Gradient Descent

Cong Ma

University of Chicago, Winter 2026

Outline

- Gradient descent algorithm
- Smooth problems
- Convex and smooth problems
- Strongly convex and smooth problems
- Backtracking line search
- Preconditioned GD

Problem Setup

We consider unconstrained optimization problems of the form

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}),$$

where:

- $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable
- Gradient $\nabla f(\boldsymbol{x})$ is available

Goal: find a point $x^\star$ such that $\nabla f(\boldsymbol{x}^\star) = 0$, which we assume exists.

Descent Directions

Definition 1 (Descent Direction)

A vector d is a *descent direction* for f at x if

$$f(x + td) < f(\boldsymbol{x})$$

for all sufficiently small $t > 0$.

A simple sufficient characterization is given by the following result.

Lemma 2

If f is continuously differentiable in a neighborhood of x, then any direction d such that

$$d^\top \nabla f(\boldsymbol{x}) < 0$$

is a descent direction.

Steepest Descent Direction

Among all unit directions:

$$\min_{\|d\|=1} \ \nabla f(\boldsymbol{x})^\top d$$

Solution:

$$d = -\frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|}$$

Therefore the steepest descent direction is:

$$d = -\nabla f(\boldsymbol{x})$$

(up to scaling)

Gradient Descent Algorithm

Basic idea: move in the direction of negative gradient

Given an initial point x^0, iterate:

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k),$$

where:

- $\alpha_k > 0$ is the step size (learning rate)
- $k = 0, 1, 2, \ldots$

Gradient descent is a first-order method: it uses only gradient information.

Proximal View of Gradient Descent

Gradient descent can be viewed as:

$$x^{k+1} = \arg\min_y \Big\{ \underbrace{f(x^k) + \nabla f(x^k)^\top (y - x^k)}_{\text{first-order approx.}} + \underbrace{\frac{1}{2\alpha_k}\|y - x^k\|_2^2}_{\text{proximal term}} \Big\}.$$

Smooth Functions

Definition 3

f is L-smooth if

$$\|\nabla f(\boldsymbol{x}) - \nabla f(y)\| \leq L\|x - y\|$$

for all x, y.

Equivalent inequality:

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

Second-order characterization

$$\|\nabla^2 f(\boldsymbol{x})\|_2 \ \leq \ L, \qquad \forall x \quad \text{(for twice differentiable functions)}$$

Descent Lemma

Lemma 4 (Smoothness Upper Bound)

Assume f is L-smooth. Then for any x, direction d, and stepsize α,

$$f(x + \alpha d) \leq f(\boldsymbol{x}) + \alpha \nabla f(\boldsymbol{x})^\top d + \frac{L\alpha^2}{2} \|d\|^2.$$

This follows from Taylor expansion and Lipschitz continuity of the gradient.

Applying the Descent Lemma

Choose the gradient descent direction:

$$d = -\nabla f(\boldsymbol{x}).$$

Substitute into the lemma:

$$f(x - \alpha \nabla f(\boldsymbol{x})) \leq f(\boldsymbol{x}) - \alpha \|\nabla f(\boldsymbol{x})\|^2 + \frac{L\alpha^2}{2} \|\nabla f(\boldsymbol{x})\|^2.$$

The right-hand side is minimized at

$$\alpha = \frac{1}{L}.$$

Then gradient descent satisfies:

$$f(x^{k+1}) = f\left(x^k - \frac{1}{L} \nabla f(x^k)\right) \leq f(x^k) - \frac{1}{2L} \|\nabla f(x^k)\|^2.$$

Aggregating the One-Step Decrements

From the one-step descent inequality,

$$f(x^{k+1}) \le f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2,$$

assume f is lower bounded:

$$f(\boldsymbol{x}) \ge \bar{f}.$$

Summing over $k = 0, \ldots, T-1$ and telescoping gives

$$f(x^T) \le f(x^0) - \frac{1}{2L}\sum_{k=0}^{T-1}\|\nabla f(x^k)\|^2.$$

Using $f(x^T) \ge \bar{f}$, we obtain

$$\sum_{k=0}^{T-1}\|\nabla f(x^k)\|^2 \le 2L\big(f(x^0) - \bar{f}\big).$$

Asymptotic Stationarity

From bounded sum:

$$\sum_{k=0}^{\infty} \|\nabla f(x^k)\|^2 < \infty$$

it follows that

$$\lim_{k\to\infty} \|\nabla f(x^k)\| = 0$$

Interpretation

Gradient descent converges to a stationary point (not necessarily global minimum).

