IFT 6085 - Lecture 2
Basics of convex analysis and gradient descent

Scribes
Winter 2019: Andrew Williams, Ankit Vani, Maximilien Le Clei
Winter 2018: Assya Trofimov, Mohammad Pezeshki, Reyhane Askari

Instructor: Ioannis Mitliagkas

1 Introduction

Many machine learning problems involve learning parameters $\theta \in \Theta$ of a function f towards achieving an objective better. Typically, such objectives are characterized by a loss function $L : \Theta \rightarrow \mathbb{R}$, and training the model corresponds to searching the optimal parameters θ^* that minimize this loss.

For example, in supervised learning, θ parameterizes a function $f : X \rightarrow Y$, where any $x \in X$ is an input and any $y \in Y$ is a target label. Then,

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i;\theta), y_i)$$

represents the loss function for a dataset containing n training examples $(x_1, y_1), \ldots, (x_n, y_n)$. Here, ℓ is a deterministic function that determines the distance between a target label y_i and the predicted label $y_i' = f(x_i;\theta)$. In this setting, learning is carried out by performing *empirical risk minimization*, which involves optimizing to find parameters $\theta^* \in \arg\min_{\theta\in\Theta} L(\theta)$.

In the first few lectures, we will dive deeper into the basics and theory of optimization that lie at the heart of machine learning. We will step back from the notation we see in machine learning, and start by considering the most general unconstrained optimization problem[1] for a function $f : \mathcal{X} \rightarrow \mathbb{R}$:

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

In most of our discussions, we will consider $\mathcal{X}$ or $\mathrm{dom}(f)$ to be the d-dimensional Euclidean space $\mathbb{R}^d$.

The optimization problem formulated above is NP-hard in general (see [2]. Section 6.6). However, for certain classes of functions f, strong theoretical guarantees and efficient optimization algorithms exist. In this lecture, we consider such a class of functions, called *convex functions* and prove convergence guarantees for an algorithm for convex optimization called *gradient descent*.

2 Convex optimization

This section introduces some concepts in convexity, and then uses them to prove convergence of gradient descent for convex functions.

Although in practice, people commonly use the same algorithms for non-convex optimization as they do for convex optimization (such as gradient descent), it is important to note that the strong theory for convex optimization algorithms

[1] More generally, this would involve an $\inf$ instead of $\min$, but in this lecture we keep the notation simple and stick with $\min$.

often breaks down without the convexity assumption. However, ideas from convex analysis and the weakening of certain results can give partial guarantees and offer generalizations for non-convex analysis.

2.1 Background

2.1.1 Lipschitz continuity

Definition 1 (Lipschitz continuity)**.** *A function* $f(x)$ *is* L-Lipschitz *iff* $\forall x, y \in \mathbb{R}, \forall L \in \mathbb{R}_{\geq 0}$,

$$|f(x) - f(y)| \leq L\|x - y\|$$

Intuitively, a Lipschitz continuous function is bounded in how fast it can change. Figure 1 illustrates two Lipschitz continuous functions with different Lipschitz constants.

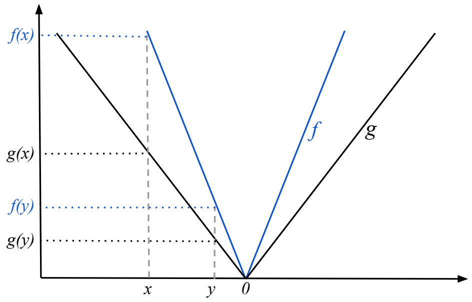

Figure 1: Consider a L_f-Lipschitz continuous function f and a L_g-Lipschitz continuous function g. If f and g are changing as fast as they can, then $L_f > L_g$.

As another example, consider the following function:

$$f(x) = \begin{cases} \exp(-\lambda x), & \text{if } x > 0 \\ 1, & \text{otherwise} \end{cases}$$

$f(x)$ here is L-lipschitz, and the value of L increases with λ. As the value of λ increases, the function gets closer to discontinuity. As λ becomes ∞, we recover a step function, which is not Lipschitz continuous for any L. This function is illustrated in Figure 2.1.1.

