CS395T: Continuous Algorithms, Part II
Gradient descent

Kevin Tian

1 Oracle model

In this lecture, and several of the following lectures, we study continuous optimization in what is commonly referred to as the "oracle model." Specifically, we assume there is a function $f : \mathcal{X} \to \mathbb{R}$ which we wish to minimize, and we can only access f through an oracle $\mathcal{O}$. Clearly, we cannot assume too strong of an oracle, e.g., an oracle which simply returns the minimizer of f would make our job uninteresting. Typically, our oracle will provide value or derivative information about f.[1] In this lecture, we only characterize algorithms accessing zeroth-order and first-order information about f, through a value oracle and a subgradient oracle (see Definition 5, Part I).

There are many reasons why the oracle model is an appealing framework for developing optimization theory. For one, it is an effective separation between the complexity of an algorithm given access to an oracle, and the complexity of implementing the oracle. From an algorithm design standpoint, this allows us to focus our attention on the different components of the algorithm in a rather modular way. Moreover, the oracle model is of practical interest, as certain operations (e.g., computing gradients of a neural network through backpropagation) may be highly-optimized through specialized hardware (e.g., cache-aware algorithms for matrix-vector multiplication). These operations hence serve as reasonable "units" of measurement for the complexity of an algorithm.

Finally, as we aim to demonstrate in this lecture, the oracle model is an effective way to benchmark different approaches to algorithm design (e.g., do we really need to query higher-order information, or are values enough?) due to the feasibility of proving lower bounds. In various situations, some of which we will soon encounter, matching lower and upper bounds are known for algorithms in the oracle model. These bounds are typically phrased in the following setting.

1. There is a family $\mathcal{F}$ of functions. These functions are typically assumed to share some type of common structure or regularity, but otherwise may be arbitrary within the family.
2. An algorithm $\mathcal{A}$ can interact with $f \in \mathcal{F}$ only through an oracle, which returns information about f such as its value or derivatives. We may also make some assumption about $\mathcal{A}$ beyond the oracle restriction, e.g., it only moves in directions suggested by the oracle.

The goal is then to characterize the number of oracle calls any algorithm $\mathcal{A}$ (with the specified additional restrictions) must use to solve the optimization task it aims to accomplish. We typically measure the oracle complexity of $\mathcal{A}$ by its worst-case performance over functions $f \in \mathcal{F}$.

2 Lipschitz optimization

Recall that in Section 3, Part I, we established the following remarkable result: all convex optimization problems admit high-accuracy polynomial-time solvers. We could ask the even more ambitious question: is any structure at all necessary to design optimization algorithms, or are there general-purpose frameworks that always work, such as the cutting-plane method for convex problems? The following simple lower bound demonstrates at least some structure is necessary, even in dimension 1 and assuming a bounded domain and range.

[1] There are interesting exceptions beyond the setting of derivatives. For example, [CJJ+20] initiated a line of work characterizing algorithm complexities accessing a *ball optimization oracle*, which returns an (approximate) minimizer of a function in a small Euclidean ball. This applies to functions exhibiting local (but not global) regularity.

Lemma 1. *No algorithm $\mathcal{A}$ which accesses a target $f : [0,1] \to [0,1]$ using a value oracle can optimize f to additive error < 1 in a finite number of queries.*

Proof. Consider a family $\{f_x\}_{x\in[0,1]}$, where $f_x = 1$ everywhere except $f_x(x) = 0$. To optimize f_x to additive error < 1, we learn what x is. Given a finite number of oracle queries, there are infinitely many f_x consistent with the answers if they are all 1, so learning x is a contradiction. □

The example of Lemma 1 is rather extreme, but already highlights a major obstacle to unstructured optimization: it is difficult to gain information about where the optimizer lies. One natural way to impose some regularity on a target function is to ask that it does not change too rapidly, so a near-optimizer is necessarily close to the true optimizer. This motivates the following definition.

Definition 1 (Lipschitzness)**.** *We say $f : \mathcal{X} \to \mathbb{R}$ is L-*Lipschitz *with respect to the norm $\|\cdot\|$ if*

$$|f(\mathbf{x}) - f(\mathbf{x}')| \le L \|\mathbf{x} - \mathbf{x}'\|, \text{ for all } \mathbf{x}, \mathbf{x}' \in \mathcal{X}.$$

If $\|\cdot\| = \|\cdot\|_2$, we simply say f is L-Lipschitz.

The following demonstrates a simple example of a Lipschitz function.

Lemma 2. *Let $f(\mathbf{x}) = \|\mathbf{x}\|$ for some norm $\|\cdot\|$. Then, f is 1-Lipschitz with respect to $\|\cdot\|$.*

Proof. By the triangle inequality, both $f(\mathbf{x}) - f(\mathbf{x}')$ and $f(\mathbf{x}') - f(\mathbf{x})$ are at most $\|\mathbf{x} - \mathbf{x}'\|$. □

Unfortunately, a small modification of Lemma 1 dashes hopes of efficiently minimizing all Lipschitz functions as well. To construct our lower bound, we use the following definition.

Definition 2 (Packings and coverings)**.** *Let $S, T \subset \mathbb{R}^d$ and $\epsilon > 0$.*

- *We say $\{T_i\}_{i\in[n]}$ is an* ϵ-packing *of S by T if there are $\{\mathbf{x}_i\}_{i\in[n]} \subset \mathbb{R}^d$ such that $T_i = \{\mathbf{x}_i\} \oplus \epsilon T$ for all $i \in [n]$, and all $T_i \cap T_j = \emptyset$ for $i \neq j$.*
- *We say $\{T_i\}_{i\in[n]}$ is an* ϵ-covering *of S by T if there are $\{\mathbf{x}_i\}_{i\in[n]} \subset \mathbb{R}^d$ such that $T_i = \{\mathbf{x}_i\} \oplus \epsilon T$ for all $i \in [n]$, and $\bigcup_{i\in[n]} T_i \supseteq S$.*

The following fundamental relationship between packings and coverings is often useful.

Lemma 3. *Let $\{T_i\}_{i\in[n]} = \{\{\mathbf{x}_i\} \oplus \epsilon T\}_{i\in[n]}$ be a maximal ϵ-packing of S by T, i.e., no $\{\mathbf{x}'\} \oplus \epsilon T$ can be added while remaining an ϵ-packing. Then $\{\{\mathbf{x}_i\} \oplus 2\epsilon T\}_{i\in[n]}$ is a 2ϵ-covering of S by T.*

Proof. Suppose there was $\mathbf{x}' \in S$ such that $\mathbf{x}$ is not covered by $\{\{\mathbf{x}_i\} \oplus 2\epsilon T\}_{i\in[n]}$. Then $\{\mathbf{x}'\} \oplus \epsilon T$ is disjoint from each of the $\{T_i\}_{i\in[n]}$, a contradiction to maximality of the packing. □

Corollary 1. *For any $\epsilon \in (0,1)$, there is an ϵ-packing of $S \subset \mathbb{R}^d$ by S of cardinality $\ge (\frac{1}{2\epsilon})^d$.*

Proof. By Lemma 3, the maximum cardinality of an ϵ-packing of S by S is larger than the minimum cardinality of a 2ϵ-covering of S by S, since any maximum cardinality packing is clearly a maximal packing. The conclusion follows by lower bounding the cardinality of any 2ϵ-covering $\{S_i = \{\mathbf{x}_i\} \oplus 2\epsilon S\}_{i\in[n]}$ of S, by using a volume argument:

$$n\,(2\epsilon)^d \operatorname{Vol}(S) = n\operatorname{Vol}(2\epsilon S) \ge \operatorname{Vol}\left(\bigcup_{i\in[n]} S_i\right) \ge \operatorname{Vol}(S) \implies n \ge \left(\frac{1}{2\epsilon}\right)^d.$$

□

Armed with Corollary 1, our lower bound on Lipschitz function minimization follows.

Lemma 4. *Let $L > \epsilon > 0$. No algorithm $\mathcal{A}$ which accesses a target L-Lipschitz function $f : \mathbb{B}(\mathbf{0}_d, 1) \to \mathbb{R}$ using a value oracle can optimize f to additive error ϵ in $< (\frac{L}{2\epsilon})^d - 1$ queries.*

Proof. Let $\{S_i = \mathbb{B}(\mathbf{x}_i, \epsilon)\}$ be an $\frac{\epsilon}{L}$-packing of $\mathbb{B}(\mathbf{0}_d, 1)$ by itself of cardinality $\geq (\frac{L}{2\epsilon})^d$, which exists by Corollary 1. Consider a family of functions $\{f_i\}_{i\in[n]}$, where

$$f_i(x) := \min\left(0, -\epsilon + L\left\|\mathbf{x} - \mathbf{x}_i\right\|_2\right).$$

We observe that f_i is L-Lipschitz by casework on Lemma 2 (i.e., considering the cases where $\mathbf{x}, \mathbf{x}'$ are variously in or outside S_i), and further outside S_i, $f_i(\mathbf{x})$ is zero. To optimize f_i to additive error ϵ, we must find a point within distance ϵ of $\mathbf{x}_i$, which uniquely identifies $i \in [n]$ by the definition of a packing, so we must learn what $i \in [n]$ is. Given $< (\frac{L}{2\epsilon})^d - 1$ oracle queries, there must be some distinct $i, j \in [n]$ such that no point in S_i or S_j is queried, so f_i and f_j are both consistent with the oracle answers (which are all zero). Hence, learning $i \in [n]$ is a contradiction. □

Lemma 4 highlights the weakness of a value oracle to distinguish functions which agree on large parts of space. Fortunately, we already have seen a structural assumption which yields a tool for learning more about the minimizer of a function: convexity, which comes with existence of subgradients. As we saw in Part I, subgradients are separating hyperplanes between iterates of an algorithm and the minimizer, enabling cutting-plane methods to solve optimization problems.

We explore in this section how convexity can help in a different way, by using subgradient information as a "descent direction." Before giving our first algorithm, we establish fundamental limits on Lipschitz convex function minimization with a subgradient oracle, from [NY83]. Our lower bound holds under the natural assumption that iterates of an algorithm are only allowed to move within the span of subgradients returned by the oracle. This assumption can be removed, and the lower bound extends to hold information-theoretically against randomized algorithms [ABRW12].

