

Interchanging the roles of x and y gives

$$|f(x) - f(y)| \le G \max_{1\le i\le d} |x_i - y_i| \le G \|x - y\|_2 .$$

Therefore f is G-Lipschitz with respect to $\|\cdot\|_2$. □

At the technical level, the proofs in this lecture use a single recurring device: *information hiding*. A first-order oracle reveals only local first-order information at the queried point. A hard instance is designed so that each query reveals only a small part of the problem. After T queries, some coordinate, or some part of a chain, remains unseen, and the algorithm cannot distinguish the true optimum well enough to be accurate. This is the same idea in both the nonsmooth and the smooth lower bound below.

The first lower bound is stated at the transcript level. Its quantifier order is

$$\exists (h, \mathsf{O}_h) \quad \forall \tau, \qquad \text{suboptimality of the last point of } \tau \ge \Delta.$$

Here τ ranges over zero-initialized span-respecting transcripts as in Definition 12.4. Since every zero-initialized linear-span method produces such a transcript, this also implies

$$\exists (h, \mathsf{O}_h) \quad \forall A_{\mathrm{zls}}, \qquad \text{suboptimality after } T \text{ oracle calls} \ge \Delta,$$

where A_{zls} ranges over zero-initialized linear-span methods.

Theorem 12.2 (Fixed max-coordinate lower bound for zero-initialized span-respecting transcripts)**.** *Let $T \in \mathbb{N}$, let $G > 0$, and let $R > 0$. Set $m := T + 1$, define*

$$h(y) := G \max_{1\le i\le m} y_i, \qquad y \in \mathbb{R}^m,$$

and equip h with the deterministic oracle

$$\mathsf{O}_h(y) := \big(h(y), Ge_{j(y)}\big), \qquad j(y) := \min \arg\max_{1\le i\le m} y_i.$$

Then h is convex and G-Lipschitz with respect to $\|\cdot\|_2$, and every length-T zero-initialized span-respecting transcript for (h, O_h) satisfies

$$h(y_{T+1}) - \min_{\|y\|_2\le R} h(y) \ge \frac{GR}{\sqrt{T+1}}.$$

Consequently, any zero-initialized linear-span method requires at least order G^2R^2/ε^2 oracle calls to guarantee error at most ε on the class of Euclidean G-Lipschitz convex objectives.

Proof of Theorem 12.2. By Lemma 12.1, h is convex and G-Lipschitz, and O_h is a valid subgradient oracle.
Fix any length-T zero-initialized span-respecting transcript for (h, O_h). Since each returned subgradient is a multiple of one standard basis vector, writing $j_t := j(y_t)$ $(= \min \arg\max_{1\le i\le m}(y_t)_i)$ we have

$$g_t = Ge_{j_t} \qquad \text{for } t = 1, \dots, T.$$

The span-respecting condition gives

$$y_{T+1} \in \operatorname{span}\{g_1, \dots, g_T\} = \operatorname{span}\{e_{j_1}, \dots, e_{j_T}\}.$$

This span uses at most T coordinate directions. Since $m = T + 1$, there exists $k \in \{1, \dots, m\}$ outside this set of directions, and hence $(y_{T+1})_k = 0$. Therefore

$$h(y_{T+1}) = G \max_{1\le i\le m} (y_{T+1})_i \ge G(y_{T+1})_k = 0.$$

Now consider the feasible point

$$y^\star := -\frac{R}{\sqrt{m}}\mathbf{1} \in \mathbb{R}^m,$$

where $\mathbf{1} = (1, \dots, 1)^\top$. Then

$$\|y^\star\|_2 = \sqrt{m \cdot \frac{R^2}{m}} = R,$$

so $y^\star$ is feasible for the Euclidean ball $\{y : \|y\|_2 \le R\}$. Moreover,

$$h(y^\star) = G \max_{1\le i\le m} \left(-\frac{R}{\sqrt{m}}\right) = -\frac{GR}{\sqrt{m}}.$$

Hence

$$h(y_{T+1}) - \min_{\|y\|_2\le R} h(y) \ge h(y_{T+1}) - h(y^\star) \ge 0 - \left(-\frac{GR}{\sqrt{m}}\right) = \frac{GR}{\sqrt{T+1}}.$$