Rate of Convergence

From averaging:

$$\min_{0 \le k \le T-1} \|\nabla f(x^k)\|^2 \le \frac{1}{T} \sum_{k=0}^{T-1} \|\nabla f(x^k)\|^2$$

Using previous bound:

$$\min_{0 \le k \le T-1} \|\nabla f(x^k)\| \le \sqrt{\frac{2L\big(f(x^0) - \bar{f}\big)}{T}}$$

This gives an $O(T^{-1/2})$ stationarity rate.

Convex and Smooth Functions

We now assume:

- f is **convex**
- f is L-**smooth**
- Global minimizer x^* exists

Define optimal value:

$$f^* = f(x^*)$$

We analyze gradient descent with constant stepsize

$$\alpha = \frac{1}{L}.$$

Convergence Rate (Convex Case)

Theorem 5

Suppose f is convex and L-smooth, and let x^* be a minimizer. Then gradient descent with stepsize $\alpha = 1/L$ satisfies:

$$f(x^T) - f^* \leq \frac{L}{2T}\|x^0 - x^*\|^2, \qquad T = 1, 2, \ldots$$

Proof

By convexity,

$$f(\boldsymbol{x}^\star) \geq f(x^k) + \nabla f(x^k)^\top (x^\star - x^k) \Rightarrow \nabla f(x^k)^\top (x^k - x^\star) \geq f(x^k) - f^\star.$$

Combining with descent:

$$f(x^{k+1}) \leq f(\boldsymbol{x}^\star) + \nabla f(x^k)^\top (x^k - x^\star) - \frac{1}{2L}\|\nabla f(x^k)\|^2$$

which implies

$$f(x^{k+1}) \leq f(\boldsymbol{x}^\star) + \frac{L}{2}\Big(\|x^k - x^\star\|^2 - \|x^{k+1} - x^\star\|^2\Big).$$

Proof

Summing over $k = 0, \ldots, T-1$ gives

$$\sum_{k=0}^{T-1} (f(x^{k+1}) - f^\star) \leq \frac{L}{2} \|x^0 - x^\star\|^2.$$

Since $f(x^k)$ is nonincreasing,

$$f(x^T) - f^\star \leq \frac{1}{T} \sum_{k=0}^{T-1} (f(x^{k+1}) - f^\star).$$

Therefore,

$$f(x^T) - f^\star \leq \frac{L}{2T} \|x^0 - x^\star\|^2.$$

Strongly Convex Functions

f is μ-strongly convex if

$$f(y) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (y - x) + \frac{\mu}{2}\|y - x\|_2^2,$$

Equivalent second-order characterization

- $$\nabla^2 f(\boldsymbol{x}) \succeq \mu I, \qquad \forall x \quad \text{(for twice differentiable functions)}$$

Linear convergence of gradient descent

Theorem 6 (Linear convergence for m-strongly convex and L-smooth f)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, m-strongly convex, and L-smooth. Consider gradient descent with constant stepsize $\eta = 1/L$:

$$x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k).$$

Let $x^\star \in \arg\min_x f(\boldsymbol{x})$ and $f^\star := f(\boldsymbol{x}^\star)$. Then for all $k \geq 0$,

$$f(x^{k+1}) - f^\star \leq \left(1 - \frac{m}{L}\right)(f(x^k) - f^\star).$$

Consequently, after T iterations,

$$f(x^T) - f^\star \leq \left(1 - \frac{m}{L}\right)^T (f(x^0) - f^\star).$$

A Key Lemma

Lemma 7

Let f be continuously differentiable and m-strongly convex. Then the following inequalities hold:

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^\star) \;\le\; \frac{\|\nabla f(\boldsymbol{x})\|^2}{2m},$$

and

$$\|x - x^\star\| \;\le\; \frac{2}{m}\|\nabla f(\boldsymbol{x})\|.$$

Proof

Suppose f is m-strongly convex. Minimizing both sides of the strong convexity inequality with respect to z, we obtain

$$f(\boldsymbol{x}^\star) \geq f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top \left(\frac{1}{m}\nabla f(\boldsymbol{x})\right) + \frac{m}{2}\left\|\frac{1}{m}\nabla f(\boldsymbol{x})\right\|^2.$$

Simplifying,

$$f(\boldsymbol{x}^\star) = f(\boldsymbol{x}) - \frac{1}{2m}\|\nabla f(\boldsymbol{x})\|^2.$$

Rearranging the previous inequality yields

$$\|\nabla f(\boldsymbol{x})\|^2 \;\geq\; 2m\big(f(\boldsymbol{x}) - f(\boldsymbol{x}^\star)\big).$$

Distance to Optimum via Gradient

We estimate the distance to the optimizer $x^\star$ using strong convexity and the Cauchy–Schwarz inequality.

From strong convexity,

$$f(\boldsymbol{x}^\star) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (x^\star - x) + \frac{m}{2}\|x - x^\star\|^2.$$

Applying Cauchy–Schwarz,

$$\nabla f(\boldsymbol{x})^\top (x^\star - x) \geq -\|\nabla f(\boldsymbol{x})\| \, \|x^\star - x\|.$$

Therefore,

$$f(\boldsymbol{x}^\star) \geq f(\boldsymbol{x}) - \|\nabla f(\boldsymbol{x})\| \, \|x^\star - x\| + \frac{m}{2}\|x - x^\star\|^2.$$

Rearranging the previous inequality yields

$$\|x - x^\star\| \; \leq \; \frac{2}{m} \, \|\nabla f(\boldsymbol{x})\|.$$

Proof of the Theorem

By smoothness, we have

$$f(x^{k+1}) = f\left(x^k - \frac{1}{L}\nabla f(x^k)\right) \le f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2.$$

By strong convexity,

$$f(x^{k+1}) \le f(x^k) - \frac{m}{L}(f(x^k) - f^\star),$$

where $f^\star = f(\boldsymbol{x}^\star)$. Subtracting $f^\star$ from both sides yields the recursion

$$f(x^{k+1}) - f^\star \le \left(1 - \frac{m}{L}\right)(f(x^k) - f^\star).$$

Thus, the function values converge linearly to the optimum.

After T iterations,

$$f(x^T) - f^\star \le \left(1 - \frac{m}{L}\right)^T (f(x^0) - f^\star).$$

Iteration Complexity of Gradient Descent

Setting	**Goal**	**Iterations Required**
Smooth nonconvex	$\|\nabla f(x^k)\| \leq \varepsilon$	$\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$
Convex	$f(x^k) - f^\star \leq \varepsilon$	$\mathcal{O}\left(\frac{1}{\varepsilon}\right)$
Strongly convex	$f(x^k) - f^\star \leq \varepsilon$	$\mathcal{O}\left(\log \frac{1}{\varepsilon}\right)$

- Higher curvature assumptions $\Rightarrow$ faster convergence.
- Strong convexity yields linear convergence.

Polyak–Łojasiewicz Inequality

Definition 8 (Polyak–Łojasiewicz (PL))

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and let $\alpha > 0$. We say that f satisfies the PL inequality with constant α if

$$\|\nabla f(\boldsymbol{x})\|_2^2 \;\geq\; 2\alpha\big(f(\boldsymbol{x}) - f^\star\big), \qquad \forall x \in \mathbb{R}^d,$$

where $f^\star := \min_x f(\boldsymbol{x})$ (equivalently, $f^\star = f(\boldsymbol{x}^\star)$ for any minimizer $x^\star$).

- A first-order condition: controls suboptimality by gradient magnitude.
- Does *not* require convexity.