Theorem 1 (Lipschitz convex lower bound)**.** *Let $\epsilon, L, R > 0$. No algorithm $\mathcal{A}$ which accesses a target L-Lipschitz, convex function $f : \mathbb{B}(\mathbf{0}_d, R) \to \mathbb{R}$ using a subgradient oracle $\mathcal{O}$ and produces iterates $\{\mathbf{x}_t\}_{0\leq t<T}$ can optimize f to additive error ϵ using $T < \min(d, (\frac{LR}{4\epsilon})^2)$ queries, subject to the restriction that*

$$\mathbf{x}_0 = \mathbf{0}_d,\ \mathbf{x}_t \in \text{Span}\left(\{\mathcal{O}(\mathbf{x}_s)\}_{0\leq s<t}\right) \text{ for all } t \in [T]. \tag{1}$$

Proof. Assume $T < d$, else there is nothing to prove. Consider, for $\gamma, \alpha > 0$ to be chosen:

$$f(\mathbf{x}) := \gamma \max_{i\in[T]} \mathbf{x}_i + \frac{\alpha}{2}\left\|\mathbf{x}\right\|_2^2. \tag{2}$$

It is straightforward to check that $f(\mathbf{x})$ is convex and $(\gamma + \alpha R)$-Lipschitz, and that

$$\partial f(\mathbf{x}) = \gamma \cdot \text{Conv}\left(\left\{\mathbf{e}_i \mid i \in \text{argmax}_{j\in[T]}\mathbf{x}_j\right\}\right) + \alpha\mathbf{x}.$$

We will use a subgradient oracle $\mathcal{O}$ which returns $\mathcal{O}(\mathbf{x}) = \gamma\mathbf{e}_i + \alpha\mathbf{x}$, for the smallest $i \in [T]$ satisfying $i \in \text{argmax}_{j\in[T]}\mathbf{x}_j$. We claim that for all $0 \leq t < T$, $\mathbf{x}_t$ is only supported in the first t coordinates, which follows inductively from our oracle definition and the assumption (1). In particular, note that either $\mathbf{x}_t$ has negative values in all of its first t coordinates (in which case $\mathcal{O}(\mathbf{x}_t) = \mathbf{e}_{t+1}$), or $\mathcal{O}(\mathbf{x}_t) = \mathbf{e}_i$ for some $i \in [t]$. Therefore, $f(\mathbf{x}_t) \geq 0$ for all $0 \leq t < T$. On the other hand,

$$\mathbf{x}^\star := \sum_{i\in[T]}\left(-\frac{\gamma}{\alpha T}\right)\mathbf{e}_i \implies f(\mathbf{x}^\star) = -\frac{\gamma^2}{2\alpha T},$$

and $\mathbf{0}_d \in \partial f(\mathbf{x}^\star)$, so $\mathbf{x}^\star$ minimizes f by Lemma 7, Part I, assuming $\left\|\mathbf{x}^\star\right\|_2 \leq R$. Therefore, $\mathcal{A}$ has not produced any iterate achieving additive error $\frac{\gamma^2}{2\alpha T}$. Choosing $\gamma = \frac{L}{2}$ and $\alpha = \frac{L}{2R\sqrt{T}}$ yields $\left\|\mathbf{x}^\star\right\|_2 \leq R$ and $\gamma + \alpha R \leq L$, and rearranging gives the claimed lower bound on T:

$$\epsilon > \frac{\gamma^2}{2\alpha T} = \frac{LR}{4\sqrt{T}} \iff T > \left(\frac{LR}{4\epsilon}\right)^2.$$

□

Remark 1. *Theorem 1 further shows that $\Omega(\min(d, \frac{L^2}{\alpha\epsilon}))$ queries are necessary in the setting where f is both L-Lipschitz and α-strongly convex (see Definition 4, to be introduced). This lower bound is also optimal; for simple derivations of a matching upper bound, see [LSB12, RSS12].*

The function in (2) is known as Nemirovski's function, and is a common example used in lower bound constructions in the subgradient oracle model. Intuitively, it exploits the inability of the subgradient oracle to reveal more than "one new coordinate" under the assumption (1), because until the algorithm sees T subgradients it cannot find an approximate minimizer.

Interestingly, up to logarithmic terms we have already established an upper bound matching the first component of the lower bound in Theorem 1. Indeed, recall that the center of gravity method from Part I uses $\approx d \log(\frac{d}{\epsilon})$ queries to achieve ϵ error.[2] In the remainder of the section, we show that a simple subgradient-based algorithm matches the other component as well. We begin by proving an alternate characterization of Lipschitzness, which will frequently be used.

Lemma 5. *Let $f : \mathcal{X} \to \mathbb{R}$ be convex and L-Lipschitz for $\mathcal{X} \subseteq \mathbb{R}^d$, and let $\mathbf{x} \in \text{relint}(\mathcal{X})$ and $\mathbf{g} \in \partial f(\mathbf{x})$ be contained in the lowest-dimensional subspace containing $\mathcal{X}$. Then $\|\mathbf{g}\|_2 \leq L$.*

Proof. Suppose for contradiction that there exists such $f, \mathbf{x}, \mathbf{g}$ with $\|\mathbf{g}\|_2 > L$. Let $\epsilon > 0$ be such that $\mathbf{x}' := \mathbf{x} + \epsilon \mathbf{g} \in \mathcal{X}$, since $\mathbf{x} \in \text{relint}(\mathcal{X})$. Then, we have a contradiction: Lipschitzness of f implies that $f(\mathbf{x}') \leq f(\mathbf{x}) + \epsilon L \|\mathbf{g}\|_2$, but $\mathbf{g} \in \partial f(\mathbf{x})$ yields

$$f(\mathbf{x}') \geq f(\mathbf{x}) + \langle \mathbf{g}, \epsilon \mathbf{g} \rangle = f(\mathbf{x}) + \epsilon \|\mathbf{g}\|_2^2 > f(\mathbf{x}) + \epsilon L \|\mathbf{g}\|_2 .$$

□

We now give an upper bound on the performance of the *projected subgradient descent method*, when applied to Lipschitz convex functions. This algorithm repeatedly iterates the update rule[3]

$$\mathbf{g}_t \in \partial f(\mathbf{x}_t), \ \mathbf{x}_{t+1} \leftarrow \mathbf{\Pi}_{\mathcal{X}} \left(\mathbf{x}_t - \eta \mathbf{g}_t \right), \text{ for all } 0 \leq t < T. \tag{3}$$

Here, $\eta > 0$ is a step size to be chosen, and $\mathbf{\Pi}_{\mathcal{X}}$ is the Euclidean projection to $\mathcal{X}$. Intuitively, (3) uses $\mathbf{g}_t$ as a suggested "descent direction" while remaining inside the feasible region $\mathcal{X}$.

Theorem 2 (Projected gradient descent)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be convex and L-Lipschitz for $\mathcal{X} \subseteq \mathbb{B}(\mathbf{0}_d, R)$. Consider iterating the update (3) for $0 \leq t < T$, from $\mathbf{x}_0 \leftarrow \mathbf{0}_d$ with $\eta = \frac{R}{LT^{1/2}}$, and let $\bar{\mathbf{x}} := \frac{1}{T} \sum_{0 \leq t < T} \mathbf{x}_t$. Then,*

$$f(\bar{\mathbf{x}}) - \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \leq \frac{LR}{\sqrt{T}}.$$

Proof. We begin by rewriting the update rule (3) in a given iteration as

$$\mathbf{x}_{t+1} = \text{argmin}_{\mathbf{x} \in \mathcal{X}} \frac{1}{2} \|\mathbf{x} - (\mathbf{x}_t - \eta \mathbf{g}_t)\|_2^2 = \text{argmin}_{\mathbf{x} \in \mathcal{X}} \langle \eta \mathbf{g}_t, \mathbf{x} \rangle + \frac{1}{2} \|\mathbf{x} - \mathbf{x}_t\|_2^2 .$$

By the first-order optimality condition (Lemma 2, Part I), we thus have for all $\mathbf{u} \in \mathcal{X}$,

$$\langle \eta \mathbf{g}_t + (\mathbf{x}_{t+1} - \mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{u} \rangle \leq 0. \tag{4}$$

By using the identity

$$\langle \mathbf{x}_t - \mathbf{x}_{t+1}, \mathbf{x}_{t+1} - \mathbf{u} \rangle = \frac{1}{2} \|\mathbf{x}_t - \mathbf{u}\|_2^2 - \frac{1}{2} \|\mathbf{x}_{t+1} - \mathbf{u}\|_2^2 - \frac{1}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|_2^2 ,$$

we hence have by rearranging (4), for all $0 \leq t < T$,

$$\begin{aligned}
\langle \eta \mathbf{g}_t, \mathbf{x}_t - \mathbf{u} \rangle &\leq \frac{1}{2} \|\mathbf{x}_t - \mathbf{u}\|_2^2 - \frac{1}{2} \|\mathbf{x}_{t+1} - \mathbf{u}\|_2^2 - \frac{1}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|_2^2 + \langle \eta \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \\
&\leq \frac{1}{2} \|\mathbf{x}_t - \mathbf{u}\|_2^2 - \frac{1}{2} \|\mathbf{x}_{t+1} - \mathbf{u}\|_2^2 - \frac{1}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|_2^2 + \eta \|\mathbf{g}_t\|_2 \|\mathbf{x}_t - \mathbf{x}_{t+1}\|_2 \\
&\leq \frac{1}{2} \|\mathbf{x}_t - \mathbf{u}\|_2^2 - \frac{1}{2} \|\mathbf{x}_{t+1} - \mathbf{u}\|_2^2 + \frac{\eta^2}{2} \|\mathbf{g}_t\|_2^2 \leq \frac{1}{2} \|\mathbf{x}_t - \mathbf{u}\|_2^2 - \frac{1}{2} \|\mathbf{x}_{t+1} - \mathbf{u}\|_2^2 + \frac{\eta^2 L^2}{2}.
\end{aligned}$$

[2] Under certain problem parameterizations, the lower bound in Theorem 1 can actually be improved by a logarithmic factor to match the upper bound from the center of gravity method, see [NY83].

[3] We ignore the issue that $\partial f(\mathbf{x})$ may be undefined on the boundary of $\mathcal{X}$, since for Lipschitz functions, moving infinitesimally into the interior negligibly affects function error.