This proves the lower bound.
For the complexity consequence, if a zero-initialized linear-span method could guarantee error at most ε uniformly over this objective class, the displayed inequality would force

$$\frac{GR}{\sqrt{T+1}} \le \varepsilon,$$

which is equivalent to

$$T + 1 \ge \frac{G^2R^2}{\varepsilon^2}.$$

□

The same hard instance (h, O_h) defeats every zero-initialized span-respecting transcript:

$$\exists (h, \mathsf{O}_h) \quad \forall \tau, \qquad \text{suboptimality of the last point of } \tau \ge \Delta.$$

An unrestricted deterministic method need not query $x_1 = 0$; the hidden-subspace construction

instead chooses the subspace so that the projected first query satisfies $U^\top x_1 = 0$. Thus, for each method, the reduction constructs a lifted hard instance:

$$\forall A \quad \exists U, f_U, \mathsf{O}_{f_U}, \qquad \text{suboptimality after } T \text{ oracle calls} \geq \Delta.$$

There is no single canonical name for the theorem below. It is a *span-to-black-box reduction*: a lower bound first proved for zero-initialized span-respecting transcripts is converted into a lower bound for arbitrary deterministic first-order methods. The construction is the hidden-subspace version of the classical resisting-oracle or orthogonalization idea.

The low-dimensional oracle (h, O_h) is fixed. What depends on the method A is the embedding $U : \mathbb{R}^m \to \mathbb{R}^D$. During the run we do not choose the full U at once. We maintain a partial isometry U_t only on the already exposed span S_t. When A queries x_t, the visible low-dimensional query is already determined by the old state:

$$y_t := (U_{t-1})^\top x_t \in S_{t-1}.$$

We then temporarily extend U_{t-1} so that the unexposed hidden directions are orthogonal to everything the method has generated so far. More precisely, after the low-dimensional oracle returns g_t, we extend U_{t-1} only to

$$U_t : S_t \to \mathbb{R}^D, \qquad S_t := \operatorname{span}(S_{t-1} \cup \{g_t\}),$$

so each round exposes at most one new hidden direction. That new direction is chosen orthogonal to all queries seen so far, and hence remains invisible to the current transcript.

Theorem 12.3 (Hidden-subspace lifting theorem)**.** *Let $T, m \in \mathbb{N}$, $R > 0$, and $\Delta > 0$. Let $D \in \mathbb{N}$ satisfy*

$$D \geq m + T + 1.$$

Let $h : \mathbb{R}^m \to \mathbb{R}$ be convex, and let O_h be a deterministic subgradient oracle

$$\mathsf{O}_h(y) = (h(y), g(y)), \qquad g(y) \in \partial h(y).$$

Assume that every length-T zero-initialized span-respecting transcript for (h, O_h) satisfies

$$h(y_{T+1}) - \min_{\|y\|_2 \leq R} h(y) \geq \Delta.$$

Then, for every deterministic first-order method A in $\mathbb{R}^D$, there exists a linear isometry $U : \mathbb{R}^m \to \mathbb{R}^D$ such that the lifted function

$$f_U(x) := h(U^\top x), \qquad x \in \mathbb{R}^D,$$

with the lifted oracle

$$\mathsf{O}_{f_U}(x) := (h(U^\top x), U g(U^\top x))$$

forces the output of A after T oracle calls to satisfy

$$f_U(x_{T+1}) - \min_{\|x\|_2 \leq R} f_U(x) \geq \Delta.$$

The elementary regularity and minimizer properties of this lifted composition are recorded in Lemma 12.4.

