CPSC 536C: Algorithms for Convex Optimization
University of British Columbia

Gradient Descent

Akshay Ram
February 2, 2025

These notes are subject to change and may contain errors.

1 Introduction

The goal of this note is to study and analyze the following simple procedure for unconstrained convex optimization:

$$x_{t+1} := x_t - \eta_t \nabla f(x_t).$$

Here and in the remainder we assume f is differentiable (so $\nabla f(x)$ is well-defined). We note that much of the analysis carries over to the non-differentiable convex functions by using any sub-gradient.

In contrast to cutting plane methods, these gradient methods usually have $\text{poly}(1/\varepsilon)$ convergence to an ε-approximate solution, but have much better dependence on the problem dimension n. At the end we will also show that such dependencies are necessary by constructing and analyzing certain hard instances for gradient descent.

1.1 Approximate Solutions

We consider more formally what it means to produce an approximate solution:

Definition 1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable convex function with optimizer $x^* := \arg\min_x f(x)$.*

- *x has ε-function gap if $f(x) \le f(x^*) + \varepsilon$;*
- *x is ε-close if $\|x - x^*\|_2 \le \varepsilon$;*
- *x is ε-critical if $\|\nabla f(x)\|_2 \le \varepsilon$.*

These three notions of approximate solutions are in general incomparable, and their importance depends on the application at hand.

Exercise 1. *Compute the function gap, distance to opt, and gradient for the following:*

1. *$f(x) := \|x\|_2^2/2$;*
2. *$f(x) := \|Ax - b\|_2^2$;*
3. *$f(Q) := -\log\det(Q) + \sum_i \lambda_i \langle x_i, Qx_i\rangle$.*

1.2 Assumptions

We need some assumptions on our initial iterate and the function in order to guarantee any reasonable approximation. The following are natural assumptions satisfied by many functions in practice.

Definition 2. *$f : \mathbb{R}^n \to \mathbb{R}$ is G-Lipschitz if*

$$\forall x, y : \quad |f(y) - f(x)| \leq G\|y - x\|_2.$$

This intuitively allows us to relate f to an affine function with slope G. We can also relate f to quadratic functions, which are the next simplest convex functions. Recall the first-order definition of convexity:

$$\forall x, y : \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle.$$

Definition 3. *$f : \mathbb{R}^n \to \mathbb{R}$ is α-strongly convex if*

$$\forall x, y : \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|_2^2.$$

f is β-smooth if

$$\forall x, y : \quad f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2}\|y - x\|_2^2.$$

Note that the first gives a stronger lower bound than standard convexity, whereas the smoothness condition gives an upper bound. In both cases this allows us to compare our function to a convex quadratic, which is much simpler to analyze.

Exercise 2.

1. *If f is α-strongly convex, then it is also α'-strongly convex for any $\alpha' \leq \alpha$. I.e. strong convexity is an increasingly strong assumption for α increasing;*
2. *If f is β-smooth, then it is also β'-strongly convex for any $\beta' \geq \beta$. I.e. smoothness is a weaker assumption for β increasing;*
3. *If f, g are α_f, α_g-strongly convex and β_f, β_g-smooth, respectively, then $f+g$ is $\alpha_f+\alpha_g$-strongly convex and $\beta_f + \beta_g$-smooth;*

Claim 4. *For quadratic $q_\gamma(x) := q_0 + \langle b, x - x_0 \rangle + \frac{\gamma}{2}\|x - x_0\|_2^2$, we can optimize*

$$\min_x q_\gamma(x) = q_0 - \frac{\|b\|_2^2}{2\gamma} \quad \textit{with} \quad x^* := \arg\min_x q_\gamma(x) = x_0 - \frac{b}{\gamma}.$$

We leave the proof as an exercise (show directly by computing gradient and setting to 0).

The point of the above assumptions are that they allow us to relate various notions of approximate solutions.

Fact 5. *Let x^* be the optimizer of f, which is (1) G-Lipschitz, (2) β-smooth, and (3) α-strongly convex, respectively. Then*

1. *$f(x) - f(x^*) \leq G\|x - x^*\|_2$;*
2. *$f(x) - f(x^*) \leq \frac{\beta}{2}\|x - x^*\|_2^2$;*
3. *$f(x) - f(x^*) \leq \frac{\|\nabla f(x)\|_2^2}{2\alpha}$ and $f(x) - f(x^*) \geq \frac{\alpha}{2}\|x - x^*\|_2^2$.*

Proof:

1. Follows directly from the definition;

2. Apply the definition of smoothness between x and x^*:

$$f(x) \leq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + \frac{\beta}{2}\|x - x^*\|_2^2 = f(x^*) + \frac{\beta}{2}\|x - x^*\|_2^2,$$

   where in the last step we used $\nabla f(x^*) = 0$ since x^* is the optimizer.