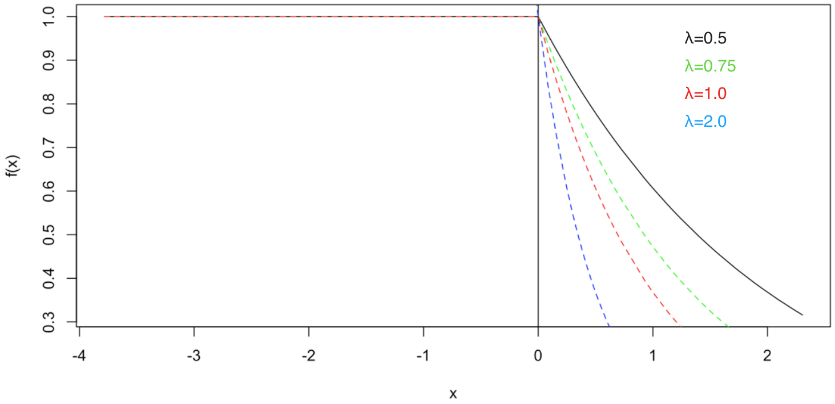

Figure 2: As λ increases in $f(x)$, the Lipschitz constant L increases, and the function gets closer to being discontinuous.

A Lipschitz continuous function does not need to be differentiable[2]. However, a corollary of f being L-Lipschitz continuous is that if it is differentiable, the norm of its gradient is bounded by L. For $\mathrm{dom}(f) \subseteq \mathbb{R}$,

$$f'(x) = \lim_{y \to x} \frac{f(x) - f(y)}{x - y} \qquad \text{(definition of derivative)}$$

$$\implies |f'(x)| = \lim_{y \to x} \frac{|f(x) - f(y)|}{|x - y|} \leq L \qquad \text{(definition of Lipschitz continuity)}$$

In general, if $\mathrm{dom}(f) \subseteq \mathbb{R}^d$ for a differentiable L-Lipschitz function f, then

$$\|\nabla f(x)\| \leq L$$

The inverse of the above statement also holds. If the norm of the gradient of a function is bounded, then the function is Lipschitz continuous.

Note that Lipschitz continuity is a special case of continuity: all Lipschitz continuous functions are continuous, but not all continuous functions are Lipschitz continuous (for more information, see [1]).

2.1.2 Convex sets

Before we define convex sets, let us first define a convex combination, which is a constrained version of a linear combination, illustrated in Figure 3 for two points.

Definition 2 (Convex combination)**.** *If $z \in \mathbb{R}^d$ is a linear combination of $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ and the coefficients are non-negative and sum to 1, then z is a convex combination of $x_1, x_2, \ldots, x_n$:*

$$z = \sum_{i=1}^{n} \theta_i x_i, \quad \text{where } \forall i \in (1, \ldots, n), \theta_i \geq 0 \text{ and } \sum_{i=1}^{n} \theta_i = 1$$

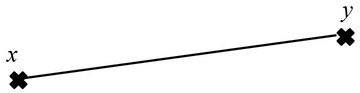

Figure 3: All convex combinations $z = \theta x + (1-\theta)y$ of two points x and y lie on the line segment from x to y. When $\theta = 0$, we get x and when $\theta = 1$, we get y.

Definition 3 (Convex set)**.** *$\mathcal{X}$ is a convex set if the convex combination any two points in $\mathcal{X}$ is also in $\mathcal{X}$. That is, for a convex set $\mathcal{X}$:*

$$\forall x, y \in \mathcal{X}, \forall \theta \in [0, 1], \quad z = \theta x + (1-\theta)y \in \mathcal{X}$$

Figure 4 gives examples of a convex set and a non-convex set.

2.1.3 Convex functions

Definition 4 (Convex function)**.** *A function $f(x)$ is* convex *iff the domain $\mathrm{dom}(f)$ is a convex set and $\forall x, y \in \mathrm{dom}(f), \forall \theta \in [0, 1]$,*

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

[2] For example, we can have integer-valued and non-smooth Lipschitz continuous functions.

