Machine Learning and Optimization

Lecture 6: Gradient Descent Convergence Theory

CSCI 4140/6960 — Spring 2026

Course website: `cs.rpi.edu/~gittea/.../mlandopt.html`

Learning Objectives

By the end of this lecture, you should be able to:

1. **Distinguish** three types of convergence and when each applies
2. **Define** strong convexity and prove its key properties
3. **Explain** why curvature enables parameter convergence
4. **Prove** convergence of GD for smooth convex functions at rate $O(1/T)$
5. **Prove** convergence of GD for strongly convex functions at rate $O((1 - 1/\kappa)^T)$
6. **Identify** when GD works well and when it doesn't (through examples)

Lecture Roadmap

Types of Convergence

Smoothness and Convergence

Strong Convexity

Convergence Results

Numerical Examples

Three Notions of Convergence

When we say an optimization algorithm "converges," what do we mean?

Three Different Notions

1. **Parameter convergence**: $\|\mathbf{w}_T - \mathbf{w}^*\|_2 \to 0$ as $T \to \infty$
2. **Objective convergence**: $f(\mathbf{w}_T) - f(\mathbf{w}^*) \to 0$ as $T \to \infty$
3. **Gradient convergence**: $\|\nabla f(\mathbf{w}_T)\|_2 \to 0$ as $T \to \infty$

Warning

These are **NOT** equivalent in general!

What we'll prove today:

- Convex + smooth $\Rightarrow$ objective convergence $O(1/T)$
- **Strongly** convex + smooth $\Rightarrow$ parameter convergence $O((1 - 1/\kappa)^T)$

β-Smoothness

To prove GD converges, we need assumptions beyond convexity

Definition: f is **β-smooth** (or **β-Lipschitz continuous gradient**) if:

$$\boxed{\|\nabla f(\mathbf{w}) - \nabla f(\mathbf{v})\|_2 \leq \beta \|\mathbf{w} - \mathbf{v}\|_2 \quad \forall \mathbf{w}, \mathbf{v}}$$

Gradient is Lipschitz continuous with constant β

Equivalent condition (for twice differentiable f):

$$\nabla^2 f(\mathbf{w}) \preceq \beta I \quad \forall \mathbf{w}$$

All eigenvalues of Hessian $\leq \beta$

Intuition:

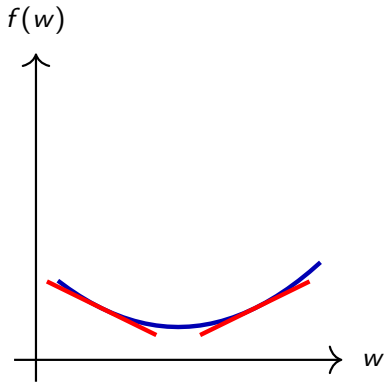
- Gradient doesn't change too rapidly
- Function curvature bounded
- No arbitrarily steep regions

Example: $f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T Q \mathbf{w}$ is β-smooth with $\beta = \lambda_{\max}(Q)$

Geometric Intuition: β-Smooth vs Non-Smooth

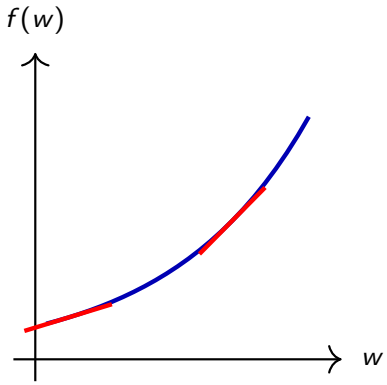
What does β-smoothness look like geometrically?

The gradient $\nabla f(\mathbf{w})$ can't change too rapidly as $\mathbf{w}$ moves



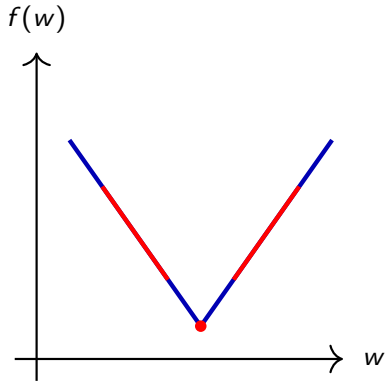
$f(w) = w^2$
β-smooth

Gradients change gradually



$f(w) = e^w$
NOT β-smooth

Curvature unbounded



$f(w) = |w|$
Not differentiable

Gradient undefined

Key Insight

β-smoothness requires: (1) **Differentiable** everywhere AND (2) **Bounded curvature** ($\nabla^2 f \preceq \beta I$). Exponential e^w fails (2), absolute value $|w|$ fails (1).

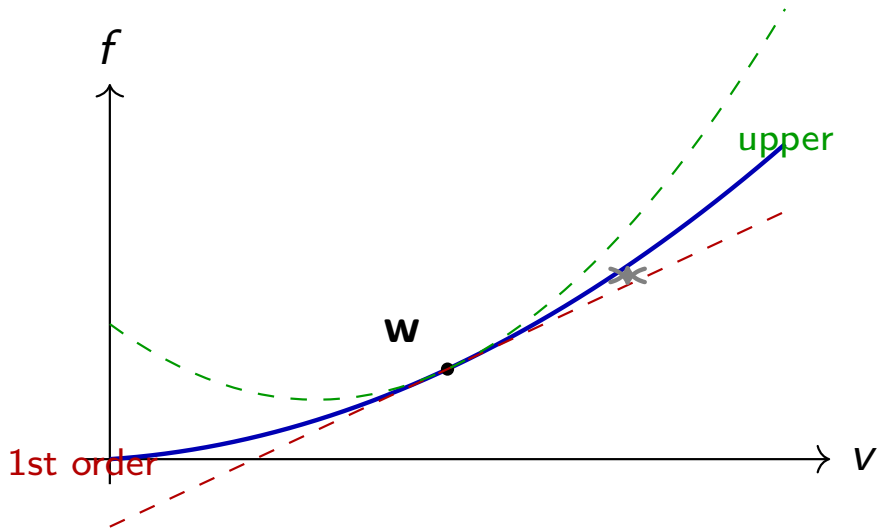
Why Smoothness Matters: Quadratic Upper Bound