Strong Convexity Implies PL Condition

Lemma 9 (Strong convexity $\Rightarrow$ PL)

Let $f : \mathbb{R}^d \to \mathbb{R}$ and let $\alpha > 0$. If f is α-strongly convex, then f satisfies the PL inequality with constant α.

PL Condition Does Not Imply Convexity

Consider

$$f(x) = x^2 + 3\sin^2(x), \qquad x \in \mathbb{R}.$$

- **Nonconvex:**

$$f''(x) = 2 + 6\cos(2x), \qquad f''(\pi/2) = -4 < 0.$$

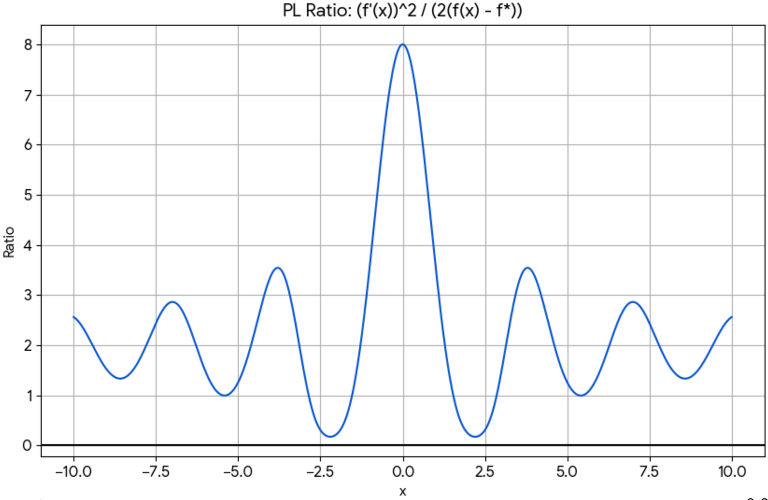
- **PL holds:**

$$f'(x) = 2x + 3\sin(2x), \quad f^\star = 0,$$

  and one can show

$$\frac{|f'(x)|^2}{2f(x)} \ \geq \ 0.0001.$$

A Picture Proof



Convergence Under the PL Condition

Theorem 10 (Linear rate under PL + smoothness)

Assume f is L-smooth and satisfies the PL inequality

$$\|\nabla f(\boldsymbol{x})\|_2^2 \geq 2\mu(f(\boldsymbol{x}) - f^\star).$$

With constant stepsize $\eta_t \equiv \eta = \frac{1}{L}$, gradient descent satisfies

$$f(x^t) - f^\star \ \leq \ \left(1 - \frac{\mu}{L}\right)^t (f(x^0) - f^\star).$$

- Linear convergence of objective values.
- PL does not imply a unique minimizer (only $f(x^t) \to f^\star$) cf. strong covnexity

Example: Over-Parameterized Linear Regression

- Data: $\{(a_i, y_i)\}_{i=1}^m$ with $a_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$.
- Least squares objective:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad f(\boldsymbol{x}) = \frac{1}{2} \sum_{i=1}^{m} (a_i^\top x - y_i)^2 \;=\; \frac{1}{2} \|Ax - y\|_2^2, \qquad A = \begin{bmatrix} a_1^\top \\ \vdots \\ a_m^\top \end{bmatrix}.$$

Over-parameterization: $n > m$ (more parameters than samples).

— a regime of particular importance in modern ML —

Example: Over-Parameterized Linear Regression

$$\nabla f(\boldsymbol{x}) = A^\top(Ax - y), \qquad \nabla^2 f(\boldsymbol{x}) = A^\top A.$$

- f is convex, but not strongly convex when $n > m$ since $A^\top A$ is rank-deficient.
- In many non-degenerate cases, the system is consistent and

$$f^\star = 0.$$

- Nevertheless, f satisfies a PL inequality (with constant depending on $AA^\top$).

$\Longrightarrow$ Gradient descent achieves a linear rate in objective value.