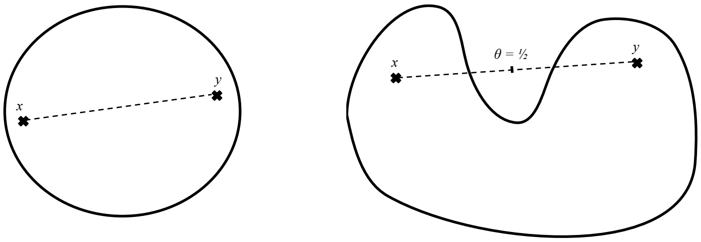

Figure 4: Examples of a convex and a non-convex set. **Left:** Convex set, **Right:** Non-convex set.

The condition above says that for any two members in the domain of f, the function's value on a convex combination does not exceed the convex combination of those values. Graphically, when f is convex, for any points x and y in f's domain, the chord connecting $f(x)$ and $f(y)$ lies above the function between those points. This is illustrated in Figure 5.

For a convex function f, the function $-f$ is defined as a *concave* function. In other words, for a concave function, the inequality condition in Definition 4 is reversed.

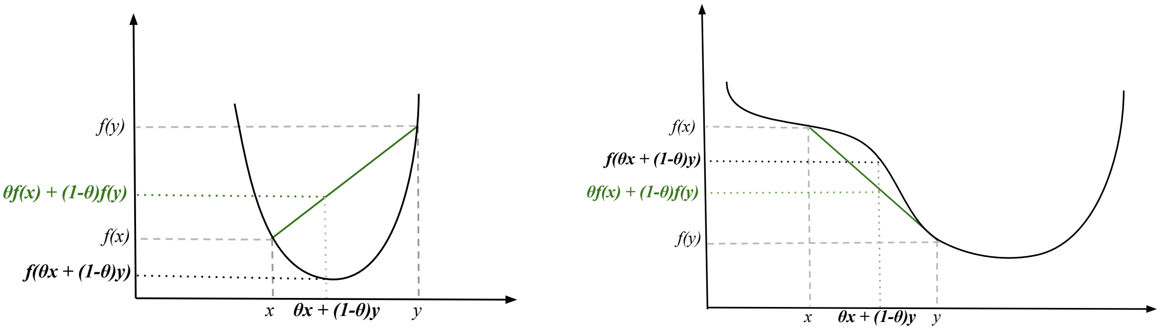

Figure 5: **Left:** Example of a convex function. For any two points $x, y \in \mathrm{dom}(f)$, the chord $\theta f(x) + (1-\theta)f(y), \theta \in [0,1]$ lies above the function value $f(\theta x + (1-\theta)y), \theta \in [0,1]$. **Right:** Example of a non-convex function[3]. We see here that there exist points x and y for which the chord lies below the curve between $f(x)$ and $f(y)$.

Moreover, for differentiable and twice differentiable functions, it is possible to define convexity in terms of first and second order conditions for convexity. Note that all the definitions of convexity are equivalent when the appropriate level of differentiability holds (for more information, see [2]).

Lemma 1 (First order condition for convexity)**.** *A differentiable function $f(x)$ is convex iff $\mathrm{dom}(f)$ is convex and $\forall x, y \in \mathrm{dom}(f)$,*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x)$$

Intuitively, this says that for a convex function, a tangent of its graph at any point must lie below the graph. This is illustrated in Figure 6.

Before discussing the second-order condition for convexity, let us review the multivariate generalization of a second derivative, namely a *Hessian*:

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d x_1} & \frac{\partial^2 f}{\partial x_d x_2} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix}$$

[3] In fact, this is an example of a quasiconvex function f, in which all sublevel sets $S_\alpha(f) = \{x \mid f(x) \leq \alpha\}$ are convex sets.

