Gradient Descent: Theory

Mikhailova Olena

27.03.2025

1 Gradient Descent

Definition 1 (Gradient Descent)**.** *An iterative optimization algorithm for minimizing a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. Starting from an initial point x_0, it updates:*

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

where $\eta_k > 0$ is the step size (learning rate).

Motivation: Follows the direction of steepest descent ($-\nabla f(x)$) to reduce $f(x)$. Can be derived from minimizing a first-order Taylor approximation with a quadratic regularization term:

$$x_{k+1} = \arg\min_y \left\{ f(x_k) + \nabla f(x_k)^T (y - x_k) + \frac{1}{2\eta} \|y - x_k\|^2 \right\}$$

2 Descent Lemma

2.1 Preliminary Concepts

Definition 2 (L-Smoothness)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is called **L-smooth** if its gradient is Lipschitz continuous with constant L:*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall x, y \in \mathbb{R}^n$$

2.2 Main Result

Lemma 3 (Descent Lemma)**.** *For any L-smooth function f and $\forall x, y \in \mathbb{R}^n$:*

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2} \|y - x\|^2$$

Proof. Proof via Taylor expansion and L-smoothness:

1. Start with Taylor's theorem with integral remainder:

$$f(y) = f(x) + \int_0^1 \nabla f(x + t(y - x))^\top (y - x)\, dt$$

2. Subtract linear approximation:

$$f(y) - f(x) - \nabla f(x)^\top (y - x) = \int_0^1 [\nabla f(x + t(y - x)) - \nabla f(x)]^\top (y - x)\, dt$$

3. Apply Cauchy-Schwarz and L-smoothness:

$$\left| f(y) - f(x) - \nabla f(x)^\top (y - x) \right| \leq \int_0^1 Lt\|y - x\|^2\, dt = \frac{L}{2}\|y - x\|^2$$

4. Combine inequalities to get final result. □

2.3 Geometric Interpretation

The Descent Lemma establishes that L-smooth functions:

- Have **quadratic upper bounds** on their growth
- Cannot deviate too far from their linear approximations
- Permit controlled descent steps in gradient methods

2.4 Connection to Gradient Descent

For gradient descent update $y = x - \eta \nabla f(x)$:

$$f(x_{k+1}) \leq f(x_k) - \eta\|\nabla f(x_k)\|^2 + \frac{L\eta^2}{2}\|\nabla f(x_k)\|^2$$

Theorem 4 (Descent Guarantee)**.** *For step size $\eta \leq 1/L$:*

$$f(x_{k+1}) \leq f(x_k) - \frac{\eta}{2}\|\nabla f(x_k)\|^2$$

2.5 Special Case: Quadratic Functions

For $f(x) = \frac{1}{2}x^\top Q x$ with $Q \succ 0$:

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2}(y - x)^\top Q (y - x)$$

Here $L = \lambda_{\max}(Q)$, and the Descent Lemma becomes exact.

2.6 Limitations and Caveats

- **Critical Step Size:** Fails for $\eta > 1/L$
- **Non-Smooth Functions:** Does not apply to non-differentiable functions
- **Local Property:** Only describes local behavior

3 Stepsize Selection

3.1 Fixed Stepsize

Definition 5 (L-Smoothness)**.** *A function f is L-smooth if:*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall x, y$$

Theorem 6 (Optimal Fixed Stepsize)**.** *For L-smooth functions, GD with $\eta = \frac{1}{L}$ guarantees:*

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2$$

Proof. From the Descent Lemma:

$$f(x_{k+1}) \leq f(x_k) - \eta\|\nabla f(x_k)\|^2 + \frac{L\eta^2}{2}\|\nabla f(x_k)\|^2$$

Substitute $\eta = \frac{1}{L}$:

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 \qquad \square$$

[Quadratic Function] For $f(x) = \frac{1}{2}x^T Q x$, optimal stepsize is $\eta = 2/(\lambda_{\max}(Q) + \lambda_{\min}(Q))$. With $\eta = 1/\lambda_{\max}(Q)$, convergence rate becomes:

$$\|x^k - x^*\| \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|x^0 - x^*\|$$

where $\kappa = \lambda_{\max}(Q)/\lambda_{\min}(Q)$.

3.2 Adaptive Stepsize Methods

3.2.1 Backtracking Line Search

Algorithm 1 Backtracking Line Search

```
Initialize η = η_init, α ∈ (0, 0.5), β ∈ (0, 1)
while f(x − η∇f(x)) > f(x) − αη‖∇f(x)‖² do
    η ← βη
end while
```

Geometric Interpretation: Maintains Armijo condition ensuring sufficient decrease:

$$f(x^+) \leq f(x) - \alpha\eta\|\nabla f(x)\|^2 \tag{1}$$

Theorem 7 (Convergence with Backtracking)**.** *For L-smooth f, backtracking GD with $\eta_{init} > 0$ achieves:*

$$\min_{1 \leq t \leq k} \|\nabla f(x_t)\|^2 \leq \frac{C(f(x_0) - f_*)}{k}$$

where C depends on α, β.