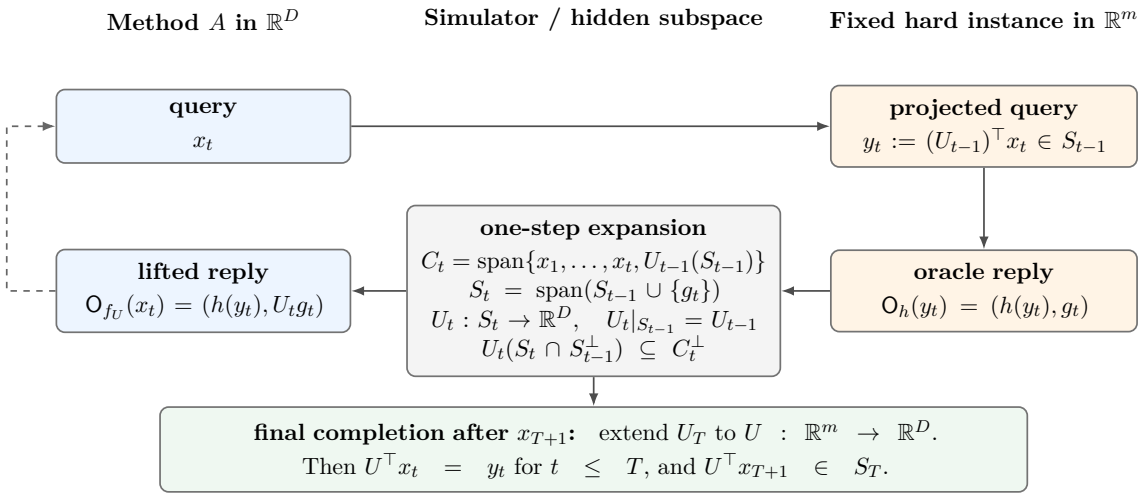

Figure 1: Invariant used in the hidden-subspace lifting proof. Here $S_t = \mathrm{span}\{g_1, \ldots, g_t\}$ is the exposed gradient span, C_t is the visited ambient subspace generated by the queries and the old lifted gradients, and U_t embeds S_t into the ambient space. Each oracle reply adds at most one new hidden direction, chosen orthogonal to C_t; the final completion defines one fixed lifted objective consistent with all oracle replies.

Proof of Theorem 12.3. We construct the hidden isometry together with the transcript seen by A.

Step 1: online construction of the partial isometry. Set

$$S_0 := \{0\} \subseteq \mathbb{R}^m.$$

The first query $x_1 = A(\varnothing) \in \mathbb{R}^D$ is fixed before any oracle reply. The construction will choose the hidden subspace orthogonal to x_1 at the first round, so the projected low-dimensional transcript starts from $y_1 = 0$.

Suppose that, before round t, we have constructed

$$S_{t-1} := \mathrm{span}\,\{g_1, \ldots, g_{t-1}\} \subseteq \mathbb{R}^m$$

and an isometry $U_{t-1} : S_{t-1} \to \mathbb{R}^D$. We call S_{t-1} the exposed gradient span. Let $r_{t-1} := \dim S_{t-1}$. The method has already determined its tth query x_t. Define the visible low-dimensional query using the old partial isometry:

$$y_t := (U_{t-1})^\top x_t \in S_{t-1},$$

where $(U_{t-1})^\top : \mathbb{R}^D \to S_{t-1}$ denotes the adjoint of the partial isometry $U_{t-1} : S_{t-1} \to \mathbb{R}^D$. Define

$$C_t := \mathrm{span}(\{x_1, \ldots, x_t\} \cup U_{t-1}(S_{t-1})) \subseteq \mathbb{R}^D.$$

We call C_t the visited ambient subspace: it contains all queries seen so far and all old lifted gradient directions. In particular, when $t = 1$, this gives $y_1 \in S_0 = \{0\}$, so $y_1 = 0$.

Query the fixed low-dimensional oracle:

$$\mathsf{O}_h(y_t) = (h(y_t), g_t).$$

Set

$$S_t := \mathrm{span}(S_{t-1} \cup \{g_t\}),$$

the updated exposed gradient span. We now extend U_{t-1} only from S_{t-1} to S_t. The new part $S_t \cap S_{t-1}^{\perp}$ has dimension at most one. Since

$$\dim C_t \le t+r_{t-1}, \qquad \text{and hence} \qquad \dim C_t^{\perp} \ge D-t-r_{t-1} \ge m-r_{t-1}+T+1-t \ge \dim(S_t \cap S_{t-1}^{\perp}),$$

we can choose an isometry

$$U_t : S_t \to \mathbb{R}^D$$

such that $U_t|_{S_{t-1}} = U_{t-1}$ and

$$U_t(S_t \cap S_{t-1}^{\perp}) \subseteq C_t^{\perp}.$$

The newly added directions are orthogonal to all queries seen so far. Hence $(U_t)^{\top} x_q = (U_{t-1})^{\top} x_q$ for every $q \le t$. For $q < t$, this common projection equals y_q by the induction invariant, and for $q = t$ it equals the current definition $y_t = (U_{t-1})^{\top} x_t$.
Return to A the lifted reply, where $\widehat{g}_t$ denotes the covector in the ambient space:

$$(\ell_t, \widehat{g}_t) := (h(y_t), U_t g_t).$$

Since A is deterministic, this reply uniquely determines the next query. This completes the induction through round T.
Step 2: final completion to a fixed hidden instance. After the Tth oracle reply, the method outputs x_{T+1}. We now complete the final hidden isometry. Let