3. Rearranging, we see that this is equivalent to a lower bound on $f(x^*)$ in terms of the gradient norm. For this we use our strong convexity assumption

$$f(x^*) \geq q_\alpha(x^*) := f(x) + \langle \nabla f(x), x^* - x \rangle + \frac{\alpha}{2}\|x^* - x\|_2^2.$$

   Then by theorem 4 we have

$$q_\alpha(x^*) \geq \min_y q_\alpha(y) = f(x) - \frac{\|\nabla f(x)\|_2^2}{2\alpha}.$$

   The result follows by the lower bound $f(x^*) \geq q_\alpha(x^*)$ and rearranging.

   For the lower bound, we can simply apply the definition

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + \frac{\alpha}{2}\|x - x^*\|_2^2 = f(x^*) + \frac{\alpha}{2}\|x - x^*\|_2^2$$

   as $\nabla f(x^*) = 0$ at the optimizer.

□

2 Analysis of Gradient Descent

Intuitively, if we have a simple quadratic function, theorem 4 shows how to find the optimum directly. In general, convexity (and strong convexity and smoothness) allows us to approximate our function f by an affine or quadratic function. The algorithm and analysis of gradient involves using these proxies to compute updates and reason about the progress made.

Theorem 6. *Let f be α-strongly convex and β-smooth, and consider update*

$$x_{t+1} := x_t - \frac{1}{\beta}\nabla f(x_t).$$

Then the function gap of iteration T can be bounded as

$$f(x_T) - f(x^*) \leq \left(1 - \frac{\alpha}{\beta}\right)^T (f(x_0) - f(x^*)),$$

*i.e. we achieve ε-*relative *function gap in $O(\kappa \log(1/\varepsilon))$ iterations, where $\kappa := \beta/\alpha$.*

Proof: The update step can be derived as the optimizer of the proxy given by smoothness:

$$x_{t+1} = \arg\min_x q_\beta(x) := \arg\min_x f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{\beta}{2}\|x - x_t\|_2^2,$$

where the minimum calculation follows by theorem 4. By smoothness we have $f \le q_\beta$, so we can show significant progress in function value:

$$f(x_{t+1}) \le q_\beta(x_{t+1}) = f(x_t) - \frac{\|\nabla f(x_t)\|_2^2}{2\beta},$$

where again the last step was by theorem 4. Next, we use strong convexity to show that this gives a significant improvement:

$$\frac{\|\nabla f(x_t)\|_2^2}{2\alpha} \ge f(x_t) - f(x^*),$$

where we applied theorem 5(3) for α-strongly convex f. Putting this together gives

$$\begin{aligned} f(x_{t+1}) - f(x^*) &= (f(x_{t+1}) - f(x_t)) + (f(x_t) - f(x^*)) \\ &\le -\frac{\|\nabla f(x_t)\|_2^2}{2\beta} + (f(x_t) - f(x^*)) \le (1 - \alpha/\beta)(f(x_t) - f(x^*)), \end{aligned}$$

where in the second step we used the upper bound for the update, and in the final step we used that theorem 5(3) to compare the gradient norm and $f(x_t) - f(x^*)$. □

Note that the above gives exponential convergence in terms of function gap. Also by theorem 5 we can show convergence in terms of the solution $\|x_t - x^*\|_2$ and gradient $\|\nabla f(x_t)\|_2$. The above setting is in some sense the most structured, since we have upper and lower bounds on f via smoothness and strong convexity. In the next settings, we remove the strong convexity assumption.

Theorem 7. *Let f be β-smooth and assume $\|x_0 - x^*\|_2 \le R$. Then for update*

$$x_{t+1} := x_t - \frac{\nabla f(x_t)}{\beta},$$

the function gap can be bounded as

$$f(x_T) - f(x^*) = \min_{t \in [T]} f(x_t) - f(x^*) \lesssim \frac{\beta R^2}{T},$$

*i.e. we achieve ε-*relative *function gap in $O(1/\varepsilon)$ iterations.*

Proof: By theorem 4 we argue that

$$x_{t+1} = \arg\min_x q_\beta(x) := \arg\min_x f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{\beta}{2}\|x - x_t\|_2^2.$$

By the smoothness condition, this guarantees improvement in the objective

$$f(x_{t+1}) \le q_\beta(x_{t+1}) = f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{\beta}{2}\|x_{t+1} - x_t\|_2^2 = f(x_t) - \frac{\|\nabla f(x_t)\|_2^2}{2\beta}.$$