3.2.2 Exact Line Search

$$\eta_k = \arg\min_{\eta>0} f(x_k - \eta \nabla f(x_k))$$

[Rayleigh Quotient] For quadratic $f(x) = \frac{1}{2}x^T Q x$, exact step:

$$\eta = \frac{\nabla f(x)^T \nabla f(x)}{\nabla f(x)^T Q \nabla f(x)}$$

3.3 Stepsize for Different Function Classes

Function Class	**Recommended Stepsize**	**Convergence Rate**
Non-convex L-smooth	$\eta = 1/L$	$O(1/\sqrt{k})$
Convex L-smooth	$\eta = 1/L$	$O(1/k)$
μ-strongly convex	$\eta = 2/(\mu + L)$	Linear $O(\gamma^k)$

Table 1: Stepsize selection guide

3.4 Practical Considerations

- For unknown L, use backtracking with $\eta_{\text{init}} = 1$
- Monitor function values: $f(x_{k+1}) < f(x_k)$
- In deep learning: Use adaptive methods (Adam, RMSProp) with step decay

Theorem 8 (Safe Initialization)**.** *For any L-smooth function, backtracking line search with $\eta_{init} \geq 1/L$ will accept $\eta \geq \beta/L$ within $\lceil \log_\beta(1/L\eta_{init}) \rceil$ steps.*

4 Metric Projection

Definition 9 (Metric Projection)**.** *Let $C \subseteq \mathbb{R}^n$ be a closed convex set. The metric projection of a point $x \in \mathbb{R}^n$ onto C is defined as:*

$$\Pi_C(x) := \arg\min_{y \in C} \|y - x\|_2$$

This is the unique point in C closest to x under the Euclidean norm.

4.1 Existence and Uniqueness

Theorem 10 (Existence and Uniqueness)**.** *For any closed convex set $C \subseteq \mathbb{R}^n$ and $x \in \mathbb{R}^n$, the metric projection $\Pi_C(x)$ exists and is unique.*

Proof. **Existence:** The function $f(y) = \|y-x\|^2$ is coercive and strictly convex. Since C is closed, a minimizer exists.
Uniqueness: Strict convexity of f guarantees uniqueness. If y_1, y_2 were both minimizers, then $\frac{y_1+y_2}{2} \in C$ (by convexity) would yield a lower function value, contradicting minimality. □

4.2 Characterizing Inequality

Theorem 11 (Projection Inequality)**.** *For any $x \in \mathbb{R}^n$ and $y \in C$:*

$$\langle x - \Pi_C(x), y - \Pi_C(x) \rangle \leq 0$$

Equivalently:

$$\|x - \Pi_C(x)\|^2 \leq \|x - y\|^2 - \|\Pi_C(x) - y\|^2$$

Proof. Let $z = \Pi_C(x)$. From the first-order optimality condition:

$$\nabla f(z)^T (y - z) \geq 0 \quad \forall y \in C$$

Since $\nabla f(z) = 2(z - x)$, this becomes:

$$\langle z - x, y - z \rangle \geq 0 \quad \Rightarrow \quad \langle x - z, y - z \rangle \leq 0$$

The equivalence follows by expanding $\|x - y\|^2 = \|x - z + z - y\|^2$. □

4.3 Normal Cone Interpretation

Theorem 12 (Normal Cone Characterization)**.** *$z = \Pi_C(x)$ if and only if:*

$$x - z \in N_C(z)$$

where $N_C(z)$ is the normal cone to C at z:

$$N_C(z) := \{v \in \mathbb{R}^n \mid \langle v, y - z \rangle \leq 0 \ \forall y \in C\}$$

4.4 Examples of Metric Projections

[Common Projections]

- **Affine set**: For $C = \{x \mid Ax = b\}$:

  $$\Pi_C(x) = x - A^\dagger (Ax - b)$$

  where $A^\dagger$ is the Moore-Penrose pseudoinverse.

- **Non-negative orthant**: For $C = \mathbb{R}^n_+$:

  $$(\Pi_C(x))_i = \max(x_i, 0)$$

- **Euclidean ball**: For $C = \{y \mid \|y\| \leq r\}$:

  $$\Pi_C(x) = \begin{cases} x & \text{if } \|x\| \leq r \\ r \frac{x}{\|x\|} & \text{otherwise} \end{cases}$$

4.5 Projected Gradient Descent

Algorithm 2 Projected Gradient Descent

1: Initialize $x_0 \in C$, step size $\eta > 0$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: $\quad y_{k+1} = x_k - \eta \nabla f(x_k)$
4: $\quad x_{k+1} = \Pi_C(y_{k+1})$
5: **end for**

Theorem 13 (Convergence Guarantee)**.** *If f is L-smooth and convex, with $\eta = 1/L$:*

$$f(x_k) - f_\star \leq \frac{L\|x_0 - x_\star\|^2}{2k}$$

4.6 Key Lemma for Analysis

Lemma 14 (Projection Contraction)**.** *For any $x \in \mathbb{R}^n$ and $z \in C$:*

$$\|\Pi_C(x) - z\|^2 \leq \|x - z\|^2 - \|\Pi_C(x) - x\|^2$$

Proof. Using the characterizing inequality with $y = z$:

$$\|x - \Pi_C(x)\|^2 \leq \|x - z\|^2 - \|\Pi_C(x) - z\|^2$$

Rearranging gives the result. □

5 Convergence Rates

5.1 Non-Convex Functions

For L-smooth f, GD achieves:

$$\min_{0 \leq t \leq k} \|\nabla f(x_t)\|^2 \leq \frac{2L(f(x_0) - f_\star)}{k}$$

Rate: $O(1/\sqrt{k})$.

5.2 Convex Functions

For L-smooth convex f, GD achieves:

$$f(x_k) - f_\star \leq \frac{L\|x_0 - x_\star\|^2}{2k}$$

Rate: $O(1/k)$.

5.3 Strongly Convex Functions

For μ-strongly convex and L-smooth f, GD achieves linear convergence:

$$f(x_k) - f_\star \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f_\star)$$

Rate: $O(\log(1/\epsilon))$ iterations to reach ϵ-accuracy.

Theorem 15 (Convergence under PL Condition)**.** *If f satisfies Polyak-Łojasiewicz inequality:*

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f_\star)$$

then GD converges linearly even without strong convexity.