$$C_{T+1} := \operatorname{span}(\{x_1, \dots, x_{T+1}\} \cup U_T(S_T)).$$

The same dimension count gives enough room to extend U_T to a full isometry

$$U : \mathbb{R}^m \to \mathbb{R}^D$$

such that $U = U_T$ on S_T and $U(S_T^{\perp}) \subseteq C_{T+1}^{\perp}$. By construction, the final projection agrees with the projection used when each query was answered:

$$U^{\top} x_t = y_t, \qquad t = 1, \dots, T,$$

and also

$$y_{T+1} := U^{\top} x_{T+1} \in S_T = \operatorname{span}\{g_1, \dots, g_T\}.$$

Therefore

$$(y_1, g_1), \dots, (y_T, g_T), y_{T+1}$$

is a length-T zero-initialized span-respecting transcript for the fixed oracle (h, O_h). The assumed fixed-instance lower bound gives

$$h(y_{T+1}) - \min_{\|y\|_2 \le R} h(y) \ge \Delta.$$

Step 3: identify the transcript with a genuine lifted oracle run. For every actual query $t \le T$, $U^{\top} x_t = y_t$, so $f_U(x_t) = h(y_t) = \ell_t$. Since U agrees with U_t on g_t, the returned ambient covector $\widehat{g}_t = U_t g_t$ is exactly $U g_t$. Hence the transcript constructed above is exactly the transcript of A on the final lifted instance.
By Lemma 12.4,

$$\min_{\|x\|_2 \le R} f_U(x) = \min_{\|y\|_2 \le R} h(y).$$

Therefore

$$f_U(x_{T+1}) - \min_{\|x\|_2 \le R} f_U(x) = h(y_{T+1}) - \min_{\|y\|_2 \le R} h(y) \ge \Delta.$$

The validity, regularity, and centered-minimizer properties of the lifted objective are exactly Lemma 12.4. □

Lemma 12.4 (Lifted composition preserves regularity, minimizers, and gaps)**.** *Let $U : \mathbb{R}^m \to \mathbb{R}^D$ be a linear isometry. Let $h : \mathbb{R}^m \to \mathbb{R}$ be convex, and let $\mathsf{O}_h(y) = (h(y), g(y))$ be a deterministic subgradient oracle for h. Define*

$$f_U(x) := h(U^\top x), \qquad \mathsf{O}_{f_U}(x) := \big(h(U^\top x), Ug(U^\top x)\big).$$

Then f_U is convex and O_{f_U} is a valid deterministic subgradient oracle for f_U. For every $R \geq 0$,

$$\inf_{\|x\|_2 \leq R} f_U(x) = \inf_{\|y\|_2 \leq R} h(y),$$

and the same equality holds with min *whenever the minima exist. If h is G-Lipschitz with respect to $\|\cdot\|_2$, then f_U is G-Lipschitz. If h is differentiable and L-smooth with respect to $\|\cdot\|_2$, then*

$$\nabla f_U(x) = U\nabla h(U^\top x).$$

Then f_U is L-smooth. Finally, if $y^\star \in \arg\min h$, then $x^\star := Uy^\star$ belongs to $\arg\min f_U$,

$$\|x^\star\|_2 = \|y^\star\|_2 ,$$

and, for every $x \in \mathbb{R}^D$,

$$f_U(x) - f_U(x^\star) = h(U^\top x) - h(y^\star).$$

Proof of Lemma 12.4. Convexity follows by composing the convex function h with the linear map $x \mapsto U^\top x$. For oracle validity, if $\mathsf{O}_h(U^\top x) = (h(U^\top x), g(U^\top x))$, then for every $z \in \mathbb{R}^D$,

$$h(U^\top z) \geq h(U^\top x) + \left\langle g(U^\top x), U^\top z - U^\top x \right\rangle = f_U(x) + \left\langle Ug(U^\top x), z - x \right\rangle .$$

Thus $Ug(U^\top x) \in \partial f_U(x)$.
For the ball minimum, $\left\|U^\top x\right\|_2 \leq \|x\|_2$, so

$$\inf_{\|x\|_2 \leq R} h(U^\top x) \geq \inf_{\|y\|_2 \leq R} h(y).$$

Conversely, every y with $\|y\|_2 \leq R$ is realized by $x = Uy$, so the reverse inequality also holds.
If h is G-Lipschitz, then

$$|f_U(x) - f_U(z)| = \left|h(U^\top x) - h(U^\top z)\right| \leq G \left\|U^\top (x - z)\right\|_2 \leq G \|x - z\|_2 .$$