In order to argue about optimality gap, we use β-strong convexity of q_β: for arbitrary x we have

$$\begin{aligned} f(x) + \frac{\beta}{2}\|x - x_t\|_2^2 &\ge f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{\beta}{2}\|x - x_t\|_2^2 = q_\beta(x) \\ &\ge q_\beta(x_{t+1}) + \langle \nabla q_\beta(x_{t+1}), x - x_{t+1} \rangle + \frac{\beta}{2}\|x - x_{t+1}\|_2^2 \\ &\ge f(x_{t+1}) + \frac{\beta}{2}\|x - x_{t+1}\|_2^2, \end{aligned}$$

where the first step was by convexity of f, the third step follows by β-strong convexity of q_β (as q_β is the sum of an affine function and $\frac{\beta}{2}\|x^* - x_t\|_2^2$), along with the fact that x_{t+1} is the optimizer of q_β so the gradient term vanishes, and in the last step we used β-smoothness condition to upper bound $f(x_{t+1})$. Plugging in $x = x^*$ above and rearranging gives

$$f(x_{t+1}) - f(x^*) \leq \frac{\beta}{2}(\|x^* - x_t\|_2^2 - \|x^* - x_{t+1}\|_2^2).$$

Then we can apply a telescoping argument to show

$$\frac{1}{T}\sum_{t=0}^{T-1}(f(x_{t+1}) - f(x^*)) \leq \frac{1}{T}\sum_{t=0}^{T-1}\frac{\beta}{2}(\|x^* - x_t\|_2^2 - \|x^* - x_{t+1}\|_2^2) \leq \frac{\beta\|x_0 - x^*\|_2^2}{T},$$

where in the last step we used telescoping and $\|x_T - x^*\|_2 \geq 0$. Finally, to argue about the last iterate, we note that plugging in $x = x_t$ into the bound above gives

$$f(x_t) = f(x_t) + \frac{\beta}{2}\|x_t - x_t\|_2^2 \geq f(x_{t+1}) + \frac{\beta}{2}\|x_t - x_{t+1}\|_2^2 \geq f(x_{t+1}),$$

i.e. the function value is non-increasing, and therefore the last iteration is minimum. □

The next case is the least structured, and will have much worse guarantees.

Theorem 8. *Let f be G-Lipschitz and assume $\|x_0 - x^*\|_2 \leq R$. Then for update*

$$x_{t+1} := x_t - \eta\nabla f(x_t) \quad \textit{where} \quad \eta := \sqrt{\frac{R^2}{G^2 T}},$$

the function gap can be bounded as

$$\min_{t\in[T]} f(x_t) - f(x^*) \lesssim \sqrt{\frac{R^2G^2}{T}},$$

*i.e. we achieve ε-*relative *function gap in $O(1/\varepsilon^2)$ iterations.*

Proof: By the same theorem 4 we argue that

$$x_{t+1} = \arg\min_x q_\eta(x) := \arg\min_x f(x_t) + \langle\nabla f(x_t), x - x_t\rangle + \frac{1}{2\eta}\|x - x_t\|_2^2.$$

But without smoothness this does not necessarily give us a direct guarantee for our function. In fact, the update may cause an increase in the objective function. We instead use convexity to argue that each iteration guarantees either function improvement or distance improvement:

$$\begin{aligned} f(x^*) + \frac{1}{2\eta}\|x^* - x_t\|_2^2 &\geq f(x_t) + \langle\nabla f(x_t), x^* - x_t\rangle + \frac{1}{2\eta}\|x^* - x_t\|_2^2 = q_\eta(x^*) \\ &\geq q_\eta(x_{t+1}) + \langle\nabla q_\eta(x_{t+1}), x^* - x_{t+1}\rangle + \frac{1}{2\eta}\|x^* - x_{t+1}\|_2^2 \\ &= f(x_t) - \eta\frac{\|\nabla f(x_t)\|_2^2}{2} + \frac{1}{2\eta}\|x^* - x_{t+1}\|_2^2, \end{aligned}$$

where the first step was by convexity of f, the third step follows by $1/\eta$-strong convexity of q_η (which can be proven directly or using that q_η is the sum of convex f and $1/\eta$-strongly convex

$\frac{1}{2\eta}\|x^* - x_t\|_2^2$), and in the final step we used theorem 4 for the proxy q_η which is optimized at x_{t+1}. Rearranging, we have a bound on the function gap in each iteration:

$$f(x_t) - f(x^*) \leq \eta \frac{\|\nabla f(x_t)\|_2^2}{2} + \frac{1}{2\eta}(\|x^* - x_t\|_2^2 - \|x^* - x_{t+1}\|_2^2).$$

We can bound the first term using the Lipschitz condition. For the second term, note that this is large iff x_{t+1} gets much closer to the optimum x^* than x_t; since we have an initial bound $\|x_0 - x^*\|_2 \leq R$, this term is bounded across all iterations. And finally, by the Lipschitz condition, if $\|x_t - x^*\|_2$ is small, then we can argue that the function gap is small by theorem 5(1).