Lemma: If f is β-smooth, then:

$$\boxed{f(\mathbf{v}) \leq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle + \frac{\beta}{2} \|\mathbf{v} - \mathbf{w}\|_2^2}$$

First-order Taylor plus quadratic correction!

Geometric picture:



Key insight:

- First-order Taylor (linear) **underestimates** convex f
- True function $f(v)$ sandwiched between
- Quadratic upper bound controls approximation error
- Error grows at most $O(\beta \|v - w\|^2)$

This is why we can trust gradient descent steps!

Quadratic Upper Bound Proof

Starting from:

$$f(\mathbf{v}) - f(\mathbf{w}) = \int_0^1 \langle \nabla f(\mathbf{w} + t(\mathbf{v} - \mathbf{w})), \mathbf{v} - \mathbf{w} \rangle \, dt$$

Add and subtract $\nabla f(\mathbf{w})$:

$$f(\mathbf{v}) - f(\mathbf{w}) = \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle + \int_0^1 \langle \nabla f(\mathbf{w} + t(\mathbf{v} - \mathbf{w})) - \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle \, dt$$

By Cauchy-Schwarz and β-smoothness:

$$|\langle \nabla f(\mathbf{w} + t(\mathbf{v} - \mathbf{w})) - \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle| \leq \beta t \|\mathbf{v} - \mathbf{w}\|_2^2$$

Therefore:

$$f(\mathbf{v}) - f(\mathbf{w}) \leq \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle + \int_0^1 \beta t \|\mathbf{v} - \mathbf{w}\|_2^2 \, dt = \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle + \frac{\beta}{2} \|\mathbf{v} - \mathbf{w}\|_2^2$$

Descent Lemma

Motivation: Once we know a function is β-smooth, we can guarantee how to choose a step size so that gradient descent decreases the function value. This is a really important result!

Lemma (Descent): If f is β-smooth, then GD with step size $\alpha = 1/\beta$ satisfies:

$$\boxed{f(\mathbf{w}_t) - f(\mathbf{w}_{t-1}) \leq -\frac{1}{2\beta}\|\nabla f(\mathbf{w}_{t-1})\|_2^2}$$

Each step decreases objective by amount proportional to gradient squared!

Proof: Apply the quadratic upper bound lemma with:

$$\mathbf{v} = \mathbf{w}_t = \mathbf{w}_{t-1} - \frac{1}{\beta}\nabla f(\mathbf{w}_{t-1})$$

Descent Lemma Proof

Starting from quadratic upper bound with $\mathbf{v} = \mathbf{w}_t = \mathbf{w}_{t-1} - \frac{1}{\beta}\nabla f(\mathbf{w}_{t-1})$:

$$f(\mathbf{w}_t) \leq f(\mathbf{w}_{t-1}) + \langle \nabla f(\mathbf{w}_{t-1}), \mathbf{w}_t - \mathbf{w}_{t-1} \rangle + \frac{\beta}{2}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2$$

Substitute $\mathbf{w}_t - \mathbf{w}_{t-1} = -\frac{1}{\beta}\nabla f(\mathbf{w}_{t-1})$:

$$f(\mathbf{w}_t) \leq f(\mathbf{w}_{t-1}) + \langle \nabla f(\mathbf{w}_{t-1}), -\frac{1}{\beta}\nabla f(\mathbf{w}_{t-1}) \rangle + \frac{\beta}{2}\|\frac{1}{\beta}\nabla f(\mathbf{w}_{t-1})\|_2^2$$

Simplify:

$$f(\mathbf{w}_t) \leq f(\mathbf{w}_{t-1}) - \frac{1}{\beta}\|\nabla f(\mathbf{w}_{t-1})\|_2^2 + \frac{1}{2\beta}\|\nabla f(\mathbf{w}_{t-1})\|_2^2$$

Combine terms:

$$f(\mathbf{w}_t) \leq f(\mathbf{w}_{t-1}) - \frac{1}{2\beta}\|\nabla f(\mathbf{w}_{t-1})\|_2^2$$

Strong Convexity - Definition

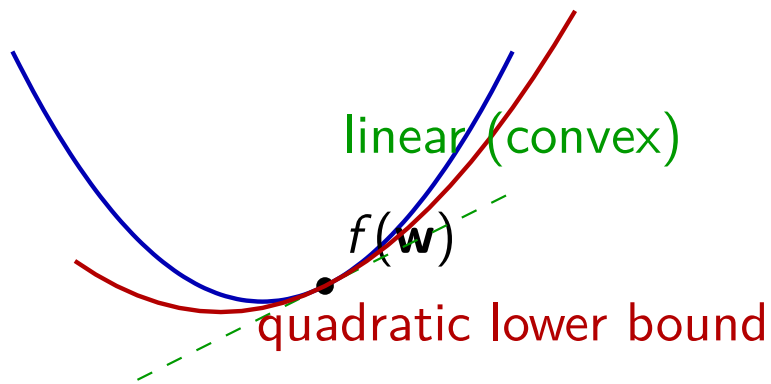
Convexity guarantees a linear lower bound. **Strong convexity** adds a quadratic term:

Definition: μ-Strong Convexity

f is **μ-strongly convex** if $\forall \mathbf{w}, \mathbf{v}$:

$$f(\mathbf{v}) \geq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle + \frac{\mu}{2} \|\mathbf{v} - \mathbf{w}\|_2^2$$

Equivalent formulation: $g(\mathbf{w}) = f(\mathbf{w}) - \frac{\mu}{2}\|\mathbf{w}\|_2^2$ is convex



Strong Convexity - Key Properties

Second-order characterization (twice differentiable):

$$\nabla^2 f(\mathbf{w}) \succeq \mu I \quad \forall \mathbf{w}$$

Minimum eigenvalue $\geq \mu$ everywhere

Unique global minimizer (if f has a minimizer)

Example

$f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T Q\mathbf{w}$ with $Q \succ 0$ is μ-strongly convex with $\mu = \lambda_{\min}(Q)$
Since $\nabla^2 f(\mathbf{w}) = Q$, and $Q \succeq \lambda_{\min}(Q)I$

Condition Number

When f is both μ-strongly convex AND β-smooth:

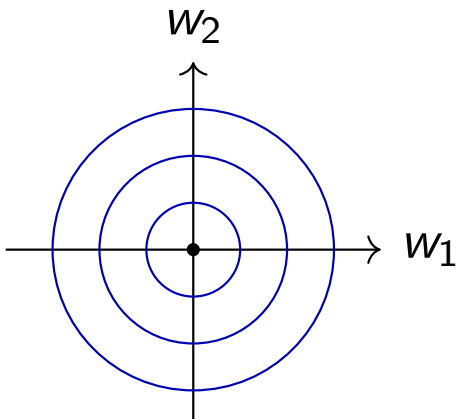
- Smoothness: $\nabla^2 f(\mathbf{w}) \preceq \beta I$ (curvature bounded above)
- Strong convexity: $\nabla^2 f(\mathbf{w}) \succeq \mu I$ (curvature bounded below)

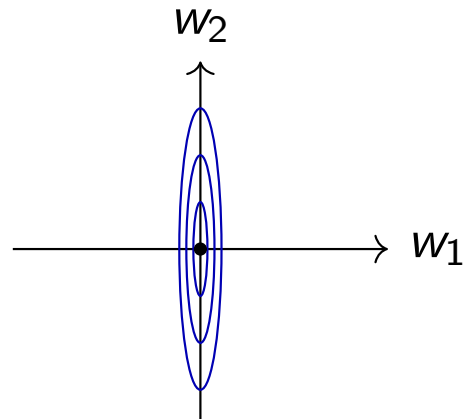
Definition: Condition Number

$$\boxed{\kappa = \frac{\beta}{\mu} \geq 1}$$

Measures problem difficulty:

- $\kappa = 1$: Perfect sphere (isotropic), easiest
- $\kappa \gg 1$: Elongated ellipsoid, harder





Convergence Theorem (Convex Case)

Theorem: Suppose f is convex and β-smooth. Run GD with constant step size $\alpha = 1/\beta$:

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \frac{1}{\beta}\nabla f(\mathbf{w}_{t-1})$$

Then after T iterations:

$$\boxed{f(\mathbf{w}_T) - f(\mathbf{w}^*) \leq \frac{\beta\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{2T}}$$

Interpretation:

- **Convergence type**: Objective gap only, rate $O(1/T)$ (sublinear)
- **Iteration complexity**: $T = O(\beta/\varepsilon)$ iterations for ε-suboptimality
- Faster for smaller β (less curved)
- Faster if $\mathbf{w}_0$ close to $\mathbf{w}^*$

Important

No parameter convergence guarantee! Without strong convexity, $\|\mathbf{w}_T - \mathbf{w}^*\|$ may not converge to zero.

Convergence for Strongly Convex Functions - Statement

Theorem: Suppose f is μ-strongly convex and β-smooth with $\kappa = \beta/\mu$. Run GD with $\alpha = 1/\beta$.

Then after T iterations:

$$\boxed{\|\mathbf{w}_T - \mathbf{w}^*\|_2^2 \leq \left(1 - \frac{1}{\kappa}\right)^T \|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}$$

Convergence type: Exponential (linear) decay in parameter distance!

Assumption	**Guarantee**	**Rate**
Convex + smooth	$f(\mathbf{w}_T) - f^* \leq O(1/T)$	Sublinear
Strongly convex + smooth[1]	$\|\mathbf{w}_T - \mathbf{w}^*\|^2 \leq O((1 - 1/\kappa)^T)$	**Exponential**

Iteration complexity: $T = O(\kappa \log(1/\varepsilon))$ for ε-accuracy in distance

[1] This rate is for using $\alpha = \frac{1}{\beta}$. You can do slightly better: see Appendix.