In the last line, we used the Cauchy-Schwarz and Young's inequalities, as well as the bound in Lemma 5. By summing the above inequality for all iterations $0 \le t < T$ and dividing by ηT, and using the assumption $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$ to lower bound the left-hand side, we have for all $\mathbf{u} \in \mathcal{X}$,

$$\frac{1}{T}\sum_{0\le t<T}(f(\mathbf{x}_t) - f(\mathbf{u})) \le \frac{1}{T}\sum_{0\le t<T}\langle \eta \mathbf{g}_t, \mathbf{x}_t - \mathbf{u}\rangle \le \frac{\|\mathbf{x}_0 - \mathbf{u}\|_2^2}{2\eta T} + \frac{\eta L^2}{2} \le \frac{R^2}{2\eta T} + \frac{\eta L^2}{2}.$$

The conclusion follows by taking u to be a minimizer of f, applying convexity once again (which yields $f(\bar{\mathbf{x}}) \le \frac{1}{T}\sum_{0\le t<T} f(\mathbf{x}_t)$), and using our choice of η. □

Pleasingly, the algorithm in Theorem 2 applies a simple update rule (3), and yet also achieves the optimal rate for Lipschitz convex function minimization (Theorem 1) up to a constant factor, when d is sufficiently larger than $\frac{LR}{\epsilon}$. Combined with the center of gravity method, we have established, up to a logarithmic factor, the complexity of Lipschitz convex optimization in the subgradient oracle model. We can think of the projected subgradient method (3) as preferable in the "low-accuracy" regime where the target accuracy ϵ is not too small, as it gives a rate independent of the dimension d, at the cost of a polynomial overhead in ϵ and other parameters, i.e., L and R. On the other hand, the center of gravity method incurs a d factor overhead, but is preferable in the "high-accuracy" regime of small ϵ due to the logarithmic dependence on ϵ in its optimization rate.

3 Smooth optimization

Theorem 2 is encouraging because of its optimality in the high-dimensional regime (despite its relative simplicity), but begs the question: can we do better assuming more structure on f? Here, we explore the utility of an additional structural assumption known in the literature as smoothness.

Definition 3 (Smoothness)**.** *We say $f : \mathbb{R}^d \to \mathbb{R}$ is* L-smooth *with respect to the norm $\|\cdot\|$, or L-smooth in $\|\cdot\|$, if f is differentiable, and ∇f is L-Lipschitz, in the sense that*[4]

$$\|\nabla f(\mathbf{x}') - \nabla f(\mathbf{x})\|_* \le L\|\mathbf{x} - \mathbf{x}'\| \text{ for all } \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d.$$

If $\|\cdot\| = \|\cdot\|_* = \|\cdot\|_2$, we simply say f is L-smooth.

To provide some motivation for Definition 3, imagine an infinitesimal version of the update rule (4) (with $\mathcal{X} = \mathbb{R}^d$), where we have a particle $\mathbf{x}_t \in \mathbb{R}^d$ evolving via the ordinary differential equation (ODE) $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}_t = \mathbf{v}_t$, for a velocity vector $\mathbf{v}_t$.[5] An application of the chain rule shows that

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{x}_t) = \langle \nabla f(\mathbf{x}_t), \mathbf{v}_t\rangle = -\|\nabla f(\mathbf{x}_t)\|_2^2 \text{ if } \mathbf{v}_t = -\nabla f(\mathbf{x}_t), \tag{5}$$

suggesting that $\mathbf{v}_t \leftarrow -\nabla f(\mathbf{x}_t)$ is the direction of steepest descent if our aim is to decrease function value. The ODE $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}_t = -\nabla f(\mathbf{x}_t)$ is known as the *gradient flow* method. However, such ODE-based update rules are not implementable in discrete time. Instead, we may use a *forward Euler discretization* which, for a time interval $t \in [t_0, t_0 + \eta]$ replaces $\mathbf{v}_t = \nabla f(\mathbf{x}_t)$ with $\mathbf{v}_t = \nabla f(\mathbf{x}_{t_0})$. It is ideal that our discretization closely approximates the ideal gradient flow process, which is true if $\nabla f(\mathbf{x})$ is stable in small regions. Definition 3 is a natural way to formalize this stability condition.

Next, recall that Definition 1 is a statement about zeroth-order stability of a function f, but implies bounds on first-order information on f through Lemma 5. Analogously, the first-order stability condition in Definition 3 yields bounds on second-order information.

Lemma 6. *If $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable and convex, then f is L-smooth iff for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$,*

$$f(\mathbf{x}') \le f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{x}' - \mathbf{x}\|_2^2. \tag{6}$$

If f is twice-differentiable (and possibly nonconvex), f is L-smooth iff $|\nabla^2 f(\mathbf{x})[\mathbf{v}, \mathbf{v}]| \le L\|\mathbf{v}\|_2^2$ for all $\mathbf{x}, \mathbf{v} \in \mathbb{R}^d$.

[4] See Definition 6 for the definition of the dual norm $\|\cdot\|_*$.

[5] We sometimes use $\mathbf{x}_t$ to denote a sequence of particles indexed by time t, as shorthand for a function $\mathbf{x}(t)$ which takes $\mathbf{x} : \mathbb{R}_{\ge 0} \to \mathbb{R}^d$. More generally, for a function Φ indexed by a variable t, we may identify $\Phi_t \equiv \Phi(t)$.

Proof. We begin with the first claim. Let f be L-smooth. By the fundamental theorem of calculus,

$$f(\mathbf{x}') = f(\mathbf{x}) + \int_0^1 \langle \nabla f(\mathbf{x}_\lambda), \mathbf{x}' - \mathbf{x} \rangle \, \mathrm{d}\lambda, \text{ where } \mathbf{x}_\lambda := (1-\lambda)\mathbf{x} + \lambda \mathbf{x}'.$$

This follows by evaluating $g(\lambda) := f(\mathbf{x}_\lambda)$ at the endpoints of $\lambda \in [0,1]$, since $g'(\lambda) = \langle \nabla f(\mathbf{x}_\lambda), \mathbf{x}' - \mathbf{x} \rangle$. By the Cauchy-Schwarz inequality and the smoothness assumption, we have the desired

$$\begin{aligned} f(\mathbf{x}') - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle &= \int_0^1 \langle \nabla f(\mathbf{x}_\lambda) - \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle \, \mathrm{d}\lambda \\ &\le \int_0^1 \left\| \nabla f(\mathbf{x}_\lambda) - \nabla f(\mathbf{x}) \right\|_2 \left\| \mathbf{x}' - \mathbf{x} \right\|_2 \mathrm{d}\lambda \\ &\le \int_0^1 L\lambda \left\| \mathbf{x}' - \mathbf{x} \right\|_2^2 \mathrm{d}\lambda = \frac{L}{2} \left\| \mathbf{x}' - \mathbf{x} \right\|_2^2. \end{aligned} \tag{7}$$

We can also check that the same proof lower bounds $f(\mathbf{x}') - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle$ by $-\frac{L}{2} \left\| \mathbf{x}' - \mathbf{x} \right\|_2^2$, regardless of convexity. Conversely, suppose (6) holds, and let $\phi(\mathbf{x}) := f(\mathbf{x}) - f(\bar{\mathbf{x}}) - \langle \nabla f(\bar{\mathbf{x}}), \mathbf{x} - \bar{\mathbf{x}} \rangle$ for all $\mathbf{x} \in \mathbb{R}^d$, and a fixed $\bar{\mathbf{x}}$. We can check that (6) still holds for ϕ, because it differs from f by a linear term which cancels out in the equation. Therefore,

$$\begin{aligned} 0 = \phi(\bar{\mathbf{x}}) &= \min_{\mathbf{x}' \in \mathbb{R}^d} \phi(\mathbf{x}') \le \min_{\mathbf{x}' \in \mathbb{R}^d} \phi(\mathbf{x}) + \langle \nabla \phi(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle + \frac{L}{2} \left\| \mathbf{x}' - \mathbf{x} \right\|_2^2 \\ &= \phi(\mathbf{x}) - \frac{1}{2L} \left\| \nabla \phi(\mathbf{x}) \right\|_2^2 = f(\mathbf{x}) - f(\bar{\mathbf{x}}) - \langle \nabla f(\bar{\mathbf{x}}), \mathbf{x} - \bar{\mathbf{x}} \rangle - \frac{1}{2L} \left\| \nabla f(\mathbf{x}) - \nabla f(\bar{\mathbf{x}}) \right\|_2^2. \end{aligned}$$

The first line used that convexity shows $\phi \ge 0$ pointwise, and also applied (6) for ϕ; the second line directly minimized the quadratic.[6] Summing this equation with itself with $\mathbf{x}, \bar{\mathbf{x}}$ interchanged, and applying the Cauchy-Schwarz inequality, we have established smoothness as desired:

$$\begin{aligned} \frac{1}{L} \left\| \nabla f(\bar{\mathbf{x}}) - \nabla f(\mathbf{x}) \right\|_2^2 \le \langle \nabla f(\bar{\mathbf{x}}) - \nabla f(\mathbf{x}), \bar{\mathbf{x}} - \mathbf{x} \rangle &\le \left\| \nabla f(\bar{\mathbf{x}}) - \nabla f(\mathbf{x}) \right\|_2 \left\| \bar{\mathbf{x}} - \mathbf{x} \right\|_2 \\ \implies \left\| \nabla f(\bar{\mathbf{x}}) - \nabla f(\mathbf{x}) \right\|_2 &\le L \left\| \bar{\mathbf{x}} - \mathbf{x} \right\|_2. \end{aligned}$$

Next, let f be twice-differentiable, and suppose f is L-smooth. By adding (6) with the same equation where $\mathbf{x}$ and $\mathbf{x}'$ are reversed, we see that for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$,

$$\langle \nabla f(\mathbf{x}') - \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle \le L \left\| \mathbf{x}' - \mathbf{x} \right\|_2^2.$$

Letting $\mathbf{x}' \leftarrow \mathbf{x} + t\mathbf{v}$ and taking a limit as $t \leftarrow 0$, we can rewrite the left-hand side as $t^2 \nabla^2 f(\mathbf{x})[\mathbf{v}, \mathbf{v}]$ and the right-hand side as $Lt^2 \left\| \mathbf{v} \right\|_2^2$, and dividing by t^2 proves the desired $\nabla^2 f(\mathbf{x})[\mathbf{v}, \mathbf{v}] \le L \left\| \mathbf{v} \right\|_2^2$. An analogous argument using the lower bound then establishes $\nabla^2 f(\mathbf{x})[\mathbf{v}, \mathbf{v}] \ge -L \left\| \mathbf{v} \right\|_2^2$.