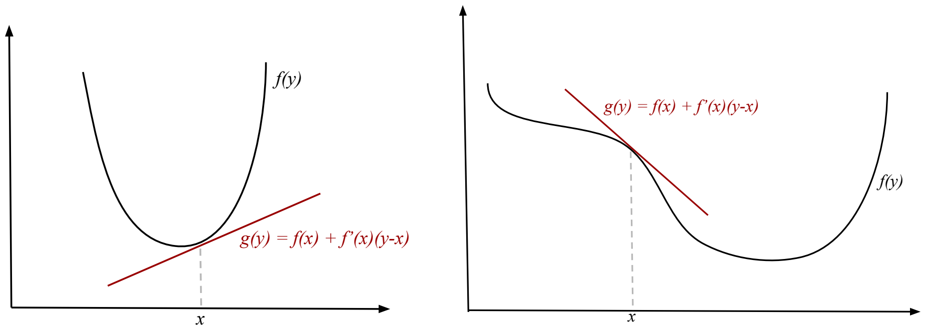

Figure 6: Example of a convex and a non-convex function illustrating the first-order condition for convexity. For the convex function on the left, all possible tangents will lie below the graph. However, for the non-convex function on the right, there exists a tangent such that it lies above the graph for some points in the function's domain.

For a function $f : \mathbb{R} \rightarrow \mathbb{R}, f(x) = \frac{h}{2}x^2$, the second derivative $f''(x) = h$ corresponds to a measure of how quickly the shape changes in the function. Similarly, the Hessian represents the curvature of a function f with $\mathrm{dom}(f) \subseteq \mathbb{R}^d$. A multivariate quadratic function f can be written as $f(x) = \frac{1}{2}x^\top Hx$, where H is the Hessian. The eigenvalues of the Hessian determines the curvature of the function along its eigenvectors. Consider the eigendecomposition

$$H = Q\Lambda Q^\top$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_4 \end{bmatrix}$$

Changing the basis to Q, we can focus on the directions described by $Q = [q_1, q_2, \ldots, q_d]$. Then, along the direction q_i, we get the curvature λ_i. Figure 7 illustrates the curvature of a quadratic function.

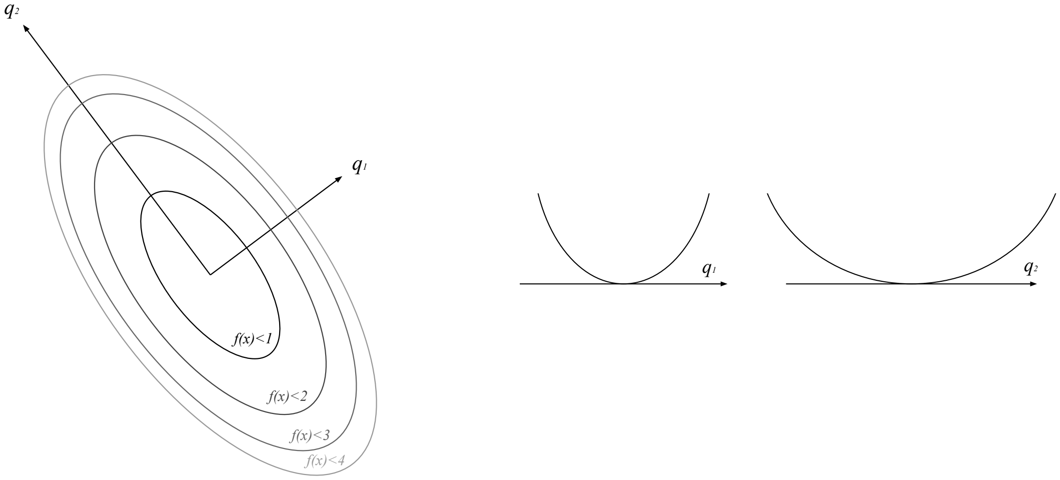

Figure 7: **Left:** Looking along the principle directions of the quadratic function $f(x) = \frac{1}{2}x^\top Hx$, we see that the curve changes faster along q_1 than q_2. Here[4], $\lambda_1 > \lambda_2$. **Right:** Cross-sections along q_1 and q_2, showing that f has higher curvature along q_1 than q_2.

[4] Most eigendecomposition algorithms return eigenvalues in non-decreasing order.

Lemma 2 (Second order condition for convexity)**.** *A twice differentiable function f is convex iff* $\mathrm{dom}(f)$ *is convex and* $\forall x \in \mathrm{dom}(f)$*,*

$$\nabla^2 f(x) \succeq 0$$

The above definition states that for a twice differentiable function to be convex, its Hessian must be positive-semidefinite. A matrix being positive-semidefinite implies that all of its eigenvalues are non-negative ($\lambda_i \geq 0$).