We will first argue about the average function gap:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) - f(x^*) &\leq \frac{1}{T}\sum_{t=0}^{T-1} \frac{\eta\|\nabla f(x_t)\|_2^2}{2} + \frac{1}{T}\sum_{t=0}^{T-1} \frac{\|x^* - x_t\|_2^2 - \|x^* - x_{t+1}\|_2^2}{2\eta} \\ &\leq \frac{\eta G^2}{2} + \frac{\|x^* - x_0\|_2^2 - \|x^* - x_T\|_2^2}{2\eta T} \leq \frac{\eta G^2}{2} + \frac{R^2}{2\eta T},\end{aligned}$$

where the first step was by the calculation above for each iterate, in the second step we bounded each gradient $\|\nabla f(x_t)\|_2 \leq G$ by the Lipschitz condition, and we bounded the sum by noting it is a telescoping series, and in the final step we used that $\|x^* - x_T\|_2 \geq 0$ and our initial bound $\|x_0 - x^*\|_2 \leq R$. Finally we can choose $\eta = \sqrt{\frac{R^2}{G^2 T}}$ to balance terms. Therefore we can bound the minimum function gap over all iterations

$$\min_t f(x_t) - f(x^*) \leq \frac{1}{T}\sum_{t=0}^{T-1} f(x_t) - f(x^*) \leq \sqrt{\frac{R^2 G^2}{T}}.$$

We can also use convexity to bound the function gap for the average iterate $\overline{x} := \frac{1}{T}\sum_{t=0}^{T-1} x_t$:

$$f(\overline{x}) - f(x^*) \leq \frac{1}{T}\sum_{t=0}^{T-1} f(x_t) - f(x^*) \leq \sqrt{\frac{R^2 G^2}{T}},$$

where the first step is by Jensen's inequality. □

We note that the step-size above requires knowledge of the time horizon T. In order to avoid this, we could consider a different step-size $\eta_t \simeq 1/\sqrt{t}$. The algorithm for the strongly convex and Lipschitz case also uses a similar time-varying step-size.

Theorem 9. *Let f be α-strongly convex and G-Lipschitz. Then for update*

$$x_{t+1} - \eta_t \nabla f(x_t) \quad \textit{with} \quad \eta_t \propto 1/\alpha t,$$

there is an appropriately averaged iterate $\bar{x} \in \mathrm{conv}\{x_t\}$ such that

$$f(\bar{x}) - f(x^*) \lesssim \frac{G^2}{\alpha T}.$$

Proof: We first use strong convexity to get a better lower bound for the function:

$$f(x) + \frac{\eta_t^{-1} - \alpha}{2}\|x - x_t\|_2^2 \geq f(x_t) + \langle \nabla f(x_t), x - x_t\rangle + \frac{1}{2\eta_t}\|x - x_t\|_2^2.$$

We can now repeat the rest of the argument to get

$$f(x) + \frac{\eta_t^{-1} - \alpha}{2}\|x - x_t\|_2^2 \geq f(x_t) - \frac{\eta_t \|\nabla f(x_t)\|_2^2}{2} + \frac{1}{2\eta_t}\|x - x_{t+1}\|_2^2$$

Rearranging this gives bound

$$f(x_t) - f(x) \leq \frac{\eta_t \|\nabla f(x_t)\|_2^2}{2} + \frac{\eta_t^{-1} - \alpha}{2}\|x - x_t\|_2^2 - \frac{\eta_t^{-1}}{2}\|x - x_{t+1}\|_2^2.$$

In order to apply telescoping when we sum up these terms, we need $\eta_t^{-1} - \alpha = \eta_{t-1}^{-1}$. We want $\eta_1^{-1} = \alpha$ so that the first error term $D(x, x_1)$ vanishes. The choice $\eta_t^{-1} = \alpha(t+1)$ satisfies these properties so we get

$$\begin{aligned}\sum_{t=1}^{T} f(x_t) - f(x) &\leq 0 + \sum_{t\geq 2}^{T-1} \frac{\|x - x_t\|_2^2}{2}(-\alpha t + \alpha(t+1) - \alpha 1) + \sum_{t=1}^{T-1} \frac{\|\nabla f(x_t)\|_2^2}{2\alpha(t+1)} \\ &\leq \frac{1}{\alpha}\sum_{t=1}^{T-1} \frac{G^2}{2(t+1)} \lesssim G^2 \log T,\end{aligned}$$

where $\eta_1^{-1} - \alpha = 0$ and $\|x - x_T\|_2 \geq 0$ so all distance terms cancel, in the second step we use the Lipschitz condition to bound the gradient, and in the final step we use $\sum_t 1/t \approx \log T$. Taking the average would give the stated convergence result up to a $\log T$ factor.