Finally, suppose $|\nabla^2 f(\mathbf{x})[\mathbf{v}, \mathbf{v}]| \le L \left\| \mathbf{v} \right\|_2^2$ for all $\mathbf{v}$, so that all eigenvalues of $\nabla^2 f(\mathbf{x})$ are in $[\pm L]$. Then, again letting $\mathbf{x}_\lambda := (1-\lambda)\mathbf{x} + \lambda \mathbf{x}'$ for $\lambda \in [0,1]$, the fundamental theorem of calculus gives

$$\begin{aligned} \left\| \nabla f(\mathbf{x}') - \nabla f(\mathbf{x}) \right\|_2 = \left\| \int_0^1 \nabla^2 f(\mathbf{x}_\lambda)(\mathbf{x}' - \mathbf{x}) \mathrm{d}\lambda \right\|_2 &\le \int_0^1 \left\| \nabla^2 f(\mathbf{x}_\lambda)(\mathbf{x}' - \mathbf{x}) \right\|_2 \mathrm{d}\lambda \\ &\le \left\| \mathbf{x}' - \mathbf{x} \right\|_2 \int_0^1 \left\| \nabla^2 f(\mathbf{x}_\lambda) \right\|_{\mathrm{op}} \mathrm{d}\lambda \le L \left\| \mathbf{x}' - \mathbf{x} \right\|_2, \end{aligned}$$

as desired. The first inequality applied the triangle inequality, and the last used that the operator norm of a symmetric matrix is the largest magnitude of any of its eigenvalues. □

Recall that convexity of f yields a lower bound on f by a linear function everywhere, using a first-order approximation centered at any point $\mathbf{x}$ (Lemma 1, Part I). Similarly, (6) can be interpreted as upper bounding f everywhere using a quadratic centered at $\mathbf{x}$. Upper bounds are very useful in the design of optimization algorithms, since they can be used to certify progress on function value. Moreover, because the form of our upper bound (6) is simple, it can be directly optimized. Both of these ideas are captured formally by the following claim.

[6] For this calculation in more detail, see the proof of Corollary 2.

Corollary 2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth. Then for any $\mathbf{x} \in \mathbb{R}^d$, letting $\mathbf{x}' \leftarrow \mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})$,*

$$f(\mathbf{x}') \le f(\mathbf{x}) - \frac{1}{2L}\left\|\nabla f(\mathbf{x})\right\|_2^2.$$

Proof. Consider minimizing the right-hand side of (6) in $\mathbf{x}'$. This is a convex function, so first-order optimality requires $\nabla f(\mathbf{x}) + L(\mathbf{x}' - \mathbf{x}) = \mathbf{0}_d \iff \mathbf{x}' = \mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})$, motivating our update rule. Directly computing the right-hand side of (6) then yields the claim:

$$\begin{aligned} f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x}\rangle + \frac{L}{2}\left\|\mathbf{x}' - \mathbf{x}\right\|_2^2 &= f(\mathbf{x}) - \frac{1}{L}\left\|\nabla f(\mathbf{x})\right\|_2^2 + \frac{1}{2L}\left\|\nabla f(\mathbf{x})\right\|_2^2 \\ &= f(\mathbf{x}) - \frac{1}{2L}\left\|\nabla f(\mathbf{x})\right\|_2^2. \end{aligned}$$

□

In the language of our earlier discussion of gradient flow (5), Corollary 2 shows that for smooth functions, following an Euler discretization $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}_t = -\nabla f(\mathbf{x}_{t_0})$ for a time interval $\eta = \frac{1}{L}$ makes comparable function progress to ideal gradient flow. Indeed, letting $\mathbf{x} := \mathbf{x}_{t_0}$ and $\mathbf{x}' := \mathbf{x}_{t_0+\eta}$ under this discretization scheme, following gradient flow would have yielded

$$f(\mathbf{x}') - f(\mathbf{x}) = -\int_{t_0}^{t_0+\eta} \left\|\nabla f(\mathbf{x}_t)\right\|_2^2 \mathrm{d}t,$$

and instead Corollary 2 shows our discrete update yields progress $-\frac{1}{2L}\left\|\nabla f(\mathbf{x})\right\|_2^2$.

Even without convexity, Corollary 2 yields strong guarantees on progress towards local optimality.

Lemma 7. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth, let $\epsilon > 0$, and suppose for $\mathbf{x}_0 \in \mathbb{R}^d$ we have $f(\mathbf{x}_0) - \min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \le \Delta$. Then iterating $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)$ for $0 \le t < T$ where $T \ge \frac{\Delta}{\epsilon^2}$,*

$$\min_{0\le t<T} \left\|\nabla f(\mathbf{x}_t)\right\|_2 \le \epsilon.$$

Proof. Suppose for contradiction that the conclusion is false. Then, telescoping Corollary 2 gives

$$f(\mathbf{x}_T) \le f(\mathbf{x}_0) - \frac{1}{2L}\sum_{t=0}^{T-1} \left\|\nabla f(\mathbf{x}_t)\right\|_2^2 \le f(\mathbf{x}_0) - T\epsilon^2 \le f(\mathbf{x}_0) - \Delta.$$

This is a contradiction, as $f(\mathbf{x}_0) - f(\mathbf{x}_T) \le \Delta$ by assumption. □

In many nonconvex optimization settings, a natural goal is to find an *ϵ-stationary point*, i.e., a point $\mathbf{x}$ where $\left\|\nabla f(\mathbf{x})\right\|_2 \le \epsilon$. The design of stationary point-finding algorithms is often complemented by a line of research which aims to establish that local minima have desirable global optimality properties, see e.g., [GLM16] for a famous example. In the smooth case, Lemma 7 in fact yields optimal guarantees for finding approximate stationary points, as shown by [CDHS20]. We next show how to combine smoothness with convexity to give stronger global convergence guarantees.

Theorem 3 (Smooth gradient descent)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and convex, and suppose for $\mathbf{x}_0 \in \mathbb{R}^d$ we have $\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|_2 \le R$ for $\mathbf{x}^\star \in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. Then iterating $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)$ for $0 \le t < T$,*

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \le \frac{2LR^2}{T}.$$

Proof. Throughout the proof, let $\Phi_t := f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ for all $0 \le t \le T$, and note that $\left\|\mathbf{x}_t - \mathbf{x}^\star\right\|_2 \le R$ for all $0 \le t \le T$ by Lemma 8. By convexity and the Cauchy-Schwarz inequality,

$$\Phi_t \le \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star\rangle \le \left\|\nabla f(\mathbf{x}_t)\right\|_2 R,$$

so combined with Corollary 2 (which shows $\Phi_t \ge \Phi_{t+1}$), we have

$$\Phi_{t+1} \le \Phi_t - \frac{1}{2L}\left(\frac{\Phi_t}{R}\right)^2 \implies \frac{1}{2LR^2} \le \frac{\Phi_t}{2LR^2\Phi_{t+1}} \le \frac{1}{\Phi_{t+1}} - \frac{1}{\Phi_t}.$$

Telescoping for T iterations shows $\frac{T}{2LR^2} \le \frac{1}{\Phi_T}$, or $\Phi_T \le \frac{2LR^2}{T}$ as claimed. □

In the proof of Theorem 3, we used the following contractivity claim.

Lemma 8. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and convex, and let $\mathbf{x}' \leftarrow \mathbf{x} - \eta \nabla f(\mathbf{x})$ for $\eta \leq \frac{1}{L}$. Then for $\mathbf{x}^\star \in \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$, $\|\mathbf{x}' - \mathbf{x}^\star\|_2 \leq \|\mathbf{x} - \mathbf{x}^\star\|_2$.*

Proof. Expanding both sides of $\|\mathbf{x}' - \mathbf{x}^\star\|_2^2 \leq \|\mathbf{x} - \mathbf{x}^\star\|_2^2$, it suffices to establish

$$\|\mathbf{x} - \mathbf{x}^\star\|_2^2 - \|(\mathbf{x} - \eta \nabla f(\mathbf{x})) - \mathbf{x}^\star\|_2^2 \geq 0 \iff 2\eta \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star \rangle \geq \eta^2 \|\nabla f(\mathbf{x})\|_2^2 .$$

The latter expression follows from Corollary 2 since $f(\mathbf{x}') \geq f(\mathbf{x}^\star)$, so convexity gives

$$\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star \rangle \geq f(\mathbf{x}) - f(\mathbf{x}^\star) \geq f(\mathbf{x}) - f(\mathbf{x}') \geq \frac{1}{2L} \|\nabla f(\mathbf{x})\|_2^2 \geq \frac{\eta}{2} \|\nabla f(\mathbf{x})\|_2^2 .$$

□

Interestingly, unlike in the Lipschitz convex setting, the gradient descent algorithm in Theorem 3 is provably suboptimal. We expand on this point in the next section, where we prove a lower bound on the performance of gradient methods in the smooth convex setting via a reduction.

4 Well-conditioned optimization

In this section, we begin to explore the question: when can we achieve dimension-free *linear convergence rates*, i.e. algorithms for minimizing f which guarantee for some C, that $f(\mathbf{x}_t) - \min_{\mathbf{x}} f(\mathbf{x}) \leq \exp(-Ct)$? Here, t should be thought of as an iteration counter, and error rates of the form $\exp(-Ct)$ imply that to achieve ϵ error, we must take $t \propto \log \frac{1}{\epsilon}$. Linear convergence rates are well-suited for the high-accuracy regime, where the overhead of achieving a very small error ϵ is only logarithmic. Note that the cutting-plane methods in Part I achieve linear convergence rates with $\mathrm{poly}(d)$ overhead, whereas the results of this lecture thus far (i.e., Theorems 2, 3) give dimension-free rates, but with polynomial overhead in the inverse target accuracy.

