Geometric Gradient Descent and Lower Bounds*

Aaron Geelon So

February 19, 2019

Given a convex body $\mathcal{X} \subset \mathcal{H}$ inside a Hilbert space and convex function $f : \mathcal{X} \to \mathbb{R}$, we aim to optimize:

$$\min_{x \in \mathcal{X}} f(x).$$

We'll denote by $x^* \in \arg\min_{x \in \mathcal{X}} f(x)$ to be any optimizer. Last time, we saw the projected gradient approach: begin with any $x_0 \in \mathcal{X}$. Then, iteratively compute:

$$\begin{aligned} y_{t+1} &= x_t - \eta g_t \\ x_{t+1} &= \Pi_{\mathcal{X}}(y_{t+1}), \end{aligned}$$

where $g_t \in \partial f(x_t)$ is a subgradient at x_t, η controls the step size and $\Pi_{\mathcal{X}} : \mathcal{H} \to \mathcal{X}$ is the projection back into $\mathcal{X}$. Assume that we obtain our subgradients from an oracle; we'd like to also give oracle complexity bounds. We saw one last time for Lipschitz convex functions that gives an oracle compleixty of $O\left(\frac{1}{\epsilon^2}\right)$:

Theorem 1 (Theorem 3.2, [B2015])**.** *Let $\mathcal{X} \subset B(0, R)$ be bounded. Let f be an L-Lipschitz convex function. Let $\eta = \frac{R}{L\sqrt{t}}$. Then, the projected gradient descent method satisfies:*

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^*) \leq \frac{RL}{\sqrt{t}}.$$

Proof. First we appeal to convexity:

$$f(x_s) - f(x^*) \leq g_s^T (x_s - x^*). \tag{1}$$

In the projected gradient approach, we can write $g_t = -\frac{1}{\eta}(y_{t+1} - x_t)$. Plug this back in and rearrange to:

$$f(x_s) - f(x^*) \leq \frac{1}{2\eta}\left(\|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2\right) + \frac{\eta}{2}\|g_s\|^2.$$

Notice that $\|x_{s+1} - x^*\|^2 \leq \|y_{s+1} - x^*\|^2$. It follows that:

$$f(x_s) - f(x^*) \leq \frac{1}{2\eta}\left(\|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2\right) + \frac{\eta}{2}\|g_s\|^2. \tag{2}$$

If we sum these terms over s from 1 to t, the blue terms will telescope. Then, we can bound it using the boundedness of $\mathcal{X}$. We can also bound the orange terms using the Lipschitz condition. And so:

$$\frac{1}{t}\sum_{s=1}^{t} f(x_s) - f(x^*) \leq \frac{1}{2\eta t}R^2 + \frac{\eta L^2}{2}.$$

The choice $\eta = \frac{R}{L\sqrt{t}}$ balances these two terms. Finally, appeal to Jensen's. □

In fact, under a 'black-box model' (defined later) restricting allowable queries to the oracle, gradient descent achieves the optimal rate. As we add more niceness conditions, not only will gradient descent converge faster, but at some point, we can take an approach, *geometric descent*, whose rates overtake gradient descent.

*This lecture covers sections 3.4–3.7 of [B2015].

1 Strong convexity

While convexity just gives a linear lower bound at every point $x \in \mathcal{X}$, strong convexity will give a stronger quadratic lower bound.

Definition 2 (α-strong convexity)**.** *A function f is α-strongly convex if it satisfies:*

$$f(x) - f(y) \leq g_x^T(x-y) - \frac{\alpha}{2}\|x-y\|^2,$$

where $g_x \in \partial f(x)$.

Notice that when $y = x^*$, α-strong convexity upper bounds how close $f(x)$ is to optimal. Also, by rearranging the terms, we see that α-strong convexity is equivalent to a quadratic lower bound on f:

$$f(y) \geq f(x) + g_x^T(y-x) + \frac{\alpha}{2}\|x-y\|^2.$$

To add to our understanding of α-strong convexity, we make three remarks:

Remark 3. *If f is smooth, α-strong convexity is equivalent to $\alpha I \preceq H(f)$, where $H(f)$ is the Hessian. In particular the function $f(x) = \frac{\alpha}{2}\|x\|^2$ is α-strongly convex. Indeed, we have:*

$$\frac{\alpha}{2}\|x\|^2 - \frac{\alpha}{2}\|y\|^2 = \alpha x^T(x-y) + \frac{\alpha}{2}\|x-y\|^2.$$

Remark 4. *If f is α-strongly convex and f' is α'-strongly convex, then their sum is $(\alpha+\alpha')$-strongly convex.*

Remark 5. *The map f is α-strongly convex if and only if $x \mapsto f(x) - \frac{\alpha}{2}\|x\|^2$ is convex.*

Proof of Remark 5. ($\Longrightarrow$) Let $g_x \in \partial f(x)$. If f is α-strongly convex, then combining with the first remark:

$$\begin{aligned} f(x) - \frac{\alpha}{2}\|x\|^2 - f(y) + \frac{\alpha}{2}\|y\|^2 &\leq g_x^T(x-y) - \frac{\alpha}{2}\|x-y\|^2 - \frac{\alpha}{2}\left(\|x\|^2 - \|y\|^2\right) \\ &= (g_x - \alpha x)^T(x-y) - \frac{\alpha}{2}\|x-y\|^2 + \frac{\alpha}{2}\|x-y\|^2 \\ &= (g_x - \alpha x)^T(x-y), \end{aligned}$$

showing convexity.