If h is differentiable, the chain rule gives

$$\nabla f_U(x) = U\nabla h(U^\top x).$$

If h is L-smooth, then

$$\|\nabla f_U(x) - \nabla f_U(z)\|_2 = \left\|U(\nabla h(U^\top x) - \nabla h(U^\top z))\right\|_2 \leq L \left\|U^\top (x - z)\right\|_2 \leq L \|x - z\|_2 .$$

Thus f_U is L-smooth.
Now suppose $y^\star \in \arg\min h$, and set $x^\star := Uy^\star$. Since $U^\top x^\star = y^\star$, for every x,

$$f_U(x) = h(U^\top x) \geq h(y^\star) = f_U(x^\star).$$

Thus $x^\star \in \arg\min f_U$, and the displayed gap equality follows from $U^\top x^\star = y^\star$. Finally,

$$\|x^\star\|_2 = \|Uy^\star\|_2 = \|y^\star\|_2 .$$

□

Remark 12.1 (What the lifting theorem proves). Theorem 12.3 proves a fixed-hard-instance reduction. It does not show that an arbitrary deterministic method is linear-span in the ambient space. Instead it shows that, after a fixed low-dimensional hard instance is hidden in an adaptively chosen subspace, the projected transcript satisfies

$$U^\top x_{t+1} \in \operatorname{span}\{g_1, \dots, g_t\}.$$

Equivalently, only the projection of x_{t+1} onto the hidden subspace must lie in the lifted span. The ambient iterate x_{t+1} may have arbitrary orthogonal components, but those components are invisible to $f_U(x) = h(U^\top x)$.

Theorem 12.5 (Unrestricted deterministic nonsmooth lower bound). *Let $T \in \mathbb{N}$, $G > 0$, $R > 0$, and let $D \geq 2T + 2$. For every deterministic first-order method A in $\mathbb{R}^D$, there exist a convex G-Lipschitz function $f : \mathbb{R}^D \to \mathbb{R}$ and a deterministic subgradient oracle for f such that, when A is run for exactly T oracle calls on this oracle, its output x_{T+1} satisfies*

$$f(x_{T+1}) - \min_{\|x\|_2 \leq R} f(x) \geq \frac{GR}{\sqrt{T+1}}.$$

Consequently, order G^2R^2/ε^2 oracle calls are necessary even without the linear-span restriction, in sufficiently high dimension.

Proof of Theorem 12.5. Apply Theorem 12.2 with $m = T + 1$. It gives a fixed low-dimensional hard instance (h, O_h) with

$$\Delta = \frac{GR}{\sqrt{T+1}}.$$

Since $D \geq 2T + 2 = m + T + 1$, Theorem 12.3 lifts this fixed instance into $\mathbb{R}^D$. By Lemma 12.4, the lifted objective is convex and G-Lipschitz, and the lifted oracle is a valid deterministic subgradient oracle. The displayed lower bound is exactly the lifted lower bound. The complexity consequence follows by requiring $GR/\sqrt{T+1} \leq \varepsilon$. □

The smooth lower-bound section repeats the same logic with a different fixed hard instance. First Theorem 12.6 proves

$$\exists (h, \mathsf{O}_h) \quad \forall \tau, \qquad h(y_{T+1}) - h(y^\star) \gtrsim \frac{LR^2}{T^2}.$$

Then Theorem 12.7 applies the lifting theorem to obtain the unrestricted deterministic statement

$$\forall A \quad \exists f, \mathsf{O}_f, x^\star, \qquad f(x_{T+1}) - f(x^\star) \gtrsim \frac{LR^2}{T^2}.$$

Theorem 12.6 (Smooth convex zero-initialized span-respecting lower bound). *For every $T \in \mathbb{N}$, every $L > 0$, and every $R > 0$, there exist an integer $d = T + 1$, an L-smooth convex quadratic function $h : \mathbb{R}^d \to \mathbb{R}$, and a minimizer $y^\star \in \arg\min h$ with*

$$\|y^\star\|_2 \leq R$$

such that every length-T zero-initialized span-respecting transcript for the exact first-order oracle