To motivate the structural assumption which we introduce in this section, we revisit the conceptual gradient flow algorithm from Section 3. Recall from (5) that if $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}_t = -\nabla f(\mathbf{x}_t)$, then $\frac{\mathrm{d}}{\mathrm{d}t}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) = \frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{x}_t) = -\|\nabla f(\mathbf{x}_t)\|_2^2$, where $\mathbf{x}^\star \in \operatorname{argmin}_{\mathbf{x}} f(\mathbf{x})$. Combined with Fact 1, this suggests a structural assumption which would yield linear convergence rates: that of the form

$$\|\nabla f(\mathbf{x})\|_2^2 \geq C(f(\mathbf{x}) - f(\mathbf{x}^\star)). \tag{8}$$

Fact 1 (Grönwall's inequality)**.** *If $\Phi_0 \geq 0$ and $\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t \leq -C\Phi_t$ for all $t \geq 0$, $\Phi_t \leq \exp(-Ct)\Phi_0$.*

Indeed, putting together (5), (8), and Fact 1 immediately implies a linear convergence rate on the optimization error of gradient flow. Another natural way to see this is to show that the potential function $V(t) := \exp(Ct)(f(\mathbf{x}_t) - f(\mathbf{x}^\star))$ is monotone, since

$$\frac{\mathrm{d}}{\mathrm{d}t} V(t) = \exp(Ct)(C\left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) - \|\nabla f(\mathbf{x}_t)\|_2^2) \leq 0,$$

using (8). Arguments of this form (where error bounds are proven via monotonicity of a potential function) are sometimes called Lyapunov arguments, and the potential $V(t)$ is called a Lyapunov function. These arguments can be used in continuous or discrete time, although sometimes significant ingenuity is required to design the correct Lyapunov function. What is left to do is to give a sufficient condition for bounds of the form (8), and to analyze a corresponding discrete-time algorithm. Fortunately, the following definition is well-suited for both of these modifications.

Definition 4 (Strong convexity)**.** *We say $f : \mathcal{X} \to \mathbb{R}$ is μ-strongly convex with respect to the norm $\|\cdot\|$, or μ-strongly convex in $\|\cdot\|$, if for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$,*

$$f((1-\lambda)\mathbf{x} + \lambda\mathbf{x}') \leq (1-\lambda)f(\mathbf{x}) + \lambda f(\mathbf{x}') - \frac{\mu\lambda(1-\lambda)}{2} \|\mathbf{x} - \mathbf{x}'\|^2 .$$

If $\|\cdot\| = \|\cdot\|_2$, we simply say f is μ-strongly convex.

Notice that Definition 4 recovers our usual notion of convexity when $\mu = 0$. Intuitively, it enforces that f is not only overestimated by linear combinations along line segments; it is overestimated by at least a quadratic factor. We begin by demonstrating several consequences of strong convexity.

Lemma 9. *If $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable, then f is μ-strongly convex iff for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$,*

$$f(\mathbf{x}') \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{x}' - \mathbf{x}\|_2^2 . \tag{9}$$

If f is twice-differentiable, f is μ-strongly convex iff $\nabla^2 f(\mathbf{x})[\mathbf{v}, \mathbf{v}] \geq \mu \|\mathbf{v}\|_2^2$ for all $\mathbf{x}, \mathbf{v} \in \mathbb{R}^d$.

Proof. For the first claim, suppose (9) holds. Then letting $\mathbf{x}_\lambda := (1-\lambda)\mathbf{x} + \lambda \mathbf{x}'$ for $\lambda \in [0, 1]$, μ-strong convexity follows by combining a $1 - \lambda$ multiple of the first equation below with a λ multiple of the second, since $\mathbf{x} - \mathbf{x}_\lambda = \lambda(\mathbf{x} - \mathbf{x}')$ and $\mathbf{x}' - \mathbf{x}_\lambda = (1-\lambda)(\mathbf{x}' - \mathbf{x})$:

$$f(\mathbf{x}) \geq f(\mathbf{x}_\lambda) + \langle \nabla f(\mathbf{x}_\lambda), \mathbf{x} - \mathbf{x}_\lambda \rangle + \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}_\lambda\|_2^2 ,$$
$$f(\mathbf{x}') \geq f(\mathbf{x}_\lambda) + \langle \nabla f(\mathbf{x}_\lambda), \mathbf{x}' - \mathbf{x}_\lambda \rangle + \frac{\mu}{2} \|\mathbf{x}' - \mathbf{x}_\lambda\|_2^2 .$$

Next, suppose f is μ-strongly convex. Then defining $\mathbf{x}_\lambda$ as above, and applying the definition of strong convexity, we have for any $\lambda \in [0, 1]$ that

$$\begin{aligned} f(\mathbf{x}') &\geq \frac{f(\mathbf{x}_\lambda) - (1-\lambda) f(\mathbf{x}) + \frac{\mu\lambda(1-\lambda)}{2} \|\mathbf{x} - \mathbf{x}'\|_2^2}{\lambda} \\ &= f(\mathbf{x}) + \frac{f(\mathbf{x}_\lambda) - f(\mathbf{x})}{\lambda} + \frac{\mu(1-\lambda)}{2} \|\mathbf{x} - \mathbf{x}'\|_2^2 . \end{aligned}$$

Taking $\lambda \to 0$ yields (9). Next, assuming the second-order lower bound and recalling (7),

$$\begin{aligned} f(\mathbf{x}') &= f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle + \int_0^1 \langle \nabla f(\mathbf{x}_\lambda) - \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle \, \mathrm{d}\lambda \\ &= f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle + \int_0^1 \left(\int_0^\lambda \nabla^2 f(\mathbf{x}_{\lambda'})[\mathbf{x}' - \mathbf{x}, \mathbf{x}' - \mathbf{x}] \mathrm{d}\lambda' \right) \mathrm{d}\lambda \\ &\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle + \int_0^1 \lambda\mu \|\mathbf{x}' - \mathbf{x}\|_2^2 \, \mathrm{d}\lambda = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{x}' - \mathbf{x}\|_2^2 . \end{aligned}$$

As we have established, this implies f is μ-strongly convex. Conversely, μ-strong convexity of f implies the second-order lower bound by an analogous argument to Lemma 6. ☐

Remark 2. *For convex, non-differentiable $f : \mathbb{R}^d \to \mathbb{R}$, μ-strong convexity is equivalent to*

$$f(\mathbf{x}') \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{x}' - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{x}' - \mathbf{x}\|_2^2 \text{ for all } \mathbf{g} \in \partial f(\mathbf{x}).$$

By taking $\mu \to 0$, Lemma 9 gives a familiar second-order characterization of (standard) convexity of a function $f : \mathbb{R}^d \to \mathbb{R}$, when f is twice-differentiable, which is that $\nabla^2 f \succeq \mathbf{0}_d$ pointwise. Symmetrically to Corollary 2, we next show that (9) indeed implies a bound of the form in (8).

Corollary 3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex, and let $\mathbf{x}^\star := \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$.*[7] *Then*

$$f(\mathbf{x}) - f(\mathbf{x}^\star) \leq \frac{1}{2\mu} \|\nabla f(\mathbf{x})\|_2^2 .$$

Proof. The minimum of the left-hand side of (9) over $\mathbf{x}'$ is at least the minimum of the right-hand side over $\mathbf{x}'$, which is achieved when $\nabla f(\mathbf{x}) + \mu(\mathbf{x}' - \mathbf{x}) = \mathbf{0}_d$. For this optimal $\mathbf{x}' = \mathbf{x} - \frac{1}{\mu} \nabla f(\mathbf{x})$, the right-hand side has value $f(\mathbf{x}) - \frac{1}{2\mu} \|\nabla f(\mathbf{x})\|_2^2$, and rearranging yields the conclusion. ☐

Corollary 3 shows that for μ-strongly convex functions, the bound (8) holds with $C = 2\mu$. Combined with Fact 1, this immediately yields a linear convergence rate on the error achieved by gradient flow, as previously discussed. It is not difficult to show that for smooth, strongly convex functions, this linear convergence rate continues to hold in discrete time.

[7] Strongly convex functions are also strictly convex, so $\mathbf{x}^\star$ is unique (Lemma 3, Part I).

Theorem 4 (Well-conditioned gradient descent)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and μ-strongly convex, and let $\mathbf{x}^\star := \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ and $\kappa := \frac{L}{\mu} \geq 1$.*[8] *Then iterating $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)$ for $0 \leq t < T$,*

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \left(1 - \frac{1}{\kappa}\right)^T (f(\mathbf{x}_0) - f(\mathbf{x}^\star))$$

Proof. We claim that for all $0 \leq t < T$,

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) \leq \left(1 - \frac{1}{\kappa}\right)(f(\mathbf{x}_t) - f(\mathbf{x}^\star)), \tag{10}$$

which inductively implied proves the claim. Indeed, (10) follows from Corollaries 2 and 3. □

Remark 3. *Interestingly, Theorem 4 only used strong convexity of f through the consequence in Corollary 3, which can hold for non-convex functions as well. In general, bounds of the form (8) are referred to in the literature as* gradient domination *or* Polyak-Łojasiewicz *conditions, and have been used to establish convergence rates for training neural networks [XLS17, HM17]. Gradient domination further implies quadratic growth bounds of the form (9) around $\mathbf{x} \leftarrow \mathbf{x}^\star$, which can be proven by tracking an appropriate Lyapunov function. For more details, see [KNS16].*

We call the setting in Theorem 4, where f is both smooth and strongly convex, the *well-conditioned* regime, and κ is called the condition number of f. Note that when $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x}$ for $\mathbf{A} \in \mathbb{S}^{d\times d}_{\succeq \mathbf{0}}$, the condition number κ is the ratio of the largest and smallest eigenvalues of $\mathbf{A}$. Theorem 4 shows that for well-conditioned functions with $\kappa \ll d$, gradient descent improves upon the convergence rate of the center of gravity method from Part I. Indeed, Theorem 4 gives a dimension-free linear convergence rate in the well-conditioned regime, as advertised at the beginning of the section.

Finally, we conclude the section with a discussion of lower bounds in the well-conditioned regime. To construct our hard instance, we take a brief detour and define a graph-theoretic object.