($\Longleftarrow$) If the map $x \mapsto f(x) - \frac{\alpha}{2}\|x\|^2$ is convex (i.e. 0-strongly convex), then the map f is α-strongly convex by the previous two remarks. □

Returning now to gradient descent, if f is not only L-Lipschitz but also α-strongly convex, then we could strengthen the bound in Equation 1 to:

$$f(x_s) - f(x^*) \leq g_s^T(x_s - x) - \frac{\alpha}{2}\|x_s - x^*\|^2.$$

As a result, the expanded bound in Equation 2 becomes:

$$f(x_s) - f(x^*) \leq \left(\frac{1}{2\eta_s} - \frac{\alpha}{2}\right)\|x_s - x^*\|^2 - \frac{1}{2\eta_s}\|x_{s+1} - x^*\|^2 + \frac{\eta_s}{2}\|g_s\|^2.$$

Unfortunately, the sum does not telescope as easily as before, but if we let $\eta_s = \frac{2}{\alpha(s+1)}$, then we get:

$$s \cdot \big(f(x_s) - f(x^*)\big) \leq \frac{\alpha}{4}\left(s(s-1)\|x_s - s^*\|^2 - (s+1)s\|x_{s+1} - x^*\|^2\right) + \frac{L^2}{\alpha},$$

which does telescope. As before, we can sum these terms up over s from 1 to t, renormalize by $\sum_{s=1}^t s = \frac{s(s+1)}{2}$, then apply Jensen's inequality. When the dust clears, we obtain:

Theorem 6 (Theorem 3.9, [B2015]). *Let f be α-strongly convex and L-Lipschitz on $\mathcal{X}$. Then, projected gradient descent with $\eta_s = \frac{2}{\alpha(s+1)}$ satisfies:*

$$f\left(\frac{2}{t(t+1)}\sum_{s=1}^{t} s \cdot x_s\right) - f(x^*) \leq \frac{2L^2}{\alpha(t+1)}.$$

And so, whereas before with just the L-Lipschitz condition, we obtained a convergence rate of $O\left(\frac{1}{\epsilon^2}\right)$, with an additional α-strong convexity assumption, we obtain a new rate of $O\left(\frac{1}{\alpha\epsilon}\right)$.

An interesting effect of including α-strong convexity is that the our oracle complexity does not require an upper bound on the size of the domain: we don't need that $\mathcal{X} \subset B(0, R)$. Note that with a quadratic lower bound $q^-(x) \leq f(x)$, any upper bound $U > f(x^*)$ will automatically restrict the 'candidate' subset $\{x : q^-(x) < U\}$ to a bounded set. In particular, this is some ball centered around $c = \arg\min q^-(x)$. This will be an important idea for geometric descent later on.

2 Strong convexity and smoothness

Recall the definition of β-smoothness, a dual condition to α-strong convexity in that it yields a quadratic upper bound to f:

Definition 7 (β-smooth). *A function f is* β-smooth *if for all $x \in \mathcal{X}$, it is upper bounded:*

$$f(y) \leq f(x) + g_x^T(y - x) + \frac{\beta}{2}\|x - y\|^2,$$

where $g_x \in \partial f(x)$.

Smoothness will be a useful structural constraint that helps us make progress closing into $f(x^*)$: if we're at some suboptimal $x_s \in \mathcal{X}$ then a quadratic upper bound ensures that if we take a step along the gradient of size related to β, we are guaranteed to find a point $x_{s+1} = \arg\min q^+(x)$ where $f(x_{s+1}) < f(x_s)$. For example, in Figure 1, let $x_s = x$ and $x_{s+1} = x^+$.

If a function is both α-strongly convex and β-smooth, then define the two quadratics:

$$q_x^-(y) = f(x) + g_x^T(y - x) + \frac{\alpha}{2}\|x - y\|^2$$
$$q_x^+(y) = f(x) + g_x^T(y - x) + \frac{\beta}{2}\|x - y\|^2.$$

These lower and upper bound f:

$$q_x^-(y) \leq f(y) \leq q_x^+(y).$$

Notice that this immediately implies that $\beta \geq \alpha$. Their ratio will be an important quantity related to how quickly gradient descent (and later, geometric descent) will converge:

Definition 8 (Condition number). *Let f be α-strongly convex and β-smooth. Its* condition number *is $\kappa = \frac{\beta}{\alpha}$.*

As a preview to geometric descent, recall our previous discussion that if f is α-strongly convex, any upper bound U on $f(x^*)$ will lead to a bound on the feasible region that contains x^*. On the other hand, consider our last note: that if f is β-smooth and $f(x_s) > f(x^*)$ is suboptimal, we can immediately obtain an improved upper bound on $q^+(x_{s+1}) > f(x^*)$. Geometric descent is an optimization strategy motivated by iteratively reducing the radius of the 'candidate' subset of points that could contain x^*.