In the online setting, where we are given a possibly different adversarially chosen f_t in each iteration, it is known that this result is tight, i.e. the $\log T$ term is necessary and the average regret diverges for $T \to \infty$. On the other hand, note that this update step does not require knowledge of the time horizon.

In our simpler setting where there is only a single objective function f, we can choose a slightly different step-size and averaging argument. Namely, we bound

$$\sum_{t=1}^{T} t(f(x_t) - f(x^*)) \leq \sum_{t=1}^{T} t\eta_t \frac{\|\nabla f(x_t)\|_2^2}{2} + \sum_{t=1} t(\frac{\eta_t^{-1} - \alpha}{2}\|x - x_t\|_2^2 - \frac{\eta_t^{-1}}{2}\|x - x_{t+1}\|_2^2.$$

Now in order for the telescoping to work, we still want $\eta_1^{-1} = \alpha$, but now we want $t(\eta_t^{-1} - \alpha) = (t-1)\eta_{t-1}^{-1}$. For this we choose $\eta_t^{-1} = \alpha(t+1)/2$ which gives

$$\sum_{t=1}^{T} t(f(x_t) - f(x^*)) \leq \sum_{t} t\frac{\|\nabla f(x_t)\|_2^2}{\alpha(t+1)} + \sum_{t\geq 2} \frac{\|x - x_t\|_2^2}{2}\bigg(- (t-1)\frac{\alpha t}{2} + \frac{\alpha(t+1) - 2\alpha}{2}\bigg) \lesssim T\frac{G^2}{\alpha},$$

where again all the distance terms cancel by our choice of η_t, and we apply the G-Lipschitz condition to bound the gradient. Applying Jensen's inequality to $\bar{x} = \sum_t (2t/T(T+1))x_t$ gives the final bound

$$f(\bar{x}) - f(x^*) \leq \sum_t \frac{2t}{T(T+1)}(f(x_t) - f(x^*)) \lesssim \frac{G^2}{\alpha T}.$$

□

In the next section we show that the guarantee for the Lipschitz setting is optimal! I.e. there exists worst-case instances such that *any* algorithm with access to function evaluation and gradients must have iteration dependence of the form $1/\varepsilon^2$. On the other hand, the guarantee for the first two cases, those involving smoothness, are surprisingly not optimal.

3 Lower Bounds

Theorem 10 (Theorem 3.13 in Bubeck)**.** *For any $T \leq n$ and any $G > 0$, there exists a convex function $f : \mathbb{R}^n \to \mathbb{R}$ that is G-Lipschitz such that any black-box algorithm with access to function evaluation and (sub-)gradients must satisfy*

$$\min_{t\in[T]} f(x_t) \geq \min_{x\in RB_2^n} f(x) + \Omega\left(\frac{RG}{\sqrt{T}}\right).$$

Theorem 11 (Theorem 3.14 in Bubeck)**.** *For any $T < (n-1)/2$ there exists a (quadratic) convex function f that is β-smooth, such that any black-box algorithm with access to function evaluation and gradients must satisfy*

$$\min_{t\in[T]} f(x_t) - f(x^*) \gtrsim \frac{\beta\|x_0 - x^*\|_2^2}{T^2}.$$

We make some remarks about the parameters before the proof. Note the assumptions that $n \gtrsim T$ in both cases. These are necessary, as for $T \gg n$ we could apply the cutting plane methods of the previous note to achieve exponential convergence, so there cannot be any lower bound of the form $1/\sqrt{T}$ or $1/T^2$ as stated. Also note that the second lower bound in theorem 11 has convergence rate $1/T^2$, whereas the analysis of gradient descent given in theorem 7 has rate $1/T$. It turns out that there is actually an improved algorithm, known as accelerated gradient descent, which has a quadratically better convergence rate, matching the lower bound. We will discuss this in a later lecture.