Definition 5 (Laplacian)**.** *Let $G = (V, E, \mathbf{w})$ be an undirected graph with vertices V, edges E, and edge weights $\mathbf{w} \in \mathbb{R}^E_{>0}$. Let $\mathbf{D}_G \in \mathbb{R}^{V\times V}$ be the diagonal degree matrix with $[\mathbf{D}_G]_{vv} = \sum_{e=(u,v)\in E} \mathbf{w}_e$ for all $v \in V$, and let $\mathbf{A} \in \mathbb{R}^{V\times V}$ be the weighted adjacency matrix where for each $e = (u, v) \in E$, we have $\mathbf{A}_{uv} = \mathbf{A}_{vu} = \mathbf{w}_e$. Then we define the* Laplacian matrix *of G by $\mathbf{L}_G := \mathbf{D}_G - \mathbf{A}_G$.*

Lemma 10. *Let $G = (V, E, \mathbf{w})$ be an undirected graph. Then for any $\mathbf{x} \in \mathbb{R}^V$,*

$$\mathbf{x}^\top\mathbf{L}_G\mathbf{x} = \sum_{e=(u,v)\in E} \mathbf{w}_e(\mathbf{x}_u - \mathbf{x}_v)^2.$$

Proof. Let $\mathbf{b}_{u,v} := \mathbf{e}_u - \mathbf{e}_v \in \{-1, 0, 1\}^V$ be a 2-sparse vector, for each edge $(u, v) \in E$.[9] It is straightforward to verify, by decomposing edgewise, that $\mathbf{D}_G - \mathbf{A}_G = \sum_{e=(u,v)\in E} \mathbf{w}_e\mathbf{b}_{u,v}\mathbf{b}_{u,v}^\top$. The conclusion follows since for all $\mathbf{x} \in \mathbb{R}^V$, we have $\langle\mathbf{b}_{u,v}, \mathbf{x}\rangle^2 = (\mathbf{x}_u - \mathbf{x}_v)^2$. □

Our well-conditioned hard instance is based off of a similar construction as (2), where each gradient query can only reveal one coordinate to the algorithm. To ensure smoothness and convexity, instead of a max-type function, our well-conditioned construction is based on the Laplacian of a path graph. The structure of the path obstructs more than one coordinate from being revealed per iteration.

Theorem 5 (Well-conditioned lower bound)**.** *Let $\kappa \geq 1$ and $\epsilon \in (0, 1)$. No algorithm $\mathcal{A}$ which accesses a target L-smooth, μ-strongly convex function f with $\kappa = \frac{L}{\mu}$ using a gradient oracle $\mathcal{O}$ can optimize f to additive error $\epsilon(f(\mathbf{x}_0) - \min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}))$ using $T < \frac{\sqrt{\kappa}-1}{2}\log(\frac{1}{\kappa\epsilon})$ queries, subject to the restriction (1).*

Proof. Let $G = (V, E, \mathbf{w})$ be an unweighted path graph (i.e., $\mathbf{w} = \mathbf{1}_E$), on d vertices labeled by $[d] \equiv V$, and with $d - 1$ edges $E = \{(i, i + 1)\}_{i\in[d-1]}$. Consider the function

$$f(\mathbf{x}) := \frac{\kappa - 1}{8}(\mathbf{x}^\top\mathbf{L}_G\mathbf{x} - 2\langle\mathbf{e}_1, \mathbf{x}\rangle) + \frac{1}{2}\|\mathbf{x}\|_2^2.$$

[8] By comparing (6) and (9), it is clear that $L < \mu$ is a contradiction.

[9] Because G is undirected, we can arbitrarily choose an orientation (u, v) or (v, u) for each edge in this proof.

By the characterizations in Lemmas 9 and 10, it is clear that f is 1-strongly convex (as $\nabla^2 f(\mathbf{x})[\mathbf{v},\mathbf{v}] \geq \nabla^2(\frac{\mu}{2}\|\mathbf{x}\|_2^2)[\mathbf{v},\mathbf{v}] = \mu\|\mathbf{v}\|_2^2$). Moreover, we claim that $\mathbf{L}_G \preceq 4\mathbf{I}_d$, which implies that f is κ-smooth by Lemma 6. To see this, applying Lemma 10 with our specific graph shows for any $\mathbf{v} \in \mathbb{R}^d$,

$$\mathbf{L}_G[\mathbf{v},\mathbf{v}] = \sum_{i\in[d-1]} (\mathbf{v}_i - \mathbf{v}_{i+1})^2 \leq \sum_{i\in[d-1]} (2\mathbf{v}_i^2 + 2\mathbf{v}_{i+1}^2) \leq 4\|\mathbf{v}\|_2^2. \tag{11}$$

Now, observe that if $\mathbf{x}$ is supported on the first $t < d$ coordinates, then both $\mathbf{x}$ and

$$\mathbf{L}_G\mathbf{x} = \sum_{i\in[d-1]} (\mathbf{x}_i - \mathbf{x}_{i+1})(\mathbf{e}_i - \mathbf{e}_{i+1})$$

are supported on the first $t+1$ coordinates. Hence, under the assumption (1), the iterates of $\mathcal{A}$ are supported on the first T coordinates, since $\nabla f(\mathbf{x})$ linearly combines $\mathbf{L}_G\mathbf{x}$, $\mathbf{e}_1$, and $\mathbf{x}$. Therefore, for any iterate $\mathbf{x}$ of $\mathcal{A}$, letting $\mathbf{x}^\star := \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$, strong convexity implies (via Lemma 9)

$$f(\mathbf{x}) - f(\mathbf{x}^\star) \geq \frac{1}{2}\|\mathbf{x} - \mathbf{x}^\star\|_2^2 \geq \frac{1}{2}\sum_{i\in[d]\setminus[T]} [\mathbf{x}^\star]_i^2. \tag{12}$$

Moreover, by first-order optimality, we have that $\frac{\kappa-1}{4}(\mathbf{L}_G\mathbf{x}^\star - \mathbf{e}_1) + \mathbf{x}^\star = \mathbf{0}_d$, so (11) shows

$$\mathbf{x}_{i-1}^\star - \frac{2(\kappa-1)+4}{\kappa-1}\mathbf{x}_i^\star + \mathbf{x}_{i+1}^\star \text{ for all } i \in [d-1],$$

where we denote $\mathbf{x}_0^\star := 1$. It is straightforward to check that

$$\mathbf{x}_i^\star = \left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^i \text{ for all } i \in [d]$$

is a solution, where we use that $r = 1 - 2(\sqrt{\kappa}-1)^{-1}$ is a root of the quadratic, $r^2 - \frac{2(\kappa+1)}{\kappa-1}r + 1 = 0$. Finally, by taking $d \to \infty$, the right-hand side of (12) is bounded by the claimed quantity, since

$$\begin{aligned}\frac{1}{2}\sum_{i\in[d]\setminus[T]} [\mathbf{x}_i^\star]^2 &=_{d\to\infty} \left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^T \left(\frac{1}{2}\sum_{i\in[d]} [\mathbf{x}_i^\star]^2\right) \\ &\geq \frac{1}{\kappa}\left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^T (f(\mathbf{x}_0) - f(\mathbf{x}^\star)) \geq \epsilon\,(f(\mathbf{x}_0) - f(\mathbf{x}^\star)).\end{aligned}$$

□

For $\kappa \gg 1$ and $\epsilon \ll \frac{1}{\kappa}$, Theorem 5 establishes that the iteration count required to achieve an ϵ-factor decrease in the function error scales as $T = \Omega(\sqrt{\kappa}\log\frac{1}{\epsilon})$. Notice that Theorem 4 implies an upper bound of $T = O(\kappa\log\frac{1}{\epsilon})$, which is suboptimal in this regime. In a later lecture, we will see how to more carefully combine two iterate sequences to design an algorithm matching the lower bound of Theorem 5. This phenomenon of using multiple iterates (which induce certain history-aware update rules viewable as adding "momentum" to our updates) to obtain faster algorithms is often called "acceleration" or "accelerated gradient descent," and was first discovered by [Nes83]. We will give a proof-of-concept sketch that acceleration is possible in Part IV, improving upon Theorem 4.

Finally, we give a reduction-based argument which shows that in the L-smooth, convex setting of Theorem 3, the best possible additive error rate scales as $\Omega(\frac{L}{T^2})$.

Lemma 11. *Suppose that there is an algorithm $\mathcal{A}$ which takes as input $\mathbf{x}_0 \in \mathbb{R}^d$ and an L-smooth, convex $f : \mathbb{R}^d \to \mathbb{R}$ (accessed through a gradient oracle) with $\mathbf{x}^\star \in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$, and produces a point $\mathbf{x} \in \mathbb{R}^d$ using T queries such that, for some constants $C, c > 0$,*

$$f(\mathbf{x}) - f(\mathbf{x}^\star) \leq \frac{CL\|\mathbf{x}_0 - \mathbf{x}^\star\|_2^2}{T^c}.$$

Then there is an algorithm $\mathcal{A}'$ which takes as input $\mathbf{x}_0 \in \mathbb{R}^d$ and an L-smooth, μ-strongly convex $f : \mathbb{R}^d \to \mathbb{R}$ (accessed through a gradient oracle) with $\mathbf{x}^\star \in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ and $\kappa := \frac{L}{\mu}$, and produces a point $\mathbf{x} \in \mathbb{R}^d$ such that $f(\mathbf{x}) - f(\mathbf{x}^\star) \leq \epsilon(f(\mathbf{x}_0) - f(\mathbf{x}^\star))$ in $O(\kappa^{1/c}\log\frac{1}{\epsilon})$ queries.

Proof. We will design a subroutine which produces $\mathbf{x}$ satisfying $f(\mathbf{x}) - f(\mathbf{x}^\star) \le \frac{1}{2}(f(\mathbf{x}_0) - f(\mathbf{x}^\star))$ in $O(\kappa^{1/c})$ iterations, which implies the claim upon recursing. Indeed, applying $\mathcal{A}$ with $T \ge (2C\kappa)^{1/c}$ implements this subroutine, since Lemma 9 shows

$$f(\mathbf{x}) - f(\mathbf{x}^\star) \le \frac{CL \left\|\mathbf{x}_0 - x^\star\right\|_2^2}{T^c} = \frac{\mu}{4} \left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|_2^2 \le \frac{1}{2}(f(\mathbf{x}_0) - f(\mathbf{x}^\star)).$$

□

Note that Theorem 3 gives an algorithm $\mathcal{A}$ achieving $c = 1$ in the statement of Lemma 11, and if $c > 2$ were achievable, then this would result in a contradiction to the lower bound of Theorem 5.

Remark 4. *Lemma 11 is a reduction from* well-conditioned *optimization to* smooth, convex *optimization. The reduction goes losslessly in the other direction as well, see e.g., [ZH16] for a proof, so in this sense the two settings are essentially equivalent. Indeed, combined with our earlier claim that there is an algorithm achieving $O(\sqrt{\kappa} \log \frac{1}{\epsilon})$ query complexity for an ϵ-multiplicative error decrease, [ZH16] immediately implies a variant of Theorem 3 with convergence rate $O(\frac{LR^2}{T^2})$.*

5 Extensions

We discuss two extensions to the previous sections. We primarily focus on modifications to Sections 3 and 4, as extending Section 2 to a more general framework is the focus of the next lecture. Specifically, we address the following two questions which are suggested by our earlier exposition.