Understanding the $(1 - 1/\kappa)$ Rate

Rate factor: $\rho = 1 - \frac{1}{\kappa} = 1 - \frac{\mu}{\beta}$

κ	ρ	Error after 100 iter	Interpretation
1	0	0	1 step converges!
2	0.5	$\approx 10^{-30}$	Very fast
10	0.9	$\approx 10^{-5}$	Fast
100	0.99	≈ 0.37	Slow
1000	0.999	≈ 0.90	Very slow

Consequence: Poorly conditioned problems (κ large) converge slowly

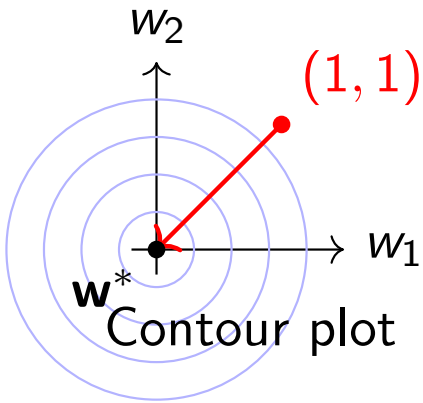
Each iteration reduces error by factor $\approx (1 - 1/\kappa)$

For $\kappa = 100$: need $T \approx 460$ iterations for 99% error reduction

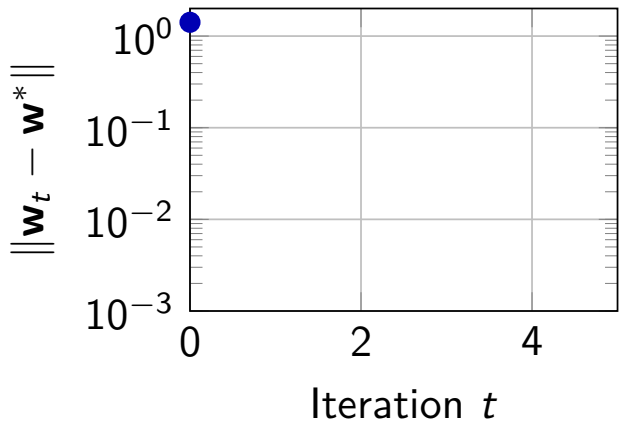
Example 1: Well-Conditioned Problem

Minimize $f(w_1, w_2) = w_1^2 + w_2^2$ starting from $(1, 1)$

Properties: $\mu = 2$, $\beta = 2$, $\kappa = 1$



Contour plot



Convergence plot

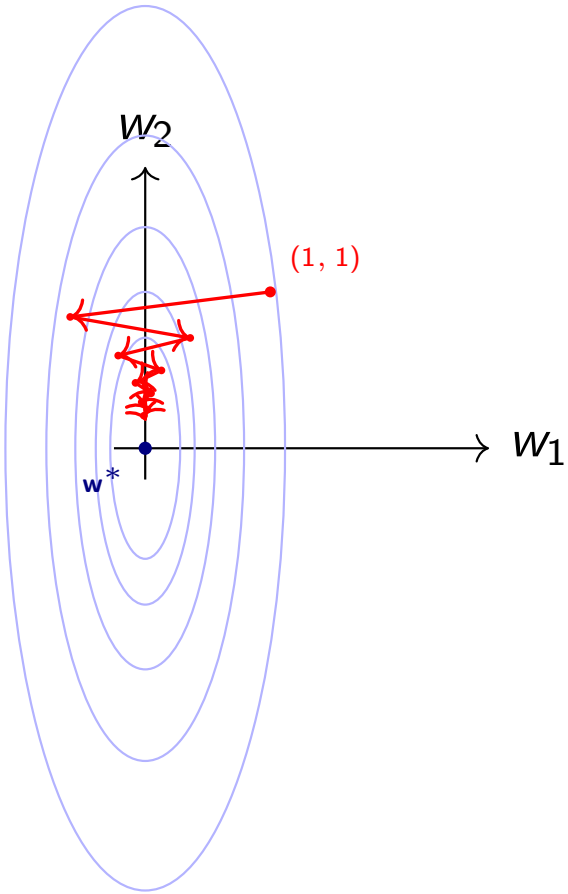
Theory predicts: $(1 - 1/\kappa)^T = 0^T$ — converges in 1 step with $\alpha = 1/\beta$!

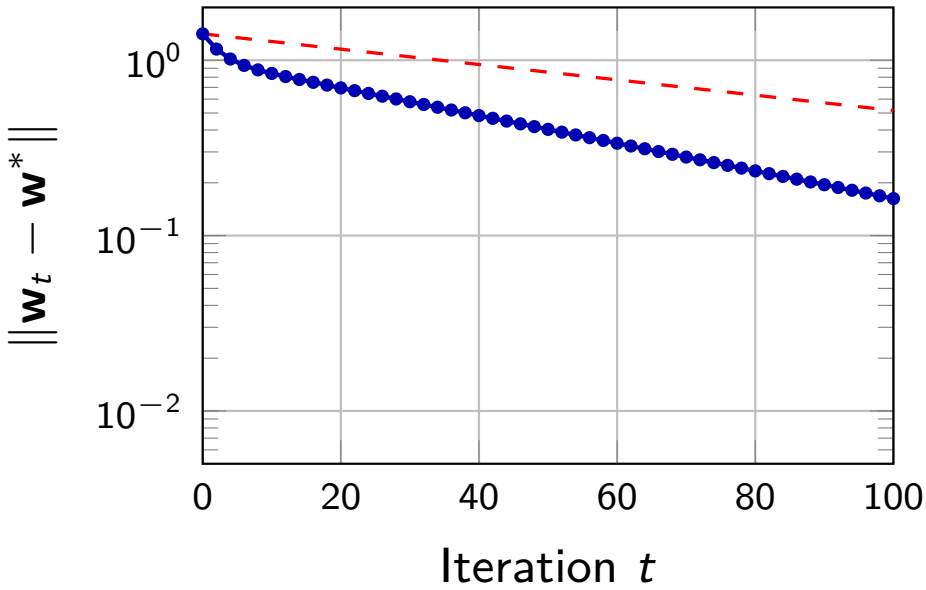
Observed: Exact convergence in 1 iteration ✓