From a technical perspective, the hard instances given in the above results are somewhat simple and can in fact be solved exactly in $O(n)$ iterations. The key will be to only reveal information about one new coordinate in each iteration; therefore any black-box algorithm will not be able to 'see' the function on the whole space until $\Omega(n)$ iterations.

Proof: [Proof of theorem 10] For this construction, we will use a non-differentiable function; therefore the requirement for the oracle is $\text{GRAD}(f, x) \in \partial f(x)$, i.e. the sub-gradient outputs an arbitrary element of the sub-gradient set at x. This freedom will be useful for us, as we will design an adversarial oracle in order to 'hide' as much information about the function as possible. We also focus on the case of $T = n - 1, G = 1, R = 1$ and discuss the simple extensions to the general parameters at the end.

The instance is as follows:

$$f(x) := \max_{i\in[n]} x_i.$$

Note that this is clearly convex since it is a max of linear functions, and it can directly verified that it is 1-Lipschitz.

Next we design an adversarial subgradient oracle. The subdifferential at x can be computed as

$$\partial f(x) = \text{conv}\{e_i \mid i \in [n], x_i = f(x) = \max_{j\in[n]} x_j\}.$$

The oracle $\text{GRAD}(f, x)$ will output e_i such that i is the first maximizing index $x_i = f(x), x_{j<i} < f(x)$. This allows us to 'hide' information as we reveal at most one new coordinate in each iteration.

We next calculate the optimizer over the ball: it is clear that for $x^* := -\vec{1}/\sqrt{n}$ we have $\|x^*\|_2^2 = 1$ and $f(x^*) = -1/\sqrt{n}$; we show that this is optimum as for $x \in B_2^n$:

$$\min_{i\in[n]} |x_i|^2 \leq \frac{\sum_{i\in[n]} |x_i|^2}{n} \leq \frac{1}{n} \implies \min_{\|x\|_2\leq 1} f(x) \geq -\min_{i\in[n]} |x_i| \geq \frac{-1}{\sqrt{n}}.$$

For the sake of our analysis, we assume $x_0 = 0$ and the algorithm only queries points that are in the span of points and gradients seen so far:

$$x_t \in \text{span}\{g_0, ..., g_{t-1}\}$$

where g_s is the output of the sub-gradient oracle for query x_s. This is a technical assumption that can be removed by slightly changing the instance (e.g. consider different linear functions instead of coordinates), but we leave this extension to the reader. Now we claim that $\text{supp}(x_t) \in [t]$ for all $t < n$. This can be shown by induction, as $x_0 = 0, g_0 = e_1$, so therefore $x_1 \in \text{span}\{e_1\} \implies \text{supp}(x_1) \subseteq [1], g_1 = e_2$; then in the next iteration $x_2 \in \text{span}\{e_1, e_2\} \implies \text{supp}(x_2) \subseteq [2], g_2 = e_3$; and so on. While $\text{supp}(x_t) \subsetneq [n]$ there must be some zero coordinate, so $f(x_t) \geq 0$ for all $t < n$. Putting this together with our optimum calculation, we have

$$\min_{t\in[T]} f(x_t) - \min_{x\in B_2^n} f(x) \geq 0 - \frac{1}{\sqrt{n}},$$

which is $\Omega(1/\sqrt{T})$ for $T = \Omega(n)$.

The extension to all $T < n$ can be accomplished by modifying the function $\tilde{f}(x) := \max_{i\in[T+1]} x_i$; similarly if we want the function to be G-Lipschitz we can consider $\tilde{f} := Gf$, and the analysis for the optimizer over $R \cdot B_2^n$ is the same with an additional R factor. □

Proof: [Proof of theorem 11] We focus on the case of $T = \Omega(n), \beta = O(1), R = O(1)$ and leave the simple extensions to the general parameters to the reader. For this construction, we will use a simple quadratic function:

$$f(x) := \frac{1}{2}\langle x, Ax\rangle - \langle x, e_1\rangle \qquad \text{where} \quad A_{ij} = \begin{cases} 2 & i = j \\ -1 & |i-j| = 1 \\ 0 & \text{otherwise} \end{cases}.$$

For some intuition, the matrix A is a small modification of the Laplacian of the path graph. We next show this function is convex:

$$\langle x, Ax\rangle = |x_1|^2 + |x_n|^2 + \sum_{i\in[n-1]} (x_{i+1} - x_i)^2,$$

which is clearly positive for all $x \neq 0$, i.e. $A \succ 0$ which implies f is (strictly) convex. We can also verify smoothness by bounding the quadratic

$$\langle x, Ax\rangle \leq |x_1|^2 + |x_n|^2 + \sum_{i\in[n-1]} (2x_{i+1}^2 + 2x_i^2) \leq O(1)\|x\|_2^2,$$

so $A \preceq O(1)I$, i.e. f is $O(1)$-smooth. f is differentiable so

$$\text{GRAD}(f, x) = \nabla f(x) = Ax - e_1.$$

Note that if $\text{supp}(x) \subseteq [k]$ then $\text{supp}(\nabla f(x)) = \text{supp}(Ax - e_1) \subseteq [k+1]$, since A has non-zero entries only on one sub- and super-diagonal. This intuitively gives the same consequence as the previous proof, that the algorithm will only see one dimension each iteration, so will not be able to find the optimum globally for $\Omega(n)$ iterations.