1. Definitions 1, 3, and 4 are stated for general norms $\|\cdot\|$,[10] but our development largely focused on the setting where $\|\cdot\| = \|\cdot\|_2$. What happens more generally when $\|\cdot\| \ne \|\cdot\|_2$?
2. Theorems 3 and 4 are stated for unconstrained optimization. What happens in the constrained setting (where the function may be non-smooth at the boundary of a convex set)?

5.1 General norms

We begin our discussion of Item 1 by recalling the definition of a dual norm.

Definition 6 (Dual norm)**.** *For a norm $\|\cdot\|$ on $\mathbb{R}^d$, the* dual norm *is $\|\cdot\|_* := \max_{\|\mathbf{x}\| \le 1} \langle \mathbf{x}, \cdot \rangle$.*[11]

We remark that the dual norm to the dual norm (when working in $\mathbb{R}^d$) is the original norm itself. We also observe the following fact from convex analysis, which is often used.

Fact 2. *For any norm $\|\cdot\|$ on $\mathbb{R}^d$, the unit norm ball $\mathbb{B}_{\|\cdot\|}(1)$ is compact. Therefore, for all $\mathbf{y} \in \mathbb{R}^d$ there is a $\mathbf{x} \in \mathbb{R}^d$ achieving $\|\mathbf{x}\| = 1$ and $\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{y}\|_*$.*

A few common examples of dual norms follow for convenience.

1. When $p \ge 1$, the dual of the ℓ_p norm $\|\cdot\|_p$ is $\|\cdot\|_q$, for $q \ge 1$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$, where $q = \infty$ if $p = 1$. For example, the ℓ_2 norm is self-dual, and in fact it is the only self-dual norm. Moreover, the dual of the Schatten-p norm (a matrix generalization of the ℓ_p norm equal to the ℓ_p norm of the singular values) is the Schatten-q norm, for q defined as above.
2. When $\mathbf{A} \in \mathbb{S}^{d \times d}_{\succ \mathbf{0}}$ and $\|\mathbf{x}\|_{\mathbf{A}} := (\mathbf{x}^\top \mathbf{A} \mathbf{x})^{1/2}$ is the induced norm, the dual of $\|\cdot\|_{\mathbf{A}}$ is $\|\cdot\|_{\mathbf{A}^{-1}}$.

The former claim follows from Hölder's inequality, and the latter can be proven by using Lagrange multipliers with Definition 6. We next present some useful facts about dual norms.

Lemma 12. *For any norm $\|\cdot\|$ on $\mathbb{R}^d$, $\|\cdot\|_*$ is a norm, and $\langle \mathbf{x}, \mathbf{y} \rangle \le \|\mathbf{x}\| \|\mathbf{y}\|_*$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Moreover, $\|\cdot\|_{**} = \|\cdot\|$.*

[10] Recall that a norm on $\mathbb{R}^d$ is a function : $\mathbb{R}^d \to \mathbb{R}_{\ge 0}$ satisfying positive definiteness, absolute homogeneity, and the triangle inequality. See Lemma 1, Part I and the corresponding footnote for the relevant definitions.

[11] Much of the discussion in this section can be generalized to arbitrary Banach spaces, but we focus on $\mathbb{R}^d$ for simplicity as this is the main setting in applications.

Proof. To see the first claim, absolute homogeneity follows from linearity of Definition 6, and positive definiteness then follows because $\|\mathbf{y}\| > 0$ for any $\mathbf{y} \neq \mathbf{0}_d$, so $\|\mathbf{y}\|_* \geq \langle \mathbf{z}, \mathbf{y} \rangle > 0$ by plugging in $\mathbf{z} = \frac{1}{\|\mathbf{y}\|}\mathbf{y}$, which has $\|\mathbf{z}\| = 1$. The triangle inequality follows since

$$\|\mathbf{y} + \mathbf{y}'\|_* = \max_{\|\mathbf{x}\| \leq 1} \langle \mathbf{x}, \mathbf{y} + \mathbf{y}' \rangle \leq \max_{\|\mathbf{x}\| \leq 1} \langle \mathbf{x}, \mathbf{y} \rangle + \max_{\|\mathbf{x}\| \leq 1} \langle \mathbf{x}, \mathbf{y}' \rangle = \|\mathbf{y}\|_* + \|\mathbf{y}'\|_*.$$

To see the second, given any $\mathbf{x} \neq \mathbf{0}_d$ we can produce a lower bound on $\|\mathbf{y}\|_*$ by applying Definition 6 with $\mathbf{x} \leftarrow \frac{\mathbf{x}}{\|\mathbf{x}\|}$; rearranging gives the claim. When $\mathbf{x} = \mathbf{0}_d$ the claim is clear from $0 \leq 0$.

To see the last claim, consider the Lagrangian formulation of the problem $\min_{\mathbf{z} \in \mathbb{R}^d, \mathbf{z} = \mathbf{x}} \|\mathbf{z}\|$ (where strong duality holds, since we can verify Slater's condition):

$$\begin{aligned} \|\mathbf{x}\| &= \min_{\mathbf{z} \in \mathbb{R}^d, \mathbf{z} = \mathbf{x}} \|\mathbf{z}\| = \min_{\mathbf{z} \in \mathbb{R}^d} \max_{\mathbf{y} \in \mathbb{R}^d} \|\mathbf{z}\| + \mathbf{y}^\top (\mathbf{x} - \mathbf{z}) \\ &= \max_{\mathbf{y} \in \mathbb{R}^d} \mathbf{y}^\top \mathbf{x} + \min_{\mathbf{z} \in \mathbb{R}^d} \|\mathbf{z}\| - \mathbf{y}^\top \mathbf{z} = \max_{\|\mathbf{y}\|_* \leq 1} \mathbf{y}^\top \mathbf{x} = \|\mathbf{x}\|_{**}. \end{aligned}$$

In the last equality, if $\|\mathbf{y}\|_* > 1$, then choosing $\mathbf{z} = C\mathbf{z}_\mathbf{y}$ for $C \to \infty$ we can make the value of the minimization problem $-\infty$, where $\mathbf{z}_\mathbf{y} \in \mathbb{R}^d$ achieves $\langle \mathbf{z}_\mathbf{y}, \mathbf{y} \rangle = \|\mathbf{y}\|_*$ and $\|\mathbf{z}_y\| = 1$ (Fact 2). ☐

By appropriately substituting Fact 2 and Lemma 12 into earlier proofs, we have the following generalizations of Lemmas 5, 6, and 9, which we state without proof here.

Lemma 13. *If $f : \mathcal{X} \to \mathbb{R}$ is convex and L-Lipschitz, $\mathbf{x} \in \text{relint}(\mathcal{X})$, and $\mathbf{g} \in \partial f(\mathbf{x})$, then $\|\mathbf{g}\|_* \leq L$.*

Lemma 14. *If $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable and convex, then f is L-smooth with respect to $\|\cdot\|$ iff for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$,*

$$f(\mathbf{x}') \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{x}' - \mathbf{x}\|^2.$$

If f is twice-differentiable (and possibly nonconvex), f is L-smooth in $\|\cdot\|$ iff $|\nabla^2 f(\mathbf{x})[\mathbf{v}, \mathbf{v}]| \leq L \|\mathbf{v}\|^2$ for all $\mathbf{x}, \mathbf{v} \in \mathbb{R}^d$.

If $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable, then f is μ-strongly convex with respect to $\|\cdot\|$ iff for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$,

$$f(\mathbf{x}') \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{x}' - \mathbf{x}\|^2.$$

If f is twice-differentiable, f is μ-strongly convex in $\|\cdot\|$ iff $\nabla^2 f(\mathbf{x})[\mathbf{v}, \mathbf{v}] \geq \mu \|\mathbf{v}\|^2$ for all $\mathbf{x}, \mathbf{v} \in \mathbb{R}^d$.

Importantly, we can prove the following generalization of Corollary 2, which was the key to establishing the progress made by gradient descent in Theorem 3.

Corollary 4. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth with respect to $\|\cdot\|$. Then for any $\mathbf{x} \in \mathbb{R}^d$, letting*

$$\mathbf{x}' \leftarrow \text{argmin}_{\mathbf{x}' \in \mathbb{R}^d} f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{x}' - \mathbf{x}\|^2, \; f(\mathbf{x}') \leq f(\mathbf{x}) - \frac{1}{2L} \|\nabla f(\mathbf{x})\|_*^2.$$

Proof. It suffices to take $\mathbf{x}' = \mathbf{x} - \frac{1}{L} \|\nabla f(\mathbf{x})\|_* \mathbf{v}$, where $\|\mathbf{v}\| = 1$ and $\langle \nabla f(\mathbf{x}), \mathbf{v} \rangle = \|\nabla f(\mathbf{x})\|_*$. ☐

By substituting Lemma 12, Lemma 14, and Corollary 4 appropriately into the proof of Theorem 3, we hence have the following generalization of gradient descent to general norms due to [KLOS14].

Theorem 6 (Smooth gradient descent in general norms)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth with respect to $\|\cdot\|$ and convex, and suppose for $\mathbf{x}_0 \in \mathbb{R}^d$, we have $\max_{\mathbf{x} \in \mathbb{R}^d, f(\mathbf{x}) \leq f(\mathbf{x}_0)} \|\mathbf{x} - \mathbf{x}^\star\| \leq R$ for $\mathbf{x}^\star \in \text{argmin}_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$. Then iterating*

$$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \frac{1}{L} \|\nabla f(\mathbf{x}_t)\|_* \mathbf{v}_t, \; \text{where } \langle \nabla f(\mathbf{x}_t), \mathbf{v}_t \rangle = \|\nabla f(\mathbf{x}_t)\|_*, \; \|\mathbf{v}_t\| = 1,$$

for $0 \leq t < T$, we have

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{2LR^2}{T}.$$

We mention that often, when $\|\cdot\|$ is explicit, we can compute the update direction $\mathbf{v}_t$ in closed form. The main difference between Theorem 6 and Theorem 3 (beyond the different choice of updates) is that we require the slightly stronger assumption that all points with function value at most $f(\mathbf{x}_0)$ are contained in $\mathbb{B}_{\|\cdot\|}(\mathbf{x}^\star, R)$, rather than just letting $R = \|\mathbf{x}_0 - \mathbf{x}^\star\|$.[12] The culprit for this is that we can no longer establish a contraction guarantee such as Lemma 8 for general norms.