Example 2: Poorly Conditioned Problem

Minimize $f(w_1, w_2) = 100w_1^2 + w_2^2$ starting from $(1, 1)$

Properties: $\mu = 2$, $\beta = 200$, $\kappa = 100$





Slow decay (blue) vs theory (red dashed)

Theory guarantees: $(1 - 1/\kappa)^T = 0.99^T$ — need $T \approx 460$ for 99% reduction

Observation: faster decay than theory, but still slow

Example 3: Large-Scale - Motivation for SGD

Least squares regression: $f(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)^2$

Gradient requires full dataset:

$$\nabla f(\mathbf{w}) = \frac{2}{n}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)\mathbf{x}_i$$

Cost per iteration: $O(nd)$ floating point operations

Timing estimates (assuming $d = 100$ features, modern CPU at $\sim$10 GFLOPS):

n	FLOPs/iter	Time/iter	Iters ($\kappa = 50$)	Total time
10^3	2×10^5	$20\mu s$	230	$4.6ms$ ✓
10^6	2×10^8	$20ms$	230	$4.6s$ ✓
10^9	2×10^{11}	$20s$	230	1.3 hours ×
10^{12}	2×10^{14}	5.5 hours	230	53 days ××

*Assuming dense matrix-vector products at 10 GFLOPS (10^{10} FLOPs/sec)

Problem

Full gradient computation prohibitive for big n!

Next lecture: Stochastic Gradient Descent uses random mini-batches!

When Does Gradient Descent Work?

Scenario	**Status**
Smooth + convex: Objective converges at $O(1/T)$ rate	✓
Smooth + strongly convex: Parameter converges at $O((1-1/\kappa)^T)$	✓
Poor conditioning ($\kappa \gg 1$): Slow exponential convergence	△
Large-scale ($n \gg 1$): Iteration cost too high	×

Coming Next

- **Lecture 7:** Stochastic Gradient Descent (random sampling for large n)
- **Lecture 8:** Subgradient methods (when f not differentiable)

Questions?

Proof of Convergence Theorem (Convex Case) - Proof Idea

We want to upper bound the final objective gap

$$f(\mathbf{w}_T) - f(\mathbf{w}^*)$$

after T gradient steps.

Strategy:

- Relate the objective gap at step t to how much closer $\mathbf{w}_t$ is to $\mathbf{w}^*$ than $\mathbf{w}_{t-1}$. Specifically, show each step gives

$$f(\mathbf{w}_t) - f(\mathbf{w}^*) \lesssim \|\mathbf{w}_{t-1} - \mathbf{w}^*\|^2 - \|\mathbf{w}_t - \mathbf{w}^*\|^2$$

- Add these inequalities for $t = 1, \ldots, T$.

When we add them, the drops in squared distance *telescope*— meaning they cancel out—, giving

$$\sum_{t=1}^{T} f(\mathbf{w}_t) - f(\mathbf{w}^*) \lesssim \|\mathbf{w}_0 - \mathbf{w}^*\|^2 - \|\mathbf{w}_T - \mathbf{w}^*\|^2$$

$$\leq \|\mathbf{w}_0 - \mathbf{w}^*\|^2$$

Finally, monotonic decrease of $f(\mathbf{w}_t)$ lets us turn the *sum* of gaps into a bound on the *last iterate*.

Proof of Convergence Theorem (Convex Case) - Part 1

By convexity at $\mathbf{w}_{t-1}$:

$$f(\mathbf{w}_{t-1}) - f(\mathbf{w}^*) \leq \langle \nabla f(\mathbf{w}_{t-1}), \mathbf{w}_{t-1} - \mathbf{w}^* \rangle$$

For convenience we let $\eta = \frac{1}{\beta}$ and expand the squared distance after one GD step $\mathbf{w}_t = \mathbf{w}_{t-1} - \eta \nabla f(\mathbf{w}_{t-1})$:

$$\begin{aligned} \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 &= \|\mathbf{w}_{t-1} - \eta \nabla f(\mathbf{w}_{t-1}) - \mathbf{w}^*\|_2^2 \\ &= \|\mathbf{w}_{t-1} - \mathbf{w}^*\|_2^2 \\ &\quad - 2\eta \langle \nabla f(\mathbf{w}_{t-1}), \mathbf{w}_{t-1} - \mathbf{w}^* \rangle \\ &\quad + \eta^2 \|\nabla f(\mathbf{w}_{t-1})\|_2^2 \end{aligned}$$

Rearranging the inner product:

$$\langle \nabla f(\mathbf{w}_{t-1}), \mathbf{w}_{t-1} - \mathbf{w}^* \rangle = \frac{1}{2\eta} \left(\|\mathbf{w}_{t-1} - \mathbf{w}^*\|_2^2 - \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 \right) + \frac{\eta}{2} \|\nabla f(\mathbf{w}_{t-1})\|_2^2$$

Proof of Convergence Theorem (Convex Case) - Part 2

Plug into the convexity bound:

$$f(\mathbf{w}_{t-1}) - f(\mathbf{w}^*) \leq \frac{1}{2\eta}\left(\|\mathbf{w}_{t-1} - \mathbf{w}^*\|_2^2 - \|\mathbf{w}_t - \mathbf{w}^*\|_2^2\right) + \frac{\eta}{2}\|\nabla f(\mathbf{w}_{t-1})\|_2^2$$

From the descent inequality,

$$f(\mathbf{w}_t) \leq f(\mathbf{w}_{t-1}) - \frac{\eta}{2}\|\nabla f(\mathbf{w}_{t-1})\|_2^2$$

Rearrange to upper bound the gradient term:

$$\frac{\eta}{2}\|\nabla f(\mathbf{w}_{t-1})\|_2^2 \leq f(\mathbf{w}_{t-1}) - f(\mathbf{w}_t)$$

Substitute this bound back in and cancel $f(\mathbf{w}_{t-1})$:

$$\boxed{f(\mathbf{w}_t) - f(\mathbf{w}^*) \leq \frac{1}{2\eta}\left(\|\mathbf{w}_{t-1} - \mathbf{w}^*\|_2^2 - \|\mathbf{w}_t - \mathbf{w}^*\|_2^2\right)}$$

This is the inequality we will sum over t.