But before describing geometric descent, let's determine the convergence rate that gradient descent can achieve on a function that is both α-strongly convex and β-smooth. In the further analyses, it will be useful to know what minimizes the quadratic bounds and what those minima are:

Lemma 9. *Let* $q(y) = g^T(y-x) + \frac{\gamma}{2}\|y-x\|^2$. *Then:*

$$q(y) = \frac{\gamma}{2}\left\|y - \left(x - \frac{1}{\gamma}g\right)\right\|^2 - \frac{\|g\|^2}{2\gamma}.$$

This lemma is easily proven, but the main takeaways are that the minimizer of q is:

$$\arg\min q(y) = x - \frac{1}{\gamma}g,$$

taking on the value:

$$\min q(y) = -\frac{\|g\|^2}{2\gamma}.$$

That is, the minimizer is found by taking a descending step of $\frac{1}{\gamma}$ along the gradient g. Furthermore, this step will decrease the value by $\|g\|^2/2\gamma$. Let $q_x^-(y) \le f(y) \le q_x^+(y)$ be quadratic bounds defined above. Then, we define x^- and x^+ to be:

$$x^- := x - \frac{1}{\alpha}g_x \qquad x^+ := x - \frac{1}{\beta}g_x. \tag{3}$$

Thus, x^- and x^+ minimizes q_x^- and q_x^+, respectively. Figure 1 visualizes this.

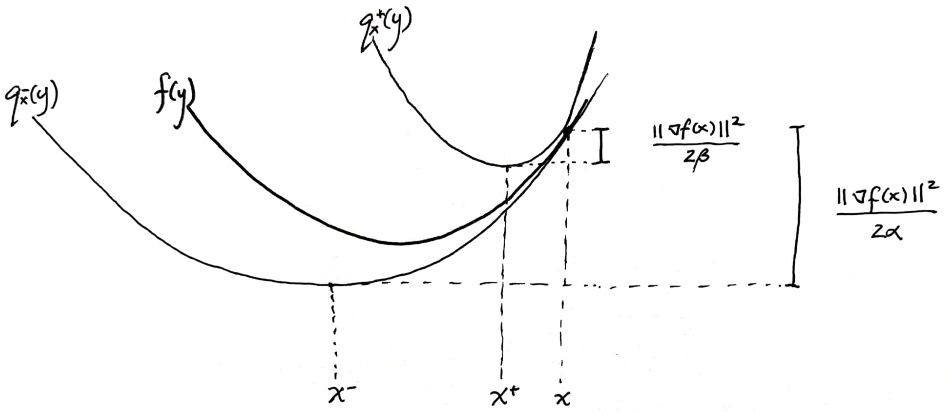

Figure 1: Quadratic lower and upper bounds q_x^- and q_x^+ on f.

With this notation, we can make precise our description above for how α-strong convexity and β-smoothness can both improve upper bounds on $f(x) - f(y)$ over convexity (e.g. in Equation 1). As before, α-strong convexity improves the bound to:

$$f(x) - f(y) \le g_x^T(x-y) - \frac{\alpha}{2}\|x-y\|^2.$$

Now, we know that $f(x^+) \le q^+(x^+) \le f(x) - \frac{\|g_x\|^2}{2\beta}$. Thus:

$$f(x^+) - f(y) \le g_x^T(x-y) - \frac{\alpha}{2}\|x-y\|^2 - \frac{\|g_x\|^2}{2\beta}. \tag{4}$$

Having both α-strong convexity and β-smoothness will let us improve the bound over L-Lipschitz and α-strong convexity, potentially exponentially.

Theorem 10 (Theorem 3.10, [B2015]). *Let f be α-strongly convex and β-smooth on $\mathcal{X}$. Then, gradient descent with $\eta = \frac{1}{\beta}$ satisfies for $t \geq 0$:*

$$\|x_{t+1} - x^*\|^2 \leq \exp\left(-\frac{t}{\kappa}\right) \|x_1 - x^*\|^2 .$$

Proof. Notice that when $\eta = \frac{1}{\beta}$, then in gradient descent, $x_{t+1} = x_t^+$. Consider Equation 4 where $x = x_t$ and $y = x^*$. Then:

$$0 \leq f(x_{t+1}) - f(x^*) \leq g_t^T(x_t - x^*) - \frac{\alpha}{2} \|x_t - x^*\|^2 - \frac{\|g_t\|^2}{2\beta}.$$

In particular, this means that we have:

$$-\frac{2}{\beta} g_t^T(x_t - x^*) + \frac{\|g_t\|^2}{\beta^2} \leq -\frac{\alpha}{\beta} \|x_t - x^*\|^2 . \tag{5}$$

We can use this inequality to show:

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &= \|x_t - \frac{1}{\beta} g_t(x_t) - x^*\|^2 \\
&= \|x_t - x^*\|^2 + \left(-\frac{2}{\beta} g_t^T(x_t - x^*) + \frac{\|g_t\|^2}{\beta^2}\right) \\
&\leq \left(1 - \frac{\alpha}{\beta}\right) \|x_t - x\|^2 \\
&\leq \left(1 - \frac{\alpha}{\beta}\right)^t \|x_1 - x^*\|^2 \leq \exp\left(-\frac{t}{\kappa}\right) \|x_1 - x\|^2 ,
\end{aligned}$$

where the next-to-last line follows from Equation 5. □

This theorem thus shows that the oracle complexity of gradient descent for a function that is both α-strongly convex and β-smooth is $O\left(\kappa \ln \frac{1}{\epsilon}\right)$. Now, we see how geometric descent improves this rate by $\sqrt{\kappa}$.

3 Geometric descent

First, in the following lemma, we formalize how α-strong convexity lets us bound the 'candidate' subset of points that can contain x^*. This is most easily seen from Figure 2. In the following, denote by $\mathbf{B}(x, r^2)$ the ball centered at $x \in \mathbb{R}^n$ with squared radius r^2 (this will help us write fewer square roots).