$\mathsf{O}_h(y) = (h(y), \nabla h(y))$ *satisfies*

$$h(y_{T+1}) - h(y^\star) \geq \frac{LR^2}{8(T+1)^2}.$$

Proof of Theorem 12.6. Set $d := T + 1$. Let $B \in \mathbb{R}^{d\times d}$ be the symmetric positive semidefinite matrix defined by the quadratic form

$$y^\top By := y_1^2 + \sum_{i=1}^{d-1}(y_i - y_{i+1})^2.$$

For every $y \in \mathbb{R}^d$,

$$y^\top By \leq y_1^2 + 2\sum_{i=1}^{d-1}(y_i^2 + y_{i+1}^2) \leq 4\|y\|_2^2.$$

Hence $\|B\|_{\mathrm{op}} \leq 4$.
Let

$$\alpha := \frac{R}{\sqrt{d}}, \qquad h(y) := \frac{L}{8}y^\top By - \frac{L\alpha}{4}e_1^\top y.$$

Since $\nabla^2 h = (L/4)B \succeq 0$, the function h is convex. Since $\|B\|_{\mathrm{op}} \leq 4$, we also have

$$\left\|\nabla^2 h\right\|_{\mathrm{op}} \leq L,$$

so h is L-smooth with respect to the Euclidean norm.
The minimizer solves

$$By = \alpha e_1.$$

The constant vector

$$y^\star := \alpha\mathbf{1}$$

satisfies this equation because

$$B\mathbf{1} = e_1.$$

Indeed, all consecutive-difference terms vanish on a constant vector, while the boundary term y_1^2 contributes one unit in the first coordinate. Thus $y^\star \in \arg\min h$, and

$$\|y^\star\|_2 = \sqrt{d\alpha^2} = R.$$

We next prove the information restriction. If a vector x is supported on coordinates $\{1, \ldots, k\}$, then Bx is supported on $\{1, \ldots, k+1\}$. Since

$$\nabla h(x) = \frac{L}{4}(Bx - \alpha e_1),$$

the gradient at such an x is also supported on $\{1, \ldots, k+1\}$. Starting from $y_1 = 0$, the span-respecting condition therefore gives by induction

$$\mathrm{supp}(y_t) \subseteq \{1, \ldots, t-1\} \qquad \text{for } t = 1, \ldots, T+1.$$

In particular,

$$(y_{T+1})_d = 0.$$

Because h is quadratic and $\nabla h(y^\star) = 0$, for every $y \in \mathbb{R}^d$,

$$h(y) - h(y^\star) = \frac{L}{8}(y - y^\star)^\top B(y - y^\star).$$

Let $z := y_{T+1} - y^\star$. Then $z_d = -\alpha$. Also, if we set $z_0 := 0$, then

$$z^\top B z = z_1^2 + \sum_{i=1}^{d-1}(z_i - z_{i+1})^2 = \sum_{i=0}^{d-1}(z_{i+1} - z_i)^2.$$

By Cauchy's inequality,

$$\alpha^2 = (z_d - z_0)^2 = \left(\sum_{i=0}^{d-1}(z_{i+1} - z_i)\right)^2 \le d\sum_{i=0}^{d-1}(z_{i+1} - z_i)^2 = d\, z^\top B z.$$

Therefore

$$z^\top B z \ge \frac{\alpha^2}{d} = \frac{R^2}{d^2}.$$

Consequently,

$$h(y_{T+1}) - h(y^\star) \ge \frac{L}{8}\frac{R^2}{d^2} = \frac{LR^2}{8(T+1)^2}.$$

This proves the theorem. □

Theorem 12.7 (Unrestricted deterministic smooth convex lower bound)**.** *Let $T \in \mathbb{N}$, $L > 0$, $R > 0$, and $D \ge 2T + 2$. For every deterministic first-order method A in $\mathbb{R}^D$, there exist a convex L-smooth quadratic function $f : \mathbb{R}^D \to \mathbb{R}$, its exact first-order oracle, and a minimizer $x^\star \in \arg\min f$ such that the problem instance satisfies*

$$\|x^\star\|_2 \le R,$$

and the output after T oracle calls satisfies

$$f(x_{T+1}) - f(x^\star) \ge \frac{LR^2}{8(T+1)^2}.$$

Proof of Theorem 12.7. Apply Theorem 12.6 with $m = d = T + 1$. Since $y^\star \in \arg\min h$ and $\|y^\star\|_2 = R$, we have

$$\min_{\|y\|_2 \le R} h(y) = h(y^\star),$$

so the theorem gives the fixed-instance lower bound required by Theorem 12.3 with