We next need to calculate the optimizer. We will also need optimizers with some given support, so for shorthand we let $A^{[k]}$ be the top $k\times k$ principal submatrix of A and $f^k(x) := \frac{1}{2}\langle x, A^k x\rangle - \langle x, e_1\rangle$,

where we assume $\text{supp}(x) \subseteq [k]$ by abuse of notation in this definition. All of these functions are (strictly) convex, so we can compute the optimizer using the critical equation

$$x^k = \arg\min_x f^k(x) \quad \iff \quad \nabla f^k(x^k) = A^k x^k - e_1 = 0.$$

We claim that A^k is invertible, so this equation has a unique solution. Explicitly $x^k_i := 1 - \frac{i}{k+1}$ as can be verified directly, and we can bound the norm

$$\|x^k\|_2^2 = \sum_{i=1}^{k} \frac{(k+1-i)^2}{(k+1)^2} = \frac{1}{(k+1)^2} \sum_{j=1}^{k} j^2 \simeq k.$$

At this point, the value of the objective is

$$f(x^k) = f^k(x^k) = \frac{1}{2}\langle x^k, A^k x^k - e_1\rangle - \frac{1}{2}\langle x^k, e_1\rangle = -\frac{1}{2}\frac{k}{k+1},$$

where in the first step we used $\text{supp}(x^k) = [k]$, and in the last step we used that $A^k x^k - e_1 = 0$.

To simplify our analysis, we assume $x_0 = 0$ and the algorithm only queries points that are in the span of points and gradients seen so far:

$$x_t \in \text{span}\{\nabla f(x_0), ..., \nabla f(x_{t-1})\}.$$

Now we claim that $\text{supp}(x_t) \in [t]$ for all $t < n$. This can be shown by induction and the observation $\text{supp}(x) \subseteq [k] \implies \text{supp}(\nabla f(x)) = \text{supp}(Ax - e_1) \subseteq [k+1]$. Therefore by iteration $T < n$, we have

$$\min_{t \in [T]} f(x_t) \geq \min_{\text{supp}(x) \subseteq [T]} f(x) = f(x^T) = -\frac{T}{2(T+1)} \geq f(x^n) = -\frac{n}{2(n+1)},$$

which gives the required gap of $T \geq \Omega(n)$ as $\|x_0 - x^n\|_2^2 \simeq n$. □

The above two constructions can be simply modified (by scaling by a constant and adding α times a quadratic) in order to give similarly tight lower bounds for the strongly convex setting. We leave this extension to the reader.

4 Projected Gradient Descent

So far we have been considering the unconstrained convex minimization problem. In this section we will describe a simple algorithm that allows us to deal with convex constraints:

$$\min f(x) \qquad \text{s.t.} \quad x \in K.$$

A natural update using the gradient could be $x_{t+1} = x_t - \eta \nabla f(x_t)$, where we choose step-size $\eta > 0$ such that $x_{t+1} \in K$. Simple examples show that this could lead to the algorithm getting stuck at the boundary of the constraint set K. The actual update involves following the gradient, but then *projecting back* to the constraint set.

Definition 12. *P_K is the projection operator for convex set $K \subseteq \mathbb{R}^n$:*

$$P_K(x) := \arg\min_{y \in K} \|y - x\|_2.$$

Note that this can equivalently be written in terms of $\|y - x\|_2^2$, which is a *strongly convex* function; therefore the optimum value is always uniquely attained.

Exercise 3. *If f is α-strongly convex for any $\alpha > 0$, then for any closed set C, the infimum $\arg\min_{x\in C} f(x)$ is always attained, and furthermore it is unique.*

With this in hand we can present the algorithm for constrained convex minimization.

Definition 13 (Projected Gradient descent)**.** *In each iteration we update*

$$\bar{x}_{t+1} := x_t - \eta \nabla f(x_t); \qquad x_{t+1} := P_K(\bar{x}_{t+1}) = \arg\min_{x\in K} \|x - \bar{x}_{t+1}\|_2^2.$$

In order to understand the projection operator better, we will analyze the optimality conditions given by the $\|y - x\|_2^2$ convex program.