Lemma 11. *Let f be α-strongly convex. Let $q_x^-(y)$ be a quadratic lower bound to f defined above. Suppose that we have an upper bound on $f(x^*) \leq q_x^-(x^-) + M$. Then, x^* is contained in the following ball with squared-radius $2M/\alpha$,*

$$x^* \in B\left(x^-, \frac{2M}{\alpha}\right).$$

The main idea of geometric descent is to combine this property guaranteed by α-strong convexity with β-smoothness: for any point x, the value $f(x)$ is an upper bound on $f(x^*)$. Thus, this gives us a bound on the radius of a ball that contains x^*. However, we can restrict the radius more because β-smoothness gives us a stronger upper bound of $q_x^+(x^+)$.

A natural algorithm using this idea is to iteratively find balls of decreasing radii that are guaranteed to contain x^*. We will rely on the following lemma:

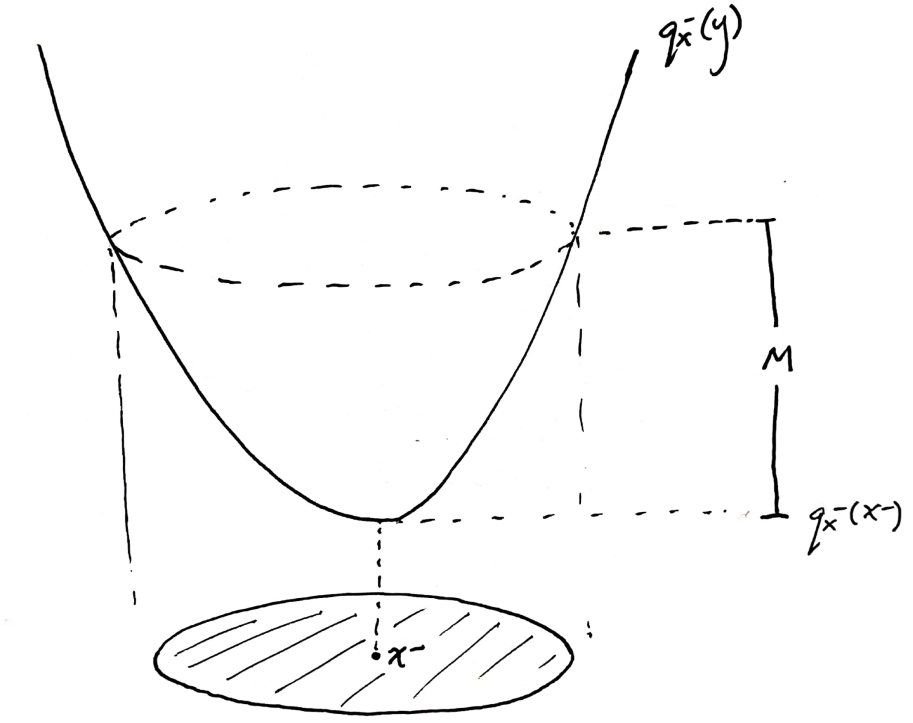

Figure 2: If $q_x^-(y)$ is a lower bound on $f(y)$, then an upper bound on $f(x^*)$ of $q_x^-(x^-)+M$ ensures that x^* is contained in the ball centered at x^- of radius $\sqrt{\frac{2M}{\alpha}}$.

Lemma 12. *Let $\mathbf{B}(x, r^2)$ be a ball centered at $x \in \mathbb{R}^n$ with radius r. Let $z \in \mathbb{R}^n$ be any vector. Then, for any $\epsilon \in (0, 1)$, there exists a ball centered at some $x' \in \mathbb{R}^n$ satisfying:*

$$\mathbf{B}\left(x, r^2\right) \cap \mathbf{B}\left(z, (1-\epsilon)\left\|x - z\right\|^2\right) \subset \mathbf{B}\left(x', (1-\epsilon)r^2\right).$$

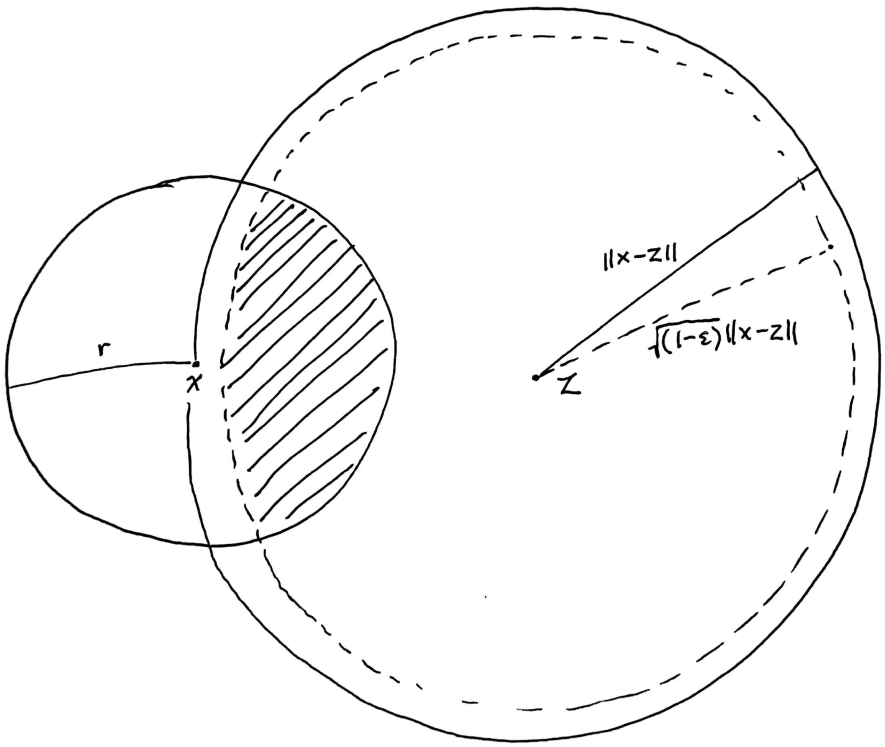

Figure 3: The shaded part is the intersection of the two balls in Lemma 12 on the left hand side. The lemma states that this shaded region can be contained in a ball of squared-radius at most $(1-\epsilon)r^2$.