For any non-negative eigenvalue of the Hessian, the curvature of the function is non-negative along the corresponding eigenvector, and thus the function is convex in that direction. On the other hand, a non-positive eigenvalue represents non-positive curvature along the eigenvector, and the function is concave in that direction. Then, we can see that for a function to be convex, it has to have non-negative curvature, and thus non-negative eigenvalues, in all directions.

2.1.4 Smoothness and strong convexity

Although the convergence proof we describe later in this lecture does not require additional assumptions of *smoothness* and *strong convexity*, we introduce these concepts here. These conditions provide even stronger theoretical guarantees for gradient descent, which we look at in the following lectures.

Definition 5 (Smoothness)**.** *A continuously differentiable function $f(x)$ is β-smooth if its gradient is β-Lipschitz. That is, for some $\beta > 0$, $\forall x, y \in \mathrm{dom}(f)$,*

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$$

Essentially, smoothness constrains the gradient of $f(x)$ from changing too rapidly.

Definition 6 (Strong convexity)**.** *A function $f(x)$ is α-strongly convex if for $\alpha > 0$, $\forall x \in \mathrm{dom}(f)$, $f(x) - \frac{\alpha}{2}\|x\|^2$ is convex.*

Strong convexity provides a lower bound for the function's curvature. In other words, all eigenvalues of the Hessian of a α-strongly convex function are lower bounded by α. We can write this in terms of positive-semidefiniteness as

$$\nabla^2 f(x) \succeq \alpha I \iff \nabla^2 f(x) - \alpha I \succeq 0$$

For example, $f : \mathbb{R} \to \mathbb{R}, f(x) = \frac{h}{2}x^2$ is h-strongly convex, but not $(h + \epsilon)$-strongly convex for $\epsilon > 0$. Figure 8 illustrates examples of two convex functions, of which only one is strongly convex.

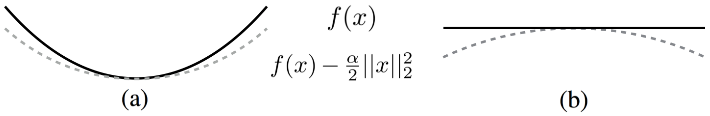

Figure 8: (a) A convex function which is also strongly convex. (b) A convex function which is not strongly convex.

2.2 Gradient descent

Gradient descent is an optimization algorithm that starts from an initial point, and moves in the direction of steepest descent. This makes the algorithm especially useful for convex optimization, since for a convex function, any local minimum is also a global minimum. Specifically, starting from an initial guess x_1, the algorithm generates the sequence $x_1, x_2, ..., x_T \in \mathbb{R}^d$ to approach the minimum of a function $f : \mathbb{R}^d \to \mathbb{R}$ using the following update rule:

$$x_{k+1} = x_k - \gamma \nabla f(x_k),$$

Here, γ is called the *step size*, also known as the *learning rate* in machine learning literature. If f is convex and γ decays at an appropriate rate, then it is guaranteed that as $T \to \infty$, $x_T \to x^*$, where $x^* \in \arg\min_{x \in \mathrm{dom}(f)} f(x)$ is an optimal value.

Lemma 3. *If f is a L-Lipschitz continuous function, then*

$$\|\nabla f(x_k)\|_2^2 \leq L^2.$$

Theorem 1. *Let $f(x)$ be convex and L-Lipschitz continuous*[5]*. If we take T steps of gradient descent with the step size*

$$\gamma = \frac{\|x_1 - x^*\|_2}{L\sqrt{T}}$$

Then the following holds:

$$f\left(\frac{1}{T}\sum_{k=1}^{T} x_k\right) - f(x^*) \leq \frac{\|x_1 - x^*\|L}{\sqrt{T}}$$

(see [2]. Theorem 3.2)