Proof of Convergence Theorem (Convex Case) - Part 3

Sum the boxed inequality from $t = 1$ to T:

$$\sum_{t=1}^{T} \big(f(\mathbf{w}_t) - f(\mathbf{w}^*)\big) \leq \frac{1}{2\eta} \sum_{t=1}^{T} \Big(\|\mathbf{w}_{t-1} - \mathbf{w}^*\|_2^2 - \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 \Big)$$

The squared-distance terms telescope:

$$\boxed{\sum_{t=1}^{T} \big(f(\mathbf{w}_t) - f(\mathbf{w}^*)\big) \leq \frac{1}{2\eta} \|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}$$

The descent lemma tells us that the function values do not increase at any iteration, so

$$f(\mathbf{w}_T) \leq f(\mathbf{w}_t) \quad \text{for all } t = 1, \ldots, T.$$

Therefore each term in the sum is at least $f(\mathbf{w}_T) - f(\mathbf{w}^*)$, giving

$$T\big(f(\mathbf{w}_T) - f(\mathbf{w}^*)\big) \leq \sum_{t=1}^{T} \big(f(\mathbf{w}_t) - f(\mathbf{w}^*)\big).$$

Proof of Convergence Theorem (Convex Case) - Part 4

Combine the last observation with the last boxed inequality:

$$T\big(f(\mathbf{w}_T) - f(\mathbf{w}^*)\big) \leq \frac{1}{2\eta}\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2$$

Divide by T:

$$f(\mathbf{w}_T) - f(\mathbf{w}^*) \leq \frac{\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{2\eta T}$$

Recall that $\eta = 1/\beta$:

$$f(\mathbf{w}_T) - f(\mathbf{w}^*) \leq \frac{\beta\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{2T}$$

□

Rates for GD on strongly convex smooth functions

Assume f is μ-strongly convex and β-smooth (so $\kappa = \frac{\beta}{\mu}$) with minimizer $\mathbf{w}^*$ and optimum value f^*.

Theorem 1: Exponential objective value convergence.
GD with step size $\alpha = \frac{1}{\beta}$ satisfies:

$$\boxed{f(\mathbf{w}_T) - f^* \leq \left(1 - \frac{1}{\kappa}\right)^T \left(f(\mathbf{w}_0) - f^*\right)}$$

Theorem 2: Exponential parameter convergence.
GD with the *optimal* constant step size

$$\alpha^\star = \frac{2}{\mu + \beta}$$

satisfies:

$$\boxed{\|\mathbf{w}_T - \mathbf{w}^*\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^T \|\mathbf{w}_0 - \mathbf{w}^*\|_2}$$

Comparing the Results

1. Since $\mu < \beta$, we have $\mu + \beta < 2\beta$, hence:

$$\alpha^\star = \frac{2}{\mu + \beta} > \frac{1}{\beta} = \alpha$$

   Knowing both μ and β enables larger, more aggressive step sizes when trying to recover parameters.

2. *Parameters can be derived at a (slightly) faster rate.*

$$\frac{\kappa - 1}{\kappa + 1} \leq \frac{\kappa - 1}{\kappa} = 1 - \frac{1}{\kappa}.$$

Obvious question: Can the objective convergence rate be (slightly) improved to match the parameter convergence rate? No! Both convergence rates are optimal.

The Polyak-Łojasiewicz (PL) Inequality

Strong convexity implies the gradient is large when you're far away from the smallest function value.

PL Inequality

If f is μ-strongly convex with minimizer $\mathbf{w}^*$, then:

$$\boxed{\frac{1}{2}\|\nabla f(\mathbf{w})\|_2^2 \geq \mu(f(\mathbf{w}) - f(\mathbf{w}^*))}$$

Interpretation:

- Gradient norm measures distance to optimum (in function values)
- If gradient is small, we must be close to optimal function value
- This is the key to proving fast convergence!

Just as quadratic upper bound (smoothness) enables monotonic descent, PL inequality (strong convexity) enables **exponential/linear convergence**!