Example: Over-Parameterized Linear Regression

Fact 2.6 (linear rate)

Assume $A \in \mathbb{R}^{m \times n}$ has rank m and take a constant stepsize

$$\eta_t \equiv \eta = \frac{1}{\lambda_{\max}(AA^\top)}.$$

Then gradient descent satisfies, for all t,

$$f(x^t) - f^\star \;\leq\; \left(1 - \frac{\lambda_{\min}(AA^\top)}{\lambda_{\max}(AA^\top)}\right)^t (f(x^0) - f^\star).$$

- Mild condition on $\{a_i\}$ (full row rank).
- No condition on $\{y_i\}$.

Proof of Fact 2.6

Key step: prove the PL inequality

$$\|\nabla f(\boldsymbol{x})\|_2^2 \;\geq\; 2\,\lambda_{\min}(AA^\top)\,f(\boldsymbol{x}). \tag{2.9}$$

Once (2.9) holds, Theorem (PL + smoothness) implies the linear rate. Here $f^\star = 0$.

Let $f(\boldsymbol{x}) = \frac{1}{2}\|Ax - y\|_2^2$, so $\nabla f(\boldsymbol{x}) = A^\top(Ax - y)$. Then

$$\begin{aligned}\|\nabla f(\boldsymbol{x})\|_2^2 &= (Ax - y)^\top AA^\top (Ax - y) \\ &\geq \lambda_{\min}(AA^\top)\,\|Ax - y\|_2^2 \\ &= 2\,\lambda_{\min}(AA^\top)\,f(\boldsymbol{x}),\end{aligned}$$

which is exactly (2.9) with $\mu = \lambda_{\min}(AA^\top)$.

Convergence in Iterates
What About $\|\boldsymbol{x}^t - \boldsymbol{x}^\star\|_2$?

Strongly Convex and Smooth Problems

Theorem 11 (Gradient Descent for Strongly Convex and Smooth Functions)

Let f be μ-strongly convex and L-smooth. If the step size is chosen as

$$\eta_t \equiv \eta = \frac{2}{\mu + L},$$

then gradient descent satisfies

$$\|x^t - x^\star\|_2 \;\leq\; \left(\frac{\kappa - 1}{\kappa + 1}\right)^t \|x^0 - x^\star\|_2,$$

where

$$\kappa := \frac{L}{\mu}$$

is the condition number and $x^\star$ is the global minimizer.

Proof

Step 1: Fundamental theorem of calculus

$$\nabla f(x^t) = \nabla f(x^t) - \nabla f(x^\star) = \left(\int_0^1 \nabla^2 f(x_\tau)\, d\tau\right)(x^t - x^\star),$$

where

$$x_\tau := x^t + \tau(x^\star - x^t),$$

which parameterizes the line segment between x^t and $x^\star$.

Step 2: One-step contraction
Using the GD update $x^{t+1} = x^t - \eta \nabla f(x^t)$:

$$\begin{aligned}
\|x^{t+1} - x^\star\|_2 &= \|x^t - x^\star - \eta \nabla f(x^t)\|_2 \\
&= \left\|\left(I - \eta \int_0^1 \nabla^2 f(x_\tau)\, d\tau\right)(x^t - x^\star)\right\|_2 \\
&\leq \sup_{0 \leq \tau \leq 1} \|I - \eta \nabla^2 f(x_\tau)\|_2 \, \|x^t - x^\star\|_2.
\end{aligned}$$

Proof

Step 3: Use smoothness and strong convexity
Since

$$\mu I \preceq \nabla^2 f(x_\tau) \preceq LI,$$

and $\eta = \frac{2}{\mu+L}$, we obtain

$$\|I - \eta \nabla^2 f(x_\tau)\|_2 \ \le \ \frac{L-\mu}{L+\mu}.$$

Conclusion:

$$\|x^{t+1} - x^\star\|_2 \ \le \ \frac{L-\mu}{L+\mu}\|x^t - x^\star\|_2.$$

Iterating yields linear convergence.

Convex and smooth problems

$\|x^t - x^\star\|_2$ is monotonically nonincreasing in t

Formal Statement

Treating f as **0-strongly convex** (i.e., convex), our previous analysis implies

$$\|x^{t+1} - x^\star\|_2 \;\leq\; \|x^t - x^\star\|_2,$$

provided the step size satisfies $\eta_t \leq \frac{1}{L}$.