$$\Delta = \frac{LR^2}{8(T+1)^2}.$$

Because $D \ge 2T + 2 = m + T + 1$, the lifting theorem constructs a lifted quadratic

$$f(x) = h(U^\top x)$$

and an exact first-order oracle for f. By Lemma 12.4, the lifted function is convex and L-smooth, and it has a minimizer $x^\star$ with $\|x^\star\|_2 \le R$. The lifting construction also gives a zero-initialized

span-respecting transcript whose last point is $y_{T+1} = U^\top x_{T+1}$. Therefore Theorem 12.6 applies to this projected transcript. The equality

$$f(x_{T+1}) - f(x^\star) = h(y_{T+1}) - h(y^\star)$$

from Lemma 12.4 gives the displayed inequality. □

Remark 12.2 (References for the smooth deterministic lower bounds). Theorem 12.6 and Theorem 12.7 prove the classical smooth-convex lower-bound scale from [Nes04, Theorem 2.1.7]. The smooth strongly-convex lower bound in the summary table below is the classical $\Omega(\sqrt{\kappa}\log(1/\varepsilon))$ benchmark from [Nes04, Theorem 2.1.13].

Rate summary

The table below hides only absolute numerical constants. In the first six rows, all Lipschitz, smoothness, strong-convexity, and radius assumptions are with respect to the Euclidean norm $\|\cdot\|_2$. For the smooth rows, the table also lists the useful non-accelerated baseline when it differs from the optimal accelerated rate. The last two rows state the normed geometry explicitly. In the stochastic nonsmooth row, G denotes the total second-moment scale of the stochastic subgradient; under the G_0, σ decomposition of Corollary 9.9, replace G by $\sqrt{G_0^2 + \sigma^2}$.[1]

[1] For a general norm $\|\cdot\|$ with dual norm $\|\cdot\|_*$, (s, L)-Hölder smoothness means

$$\|\nabla f(x) - \nabla f(y)\|_* \le L\,\|x - y\|^{s-1}, \qquad 1 < s \le 2.$$

This implies the upper bound $f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{s}\,\|y - x\|^s$. The case $s = 2$ is the usual L-smoothness assumption.

Table 1: First-order oracle lower bounds, non-accelerated baselines, and matching or near-matching upper bounds. The first six rows use the Euclidean norm; the last two rows state the normed geometry explicitly.

Setting	Lower bound	Upper bound	Upper-bound algorithm
Nonsmooth convex, G-Lipschitz, radius R	$\Omega(GR/\sqrt{T})$, Theorem 12.5	$O(GR/\sqrt{T})$, Corollary 9.5	Subgradient / mirror descent
Nonsmooth μ-strongly convex, G-Lipschitz	$\Omega(G^2/(\mu T))$, see [NY83, Nes04]	$O(G^2/(\mu T))$, Corollary 9.7	Weighted mirror descent
Smooth convex, L-smooth, radius R	$\Omega(LR^2/T^2)$, Theorem 12.7	non-accelerated $O(LR^2/T)$, Theorem 7.12; optimal $O(LR^2/T^2)$, [Nes04]	Gradient descent; accelerated gradient
Smooth μ-strongly convex, L-smooth, $\kappa = L/\mu$	$\Omega(\sqrt{\kappa}\log(1/\varepsilon))$ oracle calls, [Nes04]	non-accelerated $O(\kappa\log(1/\varepsilon))$, Theorem 7.8; optimal $O(\sqrt{\kappa}\log(1/\varepsilon))$, [Nes04]	Gradient descent; accelerated gradient
Stochastic nonsmooth convex, stochastic-subgradient second moment scale G, radius R	$\Omega(GR/\sqrt{T})$, [ABRW12]	$O(GR/\sqrt{T})$, Corollary 9.9	Stochastic mirror descent
Stochastic smooth convex, L-smooth, variance σ^2, radius R	$\Omega(LR^2/T^2 + \sigma R/\sqrt{T})$, from [Nes04, ABRW12]	$O(LR^2/T + \sigma R/\sqrt{T})$, Theorem 9.10; optimal $O(LR^2/T^2 + \sigma R/\sqrt{T})$, [Lan12]	SMD; AC-SA for the optimal rate
Smooth convex over an ℓ_p-ball of radius R, $2 \le p < \infty$, (s, L)-Hölder smooth in $\|\cdot\|_p$, $T \le n$	$\Omega\left(\frac{LR^s}{(\min\{p, \log T\})^{s-1} T^{s+s/p-1}}\right)$, [GN15]	$O\left(\frac{LR^s}{T^{s+s/p-1}}\right)$ for fixed p, see [GN15]	Non-Euclidean acceleration
Smooth convex over an ℓ_∞-ball of radius R, L-smooth in $\|\cdot\|_\infty$	$\Omega(LR^2/(T\log T))$, [GN15]	$O(LR^2/T)$, Theorem 11.3	Frank–Wolfe / conditional gradient