Lemma 14 (Optimality conditions)**.** *Let f be a convex differentiable function and $K \subseteq \mathbb{R}^n$ a convex set. Then*

$$x^* = \arg\min_{x\in K} f(x) \iff \forall x \in K : \langle \nabla f(x^*), x - x^* \rangle \geq 0.$$

Proof: The gradient condition in some sense corresponds to local optimality, so we can show

$$\forall x \in K : \quad f(x) - f(x^*) \geq \langle \nabla f(x^*), x - x^* \rangle \geq 0,$$

i.e. x^* is optimal over K.

For the converse, assume for contradiction that $\exists x \in K : \nabla f(x^*), x - x^* \rangle < 0$. By convexity we have $x_\varepsilon := (1 - \varepsilon)x^* + \varepsilon x \in K$ so we have

$$f(x_\varepsilon) = f(x^*) + \varepsilon \langle x - x^*, \nabla f(x^*) \rangle + o(\varepsilon) < f(x^*)$$

where in the first step we used differentiability of f and in the last step we used $\nabla f(x^*), x - x^* \rangle < 0$ and the first term dominates for small ε. This contradicts optimality of x^*.

In order to analyze the algorithm, we will re-interpret this two-step update in terms of a constrained version of the quadratic proxy update used in gradient descent (see theorem 4):

Claim 15. *The above update can be equivalently written as*

$$x_{t+1} = \arg\min_{x\in K} q(x) := f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta} \|x - x_t\|_2^2.$$

Proof: We can simply verify the optimality conditions from theorem 14:

$$\tilde{x} = \arg\min_{x\in K} q(x) \iff \forall x \in K : \langle \nabla q(\tilde{x}), x - \tilde{x} \rangle \geq 0;$$

$$x_{t+1} = P_K(\bar{x}_{t+1}) = \arg\min_{x\in K} \frac{\|x - \bar{x}_{t+1}\|_2^2}{2} \iff \forall x \in K : \langle x_{t+1} - \bar{x}_{t+1}, x - \bar{x}_{t+1} \rangle \geq 0.$$

Now we compute the gradients

$$\nabla q(\tilde{x}) = \nabla f(x_t) + \frac{1}{\eta}(\tilde{x} - x_t) = \frac{1}{\eta}(\tilde{x} - x_t - \eta \nabla f(x_t)) = \frac{\tilde{x} - \bar{x}_{t+1}}{\eta},$$

where in the last step we used $\bar{x}_{t+1} := x_t - \eta \nabla f(x_t)$. But this is the same as the gradient (up to scalar $\eta > 0$) of the objective function for the projection $\|x - \bar{x}_{t+1}\|_2^2$, so the optimality conditions are equivalent.

Now we can perform essentially the same analysis as gradient descent. We emphasize here that the algorithm requires computing P_K, which in itself may be computationally difficult, depending on the constraint set K. But if this can be accomplished, then the convergence guarantee is the same as the unconstrained setting.

We show just the analysis for the smooth case as an illustration, and leave the remaining cases for the reader.

Theorem 16. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and β-smooth, let $K \subseteq \mathbb{R}^n$ be a convex compact constraint set, and let*

$$x^* := \arg\min_{x \in K} f(x).$$

Then for projected gradient update as in theorem 13 with $\eta := 1/\beta$,

$$f(x_T) - f(x^*) \lesssim \frac{\beta \|x_0 - x^*\|_2^2}{T}.$$

Proof:

$$\begin{aligned} f(x) + \frac{\beta}{2}\|x - x_t\|_2^2 &\geq f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{\beta}{2}\|x - x_t\|_2^2 =: q(x) \\ &\geq q(x_{t+1}) + \langle \nabla q(x_{t+1}), x - x_{t+1} \rangle + \frac{\beta}{2}\|x - x_{t+1}\|_2^2 \\ &\geq f(x_{t+1}) + 0 + \frac{\beta}{2}\|x - x_{t+1}\|_2^2, \end{aligned}$$

where the first step was by convexity, in the third step we used β-strong convexity of q, and in the final step we used β-smoothness of f for the first term, and optimaility conditions for $x_{t+1} := \arg\min_{x \in K} q(x)$ for the second. Rearranging gives

$$f(x_{t+1}) - f(x) \leq \frac{\beta}{2}(\|x - x_t\|_2^2 - \|x - x_{t+1}\|_2^2),$$

and from here we can continue the same as in the unconstrained case to show the same $1/T$ convergence rate.