In this lemma, think of $\mathbf{B}(x, r^2)$ as a ball that contains x^*. Since $f(x)$ is an upper bound on $f(x^*)$, Lemma 11 tells us that x^* is also contained in the ball $\mathbf{B}(x^-, 2M/\alpha)$, where:

$$M = f(x) - q_x^-(x^-) = \frac{\left\|\nabla f(x)\right\|^2}{2\alpha}.$$

This is easily seen from Figure 1. If we let $z = x^-$, then this latter ball is $\mathbf{B}(z, \|x - z\|^2)$.

Now, we use β-smoothness; from Figure 1, we see that we can improve the upper bound by $\left\|\nabla f(x)\right\|^2 / 2\beta$. So, we let M' be:

$$M' = \frac{\left\|\nabla f(x)\right\|^2}{2\alpha} - \frac{\left\|\nabla f(x)\right\|^2}{2\beta} = \left(1 - \frac{\alpha}{\beta}\right) M,$$

noting that x^* is contained in the smaller ball $\mathbf{B}(x^-, 2M'/\alpha)$. Letting $\epsilon = \frac{1}{\kappa}$, this is equivalent to:

$$x^* \in \mathbf{B}\left(x^-, (1-\epsilon)\left\|x - z\right\|^2\right).$$

We can apply Lemma 12 to find a new ball with squared-radius that has shrunk from the original r^2 to $\left(1 - \frac{1}{\kappa}\right) r^2$. It follows that if we iterate t times, then we can reduce an original radius by a factor of $\left(1 - \frac{1}{\kappa}\right)^t$; this algorithm can also achieves the $O\left(\kappa \ln \frac{1}{\epsilon}\right)$ rate as gradient descent.

It is possible to further improve on this algorithm. Here, we reduced the squared-radius of the ball $\mathbf{B}\left(z, \|x-z\|^2\right)$ by $(1-\epsilon)$ by improving the upper bound on $f(x^*)$ by $(1-\epsilon)$. But notice that in the iterative algorithm we just described, the radius of $\mathbf{B}(x, r^2)$ is also assumed to be derived from an upper bound on $f(x^*)$. So, we should also be able to simultaneously shrink the former ball with the latter using the improved upper bound. Indeed, we can, and this will lead to an *accelerated* geometric descent algorithm that will achieve an $O\left(\sqrt{\kappa}\ln\frac{1}{\epsilon}\right)$ rate.

3.1 Geometric descent with acceleration

The analogous lemma to Lemma 12 that will enable acceleration is this:

Lemma 13 (Lemma 3.16, [B2015])**.** *Let $\mathbf{B}(x, r^2)$ be a ball centered at $x \in \mathbb{R}^n$ with radius r. Let $z \in \mathbb{R}^n$, and suppose that $\|x-z\| \geq s$. Then, for all $\epsilon \in (0,1)$ and all $\delta \geq 0$, there exists some $x' \in \mathbb{R}^n$ satisfying:*

$$\mathbf{B}\left(x, r^2 - \epsilon s^2 - \delta\right) \cap \mathbf{B}\left(z, (1-\epsilon)s^2 - \delta\right) \subset \mathbf{B}\left(x', (1-\sqrt{\epsilon})r^2 - \delta\right).$$

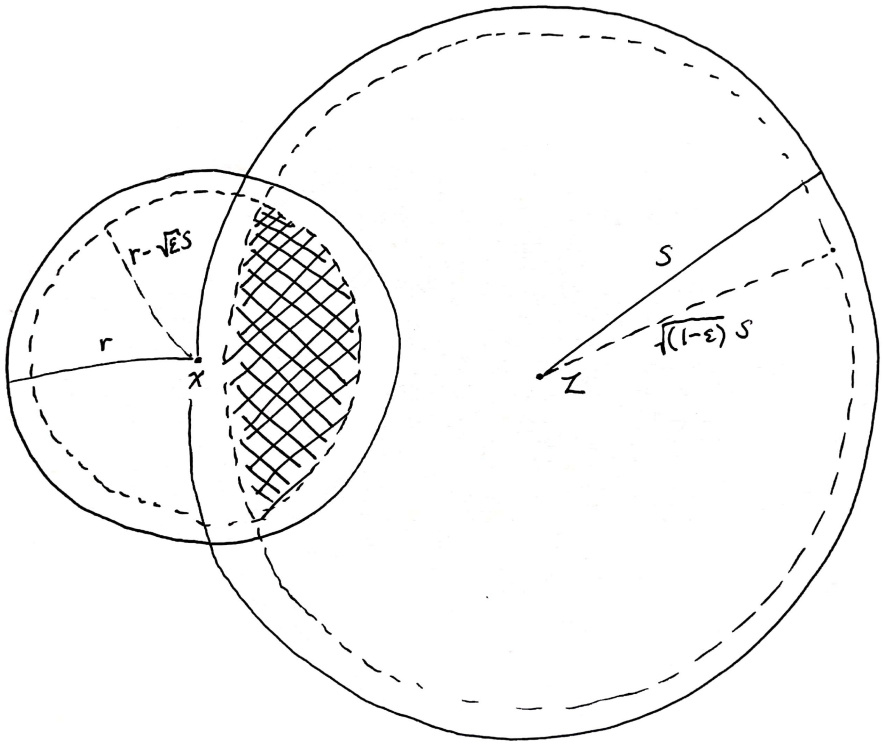

Figure 4: The shaded part is the intersection of the two balls in Lemma 13 on the left hand side. This time, we shrink both balls. The lemma states that this shaded region can be contained in a ball of squared-radius at most $(1-\sqrt{\epsilon})r^2$.