Extended reading

For deterministic Euclidean lower bounds, see the classical oracle-complexity treatment in [NY83, Nes04]; the smooth convex and smooth strongly-convex entries in Table 1 are the benchmark results used later in the course. For stochastic oracle lower bounds, see the information-theoretic framework of [ABRW12]; the accelerated stochastic upper bound in Table 1 is due to [Lan12]. For randomized nonsmooth black-box lower bounds, the right framework is distributional oracle complexity; see [BGP17]. For parallel and randomized local-oracle lower bounds, see [DG20]. For the large-scale ℓ_p and ℓ_∞ smooth convex lower bounds, see [GN15].

Dependency and proof sketch

1. Definition 12.3 is the algorithm-level zero-initialized restriction, while Definition 12.4 is the transcript-level abstraction. The latter is the right object for the lifting argument.

2. Lemma 12.1 packages the hard instance used in the lecture: the max-coordinate function is convex, nonsmooth, and has coordinate basis vectors as valid subgradients.

3. Theorem 12.2 is the complete in-class lower-bound proof. The core observation is combinatorial rather than analytic: after T oracle calls, a zero-initialized span-respecting transcript can only live in the span of at most T coordinate directions, so one coordinate remains untouched.

4. Figure 1, Theorem 12.3, Lemma 12.4, and Remark 12.1 explain how the fixed low-dimensional hard instance is hidden in a larger ambient space. The important invariant is not that x_t itself is linear-span, but that its hidden projection $U^\top x_t$ is. The two lifted-composition lemmas then isolate the routine preservation facts: oracle validity, Lipschitzness, smoothness, and the centered minimizer certificate needed for smooth lower bounds.

5. Theorem 12.5 is the unrestricted deterministic nonsmooth lower bound obtained by applying Theorem 12.3 to Theorem 12.2. It clarifies that the zero-initialized transcript restriction in Theorem 12.2 is a device for obtaining a short explicit proof, not the final word on oracle complexity.

6. Theorem 12.6 is the smooth quadratic-chain lower bound for zero-initialized span-respecting transcripts. Theorem 12.7 then applies the same lifting theorem to obtain the unrestricted deterministic smooth convex lower bound.

7. Table 1 records the smooth strongly-convex benchmark needed for the rest of the course: the optimal first-order oracle complexity has the scale $\sqrt{\kappa}\log(1/\varepsilon)$.

Exercises

1. In the proof of Theorem 12.2, identify exactly where the zero-initialized span-respecting structure is used. Convexity and G-Lipschitzness of the hard instance are class-membership constraints (the lower bound is proved against the class of G-Lipschitz convex objectives), not tools used by the proof itself.

2. Replace the Euclidean ball by the ℓ_∞ ball and redo the hard-instance calculation for the max-coordinate function.

3. Verify directly that the matrix B in the proof of Theorem 12.6 satisfies $\|B\|_{\mathrm{op}} \le 4$, and compare the chain-support mechanism with the support argument in Theorem 12.2.

4. Explain why the smooth strongly-convex row in Table 1 is the right lower bound to compare against accelerated gradient methods in the strongly convex regime.

References

[ABRW12] Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

[BGP17] Gábor Braun, Cristóbal Guzmán, and Sebastian Pokutta. Lower bounds on the oracle complexity of nonsmooth convex optimization via information theory. *IEEE Transactions on Information Theory*, 63(7):4709–4724, 2017.

[DG20] Jelena Diakonikolas and Cristóbal Guzmán. Lower bounds for parallel and randomized convex optimization. *Journal of Machine Learning Research*, 21(5):1–31, 2020.

[GN15] Cristóbal Guzmán and Arkadi Nemirovski. On lower complexity bounds for large-scale smooth convex optimization. *Journal of Complexity*, 31(1):1–14, 2015.

[Lan12] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1–2):365–397, 2012.

[Nes04] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004.

[NY83] Arkadi S. Nemirovski and David B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, New York, 1983.