Proof. Using the first order condition of convexity and rearranging terms, we can write:

$$\begin{aligned}
f(x_k) - f(x^*) &\leq \langle \nabla f(x_k), x_k - x^* \rangle \\
&= \left\langle \frac{1}{\gamma}(x_k - x_{k+1}), x_k - x^* \right\rangle && \text{(using gradient descent update rule)} \\
&= \frac{1}{2\gamma}\left(-\|x_k - x_{k+1} - (x_k - x^*)\|_2^2 + \|x_k - x_{k+1}\|_2^2 + \|x_k - x^*\|_2^2\right) \\
& && \text{(using } \|a-b\|^2 = \|a\|^a + \|b\|^2 - 2\langle a, b\rangle \text{ and rearranging terms)} \\
&= \frac{1}{2\gamma}\left(-\|x_{k+1} - x^*\|_2^2 + \|\gamma \nabla f(x_k)\|_2^2 + \|x_k - x^*\|_2^2\right) && \text{(using gradient descent update rule)} \\
&= \frac{1}{2\gamma}\left(\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2\right) + \frac{\gamma}{2}\|\nabla f(x_k)\|_2^2
\end{aligned}$$

Using Lemma 3, we can thus write:

$$f(x_k) - f(x^*) \leq \frac{1}{2\gamma}\left(\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2\right) + \frac{\gamma}{2}L^2 \tag{1}$$

Let us now perform the change of variables by defining $D_k = \|x_k - x^*\|$. Then, from Equation 1, we have:

$$\begin{aligned}
f(x_1) - f(x^*) &\leq \frac{1}{2\gamma}\left(D_1^2 - D_2^2\right) + \frac{\gamma}{2}L^2 \\
f(x_2) - f(x^*) &\leq \frac{1}{2\gamma}\left(D_2^2 - D_3^2\right) + \frac{\gamma}{2}L^2 \\
&\vdots \\
f(x_{T-1}) - f(x^*) &\leq \frac{1}{2\gamma}\left(D_{T-1}^2 - D_T^2\right) + \frac{\gamma}{2}L^2 \\
f(x_T) - f(x^*) &\leq \frac{1}{2\gamma}\left(D_T^2 - D_{T+1}^2\right) + \frac{\gamma}{2}L^2 \leq \frac{1}{2\gamma}D_T^2 + \frac{\gamma}{2}L^2
\end{aligned}$$

Adding all the terms above, we get a telescopic sum where most of the D_k terms cancel. We get:

$$\begin{aligned}
\sum_{k=1}^{T}\left(f(x_k) - f(x^*)\right) &\leq \frac{1}{2\gamma}D_1^2 + \frac{T\gamma L^2}{2} \\
\implies \left(\frac{1}{T}\sum_{k=1}^{T} f(x_k)\right) - f(x^*) &\leq \frac{1}{2\gamma T}D_1^2 + \frac{\gamma L^2}{2}
\end{aligned} \tag{2}$$

[5] Although gradient descent and the theorem rely on the notion of using gradients, the function does not need to be differentiable. The algorithm, theorem and proof still hold when any of the subgradients is used in place of a gradient at points where the function is not differentiable.

Since f is convex, Jensen's inequality tells us that $f\left(\frac{1}{T}\sum_{k=1}^{T} x_k\right) \leq \frac{1}{T}\sum_{k=1}^{T} f(x_k)$. Thus, we can rewrite Equation 2 as:

$$f\left(\frac{1}{T}\sum_{k=1}^{T} x_k\right) - f(x^*) \leq \frac{1}{2\gamma T} D_1^2 + \frac{\gamma L^2}{2} \tag{3}$$

Plugging in $\gamma = \frac{\|x_1 - x^*\|_2}{L\sqrt{T}}$ gives us the result

$$f\left(\frac{1}{T}\sum_{k=1}^{T} x_k\right) - f(x^*) \leq \frac{\|x_1 - x^*\| L}{\sqrt{T}}$$

□

To understand how we derived the optimal γ in the above theorem, notice that the RHS of Equation 3 is a convex function of γ in the domain $\mathbb{R}_{\geq 0}$. The minimizing γ would give us the tightest bound, which can be found analytically using convexity by setting the gradient to zero and solving for γ.

The convergence rate we derived here for gradient descent is $O(1/\sqrt{T})$, which is quite slow. We will see in the following lectures how stronger assumptions on the function f can guarantee significantly faster convergence rates for gradient descent.

References

[1] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[2] S. Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.