Like the previous lemma, the $-\epsilon s^2$ term comes from an improvement in upper bound based on β-smoothness. Notice that this time, the improvement is applied to the first ball in addition to the second ball. Additionally, in this lemma, the $-\delta$ term will allow us to propagate improvements back through all of the

previously constructed ball. At a high level, each iteration will reduce the radii of balls containing x^* by a factor of $(1-\sqrt{\epsilon})$, where $\epsilon = \frac{1}{\kappa}$, as before. This will lead to the promised rate of $O\left(\sqrt{\kappa}\ln\frac{1}{\epsilon}\right)$.

The geometric descent algorithm follows: let $x_0 \in \mathbb{R}^n$ be any initialization point. Following Lemma 11, we will iteratively improve upper bounds on $f(x^*)$ to reduce the radii of balls containing x^*. A first bound is $f(x^*) \leq f(x_0) - M_0$, where:

$$M_0 = \frac{\|\nabla f(x_0)\|^2}{2\beta}.$$

As we saw before, the corresponding ball that this upper bound yields is:

$$\mathbf{B}(c_0, R_0^2) := \mathbf{B}\left(x_0^-, \left(1-\frac{1}{\kappa}\right)\frac{\|\nabla f(x_0)\|^2}{\alpha^2}\right).$$

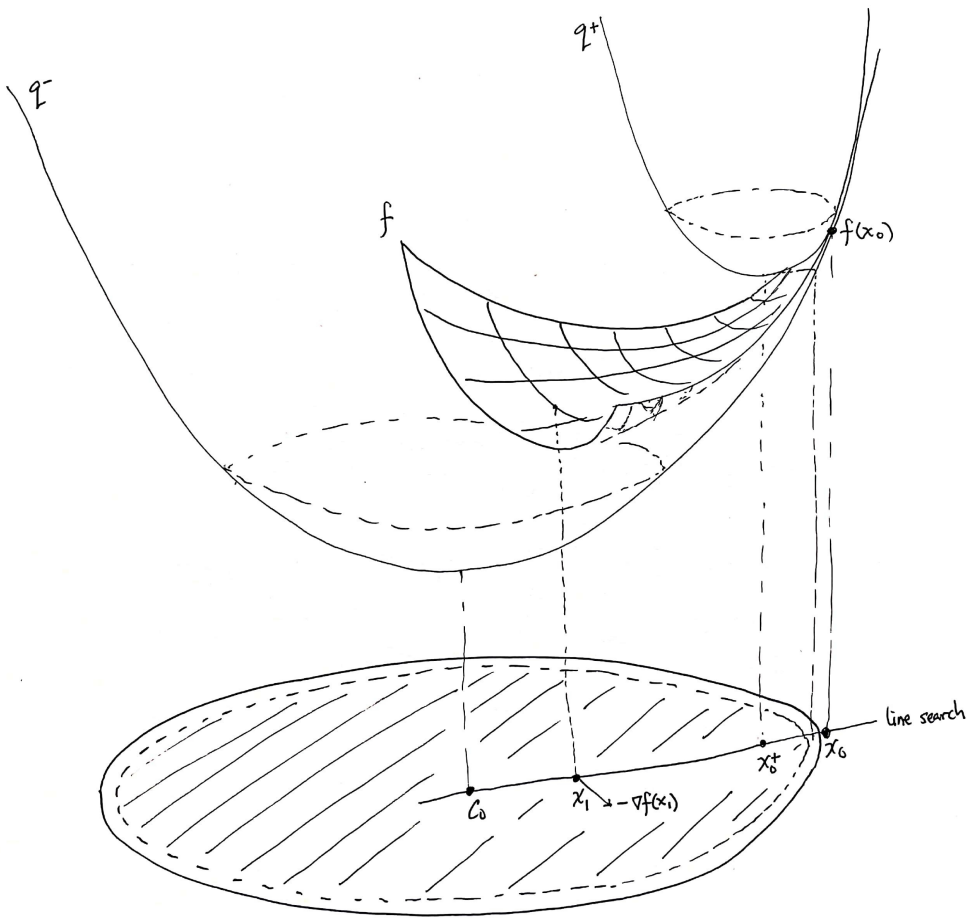

Figure 5: The first step to the accelerated geometric descent algorithm: f is lower and upper bounded by q^- and q^+. The initial point x_0 induces $c_0 = x_0^-$ as a result of α-strong convexity. Then, β-smoothness helps shrink the ball that contains x^* by a factor of $\left(1-\frac{1}{\kappa}\right)$. To obtain the next point x_1, perform line search. In the figure, the vector $-\nabla f(x_1)$ points to the direction the next ball will be chosen. In accelerated geometric descent, when we shrink the next ball, we will return to this first ball and shrink it down even further.