Proof of the PL inequality

Strong convexity says every tangent line sits below the function, with a quadratic gap. For all $\mathbf{v}$, we have

$$\begin{aligned} f(\mathbf{v}) &\geq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle + \frac{\mu}{2} \|\mathbf{v} - \mathbf{w}\|_2^2 \\ &\geq \min_{\mathbf{z}} \left(f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{z} - \mathbf{w} \rangle + \frac{\mu}{2} \|\mathbf{z} - \mathbf{w}\|_2^2 \right). \end{aligned}$$

Minimizing the right-hand side (check this!) gives

$$f(\mathbf{v}) \geq f(\mathbf{w}) - \frac{1}{2\mu} \|\nabla f(\mathbf{w})\|_2^2.$$

At $\mathbf{v} = \mathbf{v}^*$, we get that

$$f^* \geq f(\mathbf{w}) - \frac{1}{2\mu} \|\nabla f(\mathbf{w})\|_2^2.$$

Rearranging gives the PL inequality. □

Function Value Convergence via PL

Theorem: GD with step size $\alpha = 1/\beta$ on μ-strongly convex, β-smooth f satisfies:

$$\boxed{f(\mathbf{w}_T) - f^* \leq \left(1 - \frac{1}{\kappa}\right)^T (f(\mathbf{w}_0) - f^*)}$$

Proof: By descent lemma (from smoothness):

$$f(\mathbf{w}_t) \leq f(\mathbf{w}_{t-1}) - \frac{1}{2\beta}\|\nabla f(\mathbf{w}_{t-1})\|_2^2$$

By PL inequality: $\|\nabla f(\mathbf{w}_{t-1})\|_2^2 \geq 2\mu(f(\mathbf{w}_{t-1}) - f^*)$
Therefore:

$$f(\mathbf{w}_t) \leq f(\mathbf{w}_{t-1}) - \frac{1}{2\beta} \cdot 2\mu(f(\mathbf{w}_{t-1}) - f^*) = f(\mathbf{w}_{t-1}) - \frac{\mu}{\beta}(f(\mathbf{w}_{t-1}) - f^*)$$

Subtracting f^* from both sides:

$$f(\mathbf{w}_t) - f^* \leq \left(1 - \frac{1}{\kappa}\right)(f(\mathbf{w}_{t-1}) - f^*)$$

Iterating T times gives the result. $\square$

Parameter Convergence

Goal: Prove parameter convergence directly using linear algebra

Setup:

- $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and β-smooth
- Unique minimizer $\mathbf{w}^*$ with $\nabla f(\mathbf{w}^*) = 0$
- GD with constant step size α: $\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha \nabla f(\mathbf{w}_t)$
- Define error: $\mathbf{e}_t := \mathbf{w}_t - \mathbf{w}^*$

Proof sketch: Express the gradient in terms of an "averaged Hessian" along the line segment from $\mathbf{w}^*$ to $\mathbf{w}_t$. Show this averaged Hessian contracts the error at each iteration.

The Averaged Hessian

By fundamental theorem of calculus applied to ∇f:

$$\nabla f(\mathbf{w}) - \nabla f(\mathbf{w}^*) = \int_0^1 \nabla^2 f(\mathbf{w}^* + t(\mathbf{w} - \mathbf{w}^*))(\mathbf{w} - \mathbf{w}^*)\, dt$$

Define the **averaged Hessian**:

$$H(\mathbf{w}) := \int_0^1 \nabla^2 f(\mathbf{w}^* + t(\mathbf{w} - \mathbf{w}^*))\, dt$$

Then since $\nabla f(\mathbf{w}^*) = 0$:

$$\boxed{\nabla f(\mathbf{w}) = H(\mathbf{w})(\mathbf{w} - \mathbf{w}^*)}$$

Observe that since $\mu I \preceq \nabla^2 f(\mathbf{w}) \preceq \beta I$ for all $\mathbf{w}$,

$$\boxed{\mu I \preceq H(\mathbf{w}) \preceq \beta I}$$

I.e., the integral of PSD matrices with eigenvalues in $[\mu, \beta]$ has eigenvalues in $[\mu, \beta]$.

GD as Matrix Iteration

Let $H_t := H(\mathbf{w}_t)$ so that

$$\nabla f(\mathbf{w}_t) = H_t \mathbf{e}_t.$$

The GD update on the error $\mathbf{e}_t = \mathbf{w}_t - \mathbf{w}^*$ is

$$\begin{aligned}
\mathbf{e}_{t+1} &= \mathbf{w}_{t+1} - \mathbf{w}^* \\
&= \mathbf{w}_t - \alpha \nabla f(\mathbf{w}_t) - \mathbf{w}^* \\
&= \mathbf{e}_t - \alpha H_t \mathbf{e}_t \\
&= (I - \alpha H_t)\mathbf{e}_t
\end{aligned}$$

So each step multiplies the current error by a matrix:

$$\mathbf{e}_{t+1} = (I - \alpha H_t)\mathbf{e}_t.$$

Because H_t is symmetric with eigenvalues in $[\mu, \beta]$,

$$\|\mathbf{e}_{t+1}\|_2 \leq \|I - \alpha H_t\|_2 \, \|\mathbf{e}_t\|_2 = \left(\max_{\lambda \in [\mu, \beta]} |1 - \alpha \lambda| \right) \|\mathbf{e}_t\|_2.$$

One-Step Contraction

Define the contraction factor:

$$q(\alpha) := \max_{\lambda \in [\mu, \beta]} |1 - \alpha\lambda|$$

Then:

$$\|\mathbf{e}_{t+1}\|_2 \leq q(\alpha)\|\mathbf{e}_t\|_2$$

Iterating T times:

$$\boxed{\|\mathbf{w}_T - \mathbf{w}^*\|_2 \leq q(\alpha)^T \|\mathbf{w}_0 - \mathbf{w}^*\|_2}$$