Finally, we mention that the notion of the "well-conditioned setting" in optimization for general norms can be drastically different than in the Euclidean setting. For example, there are norms (such as $\|\cdot\|_\infty$) where any function which is L-smooth and μ-strongly convex in that norm necessarily has $\frac{L}{\mu} = \Omega(d)$, obviating the advantage of gradient descent-based methods over cutting-plane methods; similar challenges are met when trying to design generalized accelerated gradient descent algorithms. We will explore in a later lecture what types of linearly convergent or accelerated guarantees we can hope for in generalized settings, which requires developing new technology.

5.2 Composite objectives

We next discuss Item 2, by providing a means to optimize more general composite objectives. In particular, we give an algorithm which applies to functions of the form

$$F(\mathbf{x}) := f(\mathbf{x}) + \psi(\mathbf{x}), \text{ for convex } f, \psi : \mathbb{R}^d \to \mathbb{R}, \tag{13}$$

where f is smooth[13] and ψ is "simple." Concretely, we make the following assumption about ψ.

Definition 7 (Proximal oracle). *We say $\mathcal{O}$ is a* proximal oracle *for $\psi : \mathbb{R}^d \to \mathbb{R}$ if for any $\mathbf{v} \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}_{\geq 0}$, $\mathcal{O}(\lambda, \mathbf{v})$ returns $\operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^d} \frac{\lambda}{2} \|\mathbf{x} - \mathbf{v}\|_2^2 + \psi(\mathbf{x})$.*

To gain some intuition for Definition 7, note that when $\psi = \chi_S$ is the indicator of a convex set S, a proximal oracle returns the Euclidean projection of v onto S. In this case, (13) generalizes the setting of constrained, smooth function minimization as asked by Item 2. More generally, proximal oracle access to ψ can be viewed as a measure of how simple or explicit ψ is. A common example of when ψ admits a linear-time proximal oracle is when ψ is coordinatewise separable, as then it reduces to solving d one-dimensional problems. Common regularizers in machine learning, e.g., the Lasso (ℓ_1 regularization) or ElasticNet, indeed have this separability property.

We conclude by generalizing Theorems 3 and 4 to the setting of (13), following [Sid23].

Theorem 7 (Proximal well-conditioned gradient descent). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and convex, let $\psi : \mathbb{R}^d \to \mathbb{R}$ admit a proximal oracle $\mathcal{O}$, let $F := f + \psi$ be μ-strongly convex, and let $\mathbf{x}^\star := \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x})$ and $\kappa := \frac{L}{\mu}$. Then iterating*[14]

$$\mathbf{x}_{t+1} \leftarrow \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{x}_t\|_2^2 + \psi(\mathbf{x}) \tag{14}$$

for $0 \leq t < T$, we have

$$F(\mathbf{x}_T) - F(\mathbf{x}^\star) \leq \left(1 - \frac{1}{\kappa + 1}\right)^T \left(F(\mathbf{x}_0) - F(\mathbf{x}^\star)\right).$$

[12] This is relevant because Corollary 4 guarantees each iterate $\mathbf{x}_t$ has $f(\mathbf{x}_t) \leq f(\mathbf{x}_0)$, so $\|\mathbf{x}_t - \mathbf{x}^\star\| \leq R$.

[13] For simplicity in this section, we focus on the case of smoothness in the ℓ_2 norm, but using the aforementioned techniques in general norms much of this section can be generalized appropriately.

[14] Note that this step can be implemented with one call to $\mathcal{O}(L, \mathbf{x}_t - \frac{1}{L} \nabla f(\mathbf{x}_t))$.

Proof. Consider a single iteration t, and let $\mathbf{x}_t^{(\lambda)} := (1-\lambda)\mathbf{x}_t + \lambda\mathbf{x}^\star$ for $\lambda \in [0,1]$. We derive:

$$\begin{aligned}
F(\mathbf{x}_{t+1}) &\le f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2}\left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|_2^2 + \psi(\mathbf{x}_{t+1}) \\
&= \min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle + \frac{L}{2}\left\|\mathbf{x} - \mathbf{x}_t\right\|_2^2 + \psi(\mathbf{x}) \\
&\le \min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) + \frac{L}{2}\left\|\mathbf{x} - \mathbf{x}_t\right\|_2^2 \le \min_{\lambda\in[0,1]} F(\mathbf{x}_t^{(\lambda)}) + \frac{L}{2}\|\mathbf{x}_t^{(\lambda)} - \mathbf{x}_t\|_2^2 \\
&\le \min_{\lambda\in[0,1]} (1-\lambda)F(\mathbf{x}_t) + \lambda F(\mathbf{x}^\star) - \frac{\mu\lambda(1-\lambda)}{2}\left\|\mathbf{x}_t - \mathbf{x}^\star\right\|_2^2 + \frac{L\lambda^2}{2}\left\|\mathbf{x}_t - \mathbf{x}^\star\right\|_2^2 \\
&\le \left(1 - \frac{1}{\kappa+1}\right) F(\mathbf{x}_t) + \frac{1}{\kappa+1} F(\mathbf{x}^\star).
\end{aligned}$$

The first line used smoothness of f, the second used the definition of $\mathbf{x}_{t+1}$, the third used convexity of f and that restricting to a subset of $\mathbb{R}^d$ can only increase the minimum, the fourth used strong convexity of F, and the last used the particular choice $\lambda = \frac{1}{\kappa+1}$. Finally, rearranging yields the claim $F(\mathbf{x}_{t+1}) - F(\mathbf{x}^\star) \le (1 - \frac{1}{\kappa+1})(F(\mathbf{x}_t) - F(\mathbf{x}^\star))$, and iterating gives the conclusion. □

Theorem 8 (Proximal smooth gradient descent)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and convex, let $\psi : \mathbb{R}^d \to \mathbb{R}$ admit a proximal oracle $\mathcal{O}$, and let $F := f + \psi$ be convex. Suppose for $\mathbf{x}_0 \in \mathbb{R}^d$, we have $\max_{\mathbf{x}\in\mathbb{R}^d, F(\mathbf{x})\le F(\mathbf{x}_0)} \|x - x^\star\|_2 \le R$ for $\mathbf{x}^\star \in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x})$. Then iterating (14) for $0 \le t < T$, we have*

$$F(\mathbf{x}_T) - F(\mathbf{x}^\star) \le \frac{2LR^2}{T-1}.$$

Proof. We first observe that by convexity and the definitions of x_1 and R,

$$F(\mathbf{x}_1) \le f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x}^\star - \mathbf{x}_0 \rangle + \frac{L\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|_2^2}{2} + \psi(\mathbf{x}^\star) \le F(\mathbf{x}^\star) + \frac{LR^2}{2}.$$

Thus, denoting $\Phi_t := F(\mathbf{x}_t) - F(\mathbf{x}^\star)$ and $R_t := \left\|\mathbf{x}_t - \mathbf{x}^\star\right\|_2$ for all $0 \le t \le T$, we have shown $\Phi_1 \le \frac{LR^2}{2}$. Next, repeating the proof of Theorem 7 up to where we used strong convexity, we have

$$\begin{aligned}
F(\mathbf{x}_{t+1}) &\le F(\mathbf{x}_t) + \min_{\lambda\in[0,1]} \left(-\lambda\Phi_t + \frac{L\lambda^2 R_t^2}{2}\right) \\
&\le F(\mathbf{x}_t) - \min\left(\frac{\Phi_t^2}{2LR_t^2}, \Phi_t - \frac{LR_t^2}{2}\right) \\
&\le F(\mathbf{x}_t) - \min\left(\frac{\Phi_t^2}{2LR^2}, \frac{\Phi_t}{2}\right) = F(\mathbf{x}_t) - \frac{\Phi_t^2}{2LR^2},
\end{aligned}$$

for all $t \ge 1$. In the second inequality, the minimizing $\lambda \in \mathbb{R}$ is at $\lambda^\star := \frac{\Phi_t}{LR_t^2}$, so we take $\lambda \leftarrow \min(\lambda^\star, 1)$; the third used that if $\lambda = 1$, we have $LR_t^2 \le \Phi_t$. The last equality used $\Phi_t \le \frac{LR^2}{2}$ for all $t \ge 1$. The remainder of the proof is identical to Theorem 3, offsetting indices by 1. □

Source material

Portions of this lecture are based on reference material in [NY83, Nes03, Bub15, Tib16, Sid23], as well as the author's own experience working in the field.

References

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

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[CDHS20] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Math. Program.*, 184(1):71–120, 2020.

[CJJ+20] Yair Carmon, Arun Jambulapati, Qijia Jiang, Yujia Jin, Yin Tat Lee, Aaron Sidford, and Kevin Tian. Acceleration with a ball optimization oracle. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, 2020.

[GLM16] Rong Ge, Jason D. Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 2973–2981, 2016.

[HM17] Moritz Hardt and Tengyu Ma. Identity matters in deep learning. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[KLOS14] Jonathan A. Kelner, Yin Tat Lee, Lorenzo Orecchia, and Aaron Sidford. An almost-linear-time algorithm for approximate max flow in undirected graphs, and its multicommodity generalizations. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014*, pages 217–226. SIAM, 2014.

[KNS16] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016*, volume 9851 of *Lecture Notes in Computer Science*, pages 795–811. Springer, 2016.

[LSB12] Simon Lacoste-Julien, Mark Schmidt, and Francis R. Bach. A simpler approach to obtaining an o(1/t) convergence rate for the projected stochastic subgradient method. *CoRR*, abs/1212.2002, 2012.

[Nes83] Yurii Nesterov. A method for solving a convex programming problem with convergence rate $o(1/k^2)$. *Doklady AN SSSR*, 269:543–547, 1983.

[Nes03] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course, volume I.* 2003.

[NY83] A. Nemirovski and D.B̃. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

[RSS12] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 2012.

[Sid23] Aaron Sidford. *Optimization Algorithms*. 2023.

[Tib16] Ryan Tibshirani. Lecture 13: Duality uses and correspondences. class notes, cmu 10-725/36-725: Convex optimization. https://www.stat.cmu.edu/ ryantibs/convexopt-F16/scribes/dual-corres-scribed.pdf, 2016.

[XLS17] Bo Xie, Yingyu Liang, and Le Song. Diverse neural network learns true target functions. In *Proceedings of the 20th International Conference on Artificial Intelligence and*

Statistics, AISTATS 2017, volume 54 of *Proceedings of Machine Learning Research*, pages 1216–1224. PMLR, 2017.

[ZH16] Zeyuan Allen Zhu and Elad Hazan. Optimal black-box reductions between optimization objectives. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 1606–1614, 2016.