For each iteration, let x_{t+1} as the minimizer of f on the line between c_t and x_t, found via line search:

$$x_{t+1} = \operatorname*{arg\,min}_{\{(1-\lambda)c_t + \lambda x_t : \lambda \in \mathbb{R}\}} f(x).$$

Again, we can improve on the upper bound $f(x^*) \leq f(x_{t+1})$ by $M_{t+1} = \|\nabla f(x_{t+1})\|^2 / 2\beta$. But in fact, because we've performed the line search first, we can propagate this improvement backwards, obtaining:

$$f(x^*) \leq f(x_s) - \sum_{j=s}^{t+1} M_j,$$

for all $s \leq t$. In other words, the line search implies that $f(x_{t+1})$ will be less than all previously obtained upper bounds: $f(x_{t+1}) \leq f(x_s) - \sum_{j=s}^{t} M_j$.

Using Lemma 13, we can shrink all previous balls by a squared-radius of δ_{t+1} by applying Lemma 11,

$$\delta_{t+1} = \frac{2}{\alpha} M_{t+1} = \frac{1}{\kappa} \frac{\|\nabla f(x_{t+1})\|^2}{\alpha^2}.$$

Finally, we let c_{t+1} and R_{t+1} be the point and radius found by Lemma 13 applied to:

$$\mathbf{B}\left(c_t, R_t^2 - \frac{1}{\kappa}\frac{\|\nabla f(x_{t+1})\|^2}{\alpha^2}\right) \cap \mathbf{B}\left(x_{t+1}^-, \left(1 - \frac{1}{\kappa}\right)\frac{\|\nabla f(x_{t+1})\|^2}{\alpha^2}\right) \subset \mathbf{B}\left(c_{t+1}, R_{t+1}^2\right).$$

Formally, we have the following theorem to bound the convergence rate:

Theorem 14 (Theorem 3.17, [B2015])**.** *Let $t \geq 0$. Then $x^* \in \mathbf{B}\left(c_t, R_t^2\right)$ and $R_{t+1}^2 \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right) R_t^2$. Thus:*

$$\|x^* - c_t\|^2 \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^t R_0^2.$$

This rate is tight under the so-called *black-box model*, which we will discuss now.

4 Lower bounds

Now that we have some upper bounds on the oracle complexity, let's look at lower bounds. We assume the following "black-box procedure", where given a *history* of queries and gradients:

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

where $g_s \in \partial f(x_s)$, the black-box will query the next point x_{t+1}. We assume that $x_1 = 0$ and that:

$$x_{t+1} \in \operatorname{Span}(g_1, \ldots, g_t). \tag{6}$$

And we will prove lower bounds with respect to convex optimization procedures that satisfy this black-box model. Note that this means that these bounds do not hold for randomized algorithms where the next query point x_{t+1}, being chosen randomly, may not necessarily fall into the span of all the previous gradients.

The general proof technique will be to produce an optimization problem and an oracle such that after t queries (for a fixed t), we will always have the lower bound:

$$\min_{1 \leq s \leq t} f(x_s) - f(x^*) > L.$$

In this lecture, we'll just consider one such lower bound for f that is β-smooth (it is possible to extend the analysis for this to f both α-strongly convex and β-smooth). Our goal is the following theorem:

Theorem 15 (Theorem 3.14, [B2015]). *Let $t \leq (n-1)/2$ and $\beta > 0$. Then there exists a β-smooth convex function f such that for any black-box procedure satisfying Equation 6,*

$$\min_{1 \leq s \leq t} f(x_s) - f(x^*) \geq \frac{3\beta}{32} \frac{\|x_1 - x^*\|^2}{(t+1)^2}.$$

This theorem shows that under the black-box model, there is an oracle complexity lower bound of $\Omega\left(\frac{1}{\sqrt{\epsilon}}\right)$. This is in contrast to the upper bound from gradient descent we saw last lecture of $O\left(\frac{1}{\epsilon}\right)$. [Question: is this gap between the lower bound and upper bound for gradient descent because gradient descent is not strong enough? Or because the analysis is not tight enough?]

Theorem 16 (Theorem 3.3, [B2015]). *Let f be convex and β-smooth. Gradient descent satisfies*

$$f(x_t) - f(x^*) \leq \frac{2\beta \|x_1 - x^*\|^2}{t-1}.$$

Likewise, there is a gap for gradient descent for f that is both α-strongly convex and β-smooth. The lower bound is $\Omega\left(\sqrt{\kappa} \ln \frac{1}{\epsilon}\right)$, following:

Theorem 17 (Theorem 3.15, [B2015]). *Let $\kappa > 1$. There exists an α-strongly convex and β-smooth function $f : \ell_2 \to \mathbb{R}$ with $\kappa = \beta/\alpha$ such that for all $t \geq 1$ and black-box procedure satisfying Equation 6,*

$$f(x_t) - f(x^*) \geq \frac{\alpha}{2} \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^{2(t-1)} \|x_1 - x^*\|^2.$$

We saw in Theorem 10 that the gradient descent achieves convergence rate $O\left(\kappa \ln \frac{1}{\epsilon}\right)$, while geometric descent closes the gap and achieves the lower bound.

In this lecture, we'll just consider the setup and approach to proving Theorem 15.

4.1 Example lower bound

Consider the following 1-dimensional mass-spring optimization problem. Assume that there are n masses connected in series by springs. Thus, they form the path graph on n vertices. Using springs, we'll connect the first and last masses to a wall at the origin.