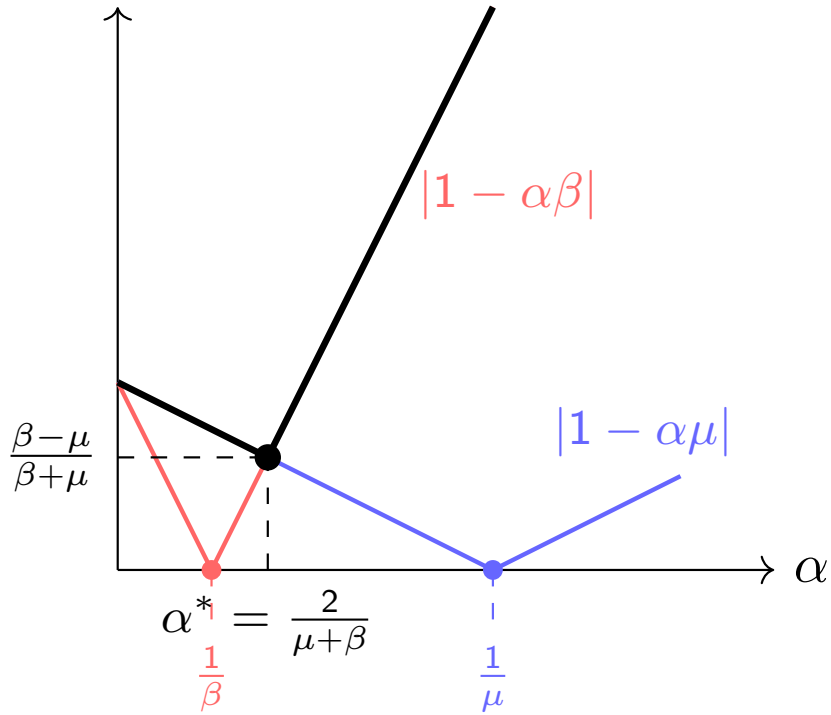
Question: What choice of α minimizes $q(\alpha)$?

Since $|1 - \alpha\lambda|$ is convex in λ, the maximum on $[\mu, \beta]$ occurs at an endpoint:

$$q(\alpha) = \max\{|1 - \alpha\mu|, |1 - \alpha\beta|\}$$

Optimal Step Size

We have: $q(\alpha) = \max\{|1 - \alpha\mu|, |1 - \alpha\beta|\}$



To minimize $q(\alpha)$, balance the two terms at their intersection:

$$1 - \alpha\mu = \alpha\beta - 1 \implies \alpha^* = \frac{2}{\mu + \beta}, \quad q(\alpha^*) = \frac{\beta - \mu}{\beta + \mu}$$

Optimal Contraction Factor

With $\alpha^* = \frac{2}{\mu+\beta}$:

$$q(\alpha^*) = |1 - \alpha^*\mu| = 1 - \frac{2\mu}{\mu+\beta} = \frac{(\mu+\beta) - 2\mu}{\mu+\beta} = \frac{\beta-\mu}{\beta+\mu}$$

Multiply numerator and denominator by $\frac{1}{\mu}$:

$$q(\alpha^*) = \frac{\beta/\mu - 1}{\beta/\mu + 1} = \frac{\kappa-1}{\kappa+1}$$

Observation: We could also take $\alpha = 1/\beta$. This gives

$$q(1/\beta) = \max\{|1 - \mu/\beta|, |1-1|\} = 1 - \mu/\beta = 1 - 1/\kappa$$

However, $\frac{\kappa-1}{\kappa+1} < 1 - \frac{1}{\kappa}$ for $\kappa > 1$
So $\alpha^* = \frac{2}{\mu+\beta}$ is actually better than $\alpha = 1/\beta$, as we saw before.

Parameter Convergence (Optimal Step Size)

Theorem: GD with optimal step size $\alpha^* = \frac{2}{\mu+\beta}$ satisfies:

$$\boxed{\|\mathbf{w}_T - \mathbf{w}^*\|_2 \leq \left(\frac{\kappa-1}{\kappa+1}\right)^T \|\mathbf{w}_0 - \mathbf{w}^*\|_2}$$

With the smoothness-derived step size $\alpha = 1/\beta$:

$$\boxed{\|\mathbf{w}_T - \mathbf{w}^*\|_2 \leq \left(1 - \frac{1}{\kappa}\right)^T \|\mathbf{w}_0 - \mathbf{w}^*\|_2}$$

Optimality of the Rate

Question: Is this the best possible for gradient descent?

Answer: Yes! Consider the quadratic $f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T Q\mathbf{w}$ where Q has eigenvalues $\{\mu, \beta\}$
For this function:

- $\nabla f(\mathbf{w}) = Q\mathbf{w}$
- GD update: $\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha Q\mathbf{w}_t = (I - \alpha Q)\mathbf{w}_t$
- Error: $\mathbf{e}_{t+1} = (I - \alpha Q)\mathbf{e}_t$

The worst-case contraction is exactly:

$$\max_{\lambda \in \{\mu, \beta\}} |1 - \alpha\lambda|$$

No constant step size can do better than $q(\alpha^*) = \frac{\kappa-1}{\kappa+1}$ on this example.

The rate $(\frac{\kappa-1}{\kappa+1})^T$ is optimal for first-order methods on this function class (convex, smooth, κ-conditioned)

Summary: Optimal Convergence Rates

For μ-strongly convex, β-smooth f with $\kappa = \beta/\mu$:

Function values (via PL):

$$\boxed{f(\mathbf{w}_T) - f^* \leq \left(1 - \frac{1}{\kappa}\right)^T (f(\mathbf{w}_0) - f^*)}$$

Parameters (via spectral analysis):

$$\boxed{\|\mathbf{w}_T - \mathbf{w}^*\|_2^2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^{2T} \|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}$$

Iteration complexity: $T = O(\kappa \log(1/\varepsilon))$ for ε-accuracy

Both rates are optimal for constant-step-size gradient descent!