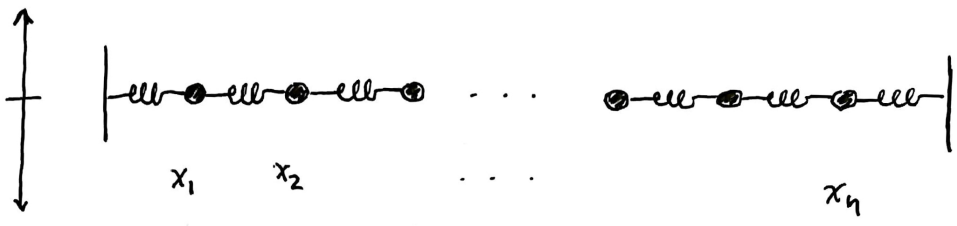

Figure 6: A visualization of the mass-spring system at rest. Note that for clarity, we've expanded the springs on a different axis (they should technically be all on top of each other). The problem is a 1-dimensional spring system where the axis of motion is up and down.

We will apply a constant force upwards on the first mass (e.g. imagine the force of gravity acting only on the first mass). Our goal will be to minimize the energy of the spring system by placing the mass along the

up-and-down direction. Let $x \in \mathbb{R}^n$ be denote the configuration of this system, where x_i is the position of the ith mass. We will attempt to achieve this through any black-box model with the assumption that the first query to the oracle is $x^{(0)} = 0$ and that the tth query $x^{(t)}$ is contained in $\text{Span}(g_1, \ldots, g_{t-1})$.

The energy of the above mass-spring system in configuration x can be defined by:

$$E(x) = \frac{1}{2} x^T A x,$$

where $A = L + \delta_{11} + \delta_{nn}$. Here, L is the Laplacian of the path graph and δ_{ii} is the matrix of all 0's except for the (i, i)-th entry, which is 1. Recall that:

$$x^T L x = \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2$$

gives the sum of all the squared-lengths of the stretched springs (and energy is proportional to the squared-lengths). The additional δ_{11} and δ_{nn} terms captures the energy stored in the springs attaching the first and last masses to the wall. In other words, A is the following tri-diagonal matrix:

$$A = \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & & & & \\ & & & \ddots & & \\ & & & & & -1 \\ & & & & -1 & 2 \end{bmatrix}$$

Now, we define our objective function:

$$f(x) = \frac{\beta}{4} \left[\frac{1}{2} x^T A x - e_1^T x \right],$$

which corresponds to the energy of the spring system (scaled by $\beta/4$ so as to ensure that f is β-smooth) where there is an upward force on the first mass.

Claim 18. *If the $k+1, \ldots, n$th vertices are fixed at 0, then the minimizer subject to this constraint satisfies:*

$$x_i^* = 1 - \frac{i}{k+1}.$$

This corresponds to the physical intuition that if we were to lift the first mass and hold all the masses from $k+1$ on at 0, then the rest of the masses would distribute themselves to decrease at equal intervals, as in Figure 7. Now, if we were optimizing f in the black-box approach, then at step s, we would only be allowed to shift the first s masses away from 0. We know what configuration given these constraints would minimize f, and so it is possible to compute lower bounds on $f(x^{(s)})$. We could also compute $f(x^*)$ by the same formula. Straightforward algebra is performed to show Theorem 15. See [B2015] for details.

The last thing to note is that f is indeed β-smooth, just by considering the Hessian of f and noting that:

$$x^T A x = 2 \|x\|^2 - 2 \sum_{i=1}^{k-1} x_i x_{i+1} \leq 4 \|x\|^2.$$

References

[B2015] Bubeck, S. *Convex optimization: algorithms and complexity.* Foundations and Trends in Machine Learning. (2015).

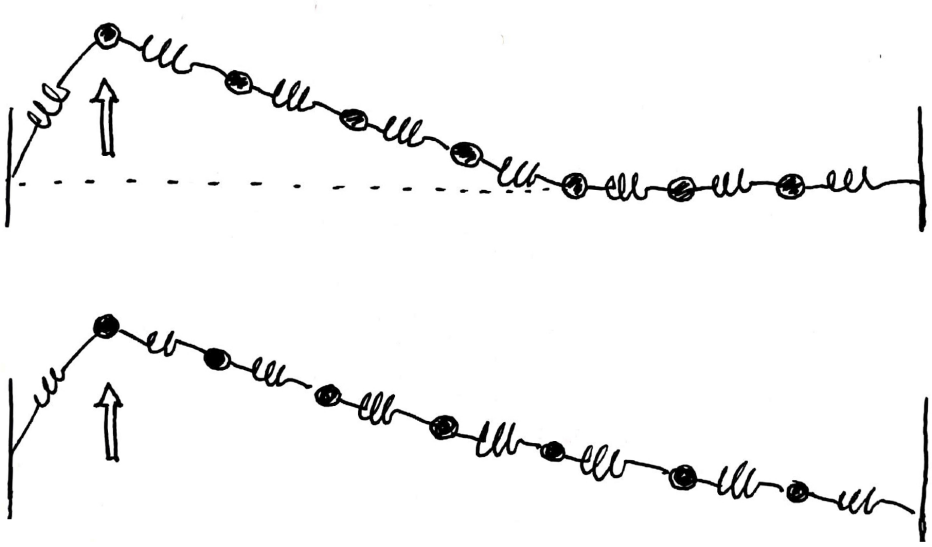

Figure 7: Top: an upward force is applied to the first mass, while the $k+1$ to nth masses are fixed at 0. The configuration that minimizes energy subject to this constraint is where the intermediate masses arrange themselves linearly. Bottom: it follows that if there is no constraint, then all the masses would arrange themselves linearly. This corresponds to x^*.