Distance Decrease for Convex and Smooth Functions

Fact (Monotonic Improvement of Iterates)

Let f be convex and L-smooth. If the step size is chosen as

$$\eta_t \equiv \eta = \frac{1}{L},$$

then gradient descent satisfies

$$\|x^{t+1} - x^\star\|_2^2 \;\le\; \|x^t - x^\star\|_2^2 \;-\; \frac{1}{L^2}\,\|\nabla f(x^t)\|_2^2,$$

where $x^\star$ is any minimizer of f.

Proof of Distance Decrease (Fact 2.8)

Since $\nabla f(x^\star) = 0$, we have

$$\begin{aligned}\|x^{t+1} - x^\star\|_2^2 &= \|x^t - x^\star - \eta(\nabla f(x^t) - \nabla f(x^\star))\|_2^2 \\ &= \|x^t - x^\star\|_2^2 - 2\eta\langle x^t - x^\star, \nabla f(x^t) - \nabla f(x^\star)\rangle \\ &\quad + \eta^2\|\nabla f(x^t) - \nabla f(x^\star)\|_2^2.\end{aligned}$$

Use convexity + smoothness:
For convex and L-smooth f,

$$\langle x^t - x^\star, \nabla f(x^t) - \nabla f(x^\star)\rangle \ \geq \ \frac{1}{L}\|\nabla f(x^t) - \nabla f(x^\star)\|_2^2.$$

Therefore,

$$\begin{aligned}\|x^{t+1} - x^\star\|_2^2 \leq \|x^t - x^\star\|_2^2 &- \frac{2\eta}{L}\|\nabla f(x^t) - \nabla f(x^\star)\|_2^2 \\ &+ \eta^2\|\nabla f(x^t) - \nabla f(x^\star)\|_2^2.\end{aligned}$$

Plug in $\eta = 1/L$:

$$\|x^{t+1} - x^\star\|_2^2 = \|x^t - x^\star\|_2^2 - \frac{1}{L^2}\|\nabla f(x^t)\|_2^2.$$

Backtracking Line Search

Why Backtracking Line Search?

- Constant step size $\eta = 1/L$ needs (an estimate of) L.
- In practice L is unknown; too large η can cause oscillation/divergence.
- **Idea:** start with a candidate step size and shrink it until we get a guaranteed decrease in f.

Goal

Pick η_t adaptively so that each step makes **measurable progress**:

$$f(x^{t+1}) \leq f(x^t) - \text{(something positive)}.$$

Backtracking Line Search: Armijo Condition

Descent direction

Take the gradient step direction

$$d^t = -\nabla f(x^t), \qquad x^{t+1} = x^t + \eta_t d^t.$$

Armijo (sufficient decrease) condition

Choose η_t such that

$$f(x^t + \eta_t d^t) \;\le\; f(x^t) + c\,\eta_t \langle \nabla f(x^t), d^t \rangle,$$

where $c \in (0, 1)$.

- For $d^t = -\nabla f(x^t)$, this becomes

$$f(x^t - \eta_t \nabla f(x^t)) \;\le\; f(x^t) - c\,\eta_t \|\nabla f(x^t)\|^2.$$

- Easy to check: evaluate $f(\cdot)$ at the trial point.

Algorithm: Backtracking (Gradient Descent)

Inputs

Initial step $\eta_0 > 0$, shrink factor $\beta \in (0, 1)$, Armijo parameter $c \in (0, 1)$.

At iteration t

1. Set $\eta \leftarrow \eta_0$ (or reuse previous step size).
2. While Armijo fails:

$$f(x^t - \eta \nabla f(x^t)) \; > \; f(x^t) - c\,\eta \|\nabla f(x^t)\|^2,$$

   update $\eta \leftarrow \beta\eta$.
3. Set $x^{t+1} = x^t - \eta \nabla f(x^t)$ and $\eta_t = \eta$.

Typical choices

$\beta = 0.5$ or 0.8, $\quad c = 10^{-4}$.

Why It Works: Finite Termination (Smooth Case)

Assume f is L-smooth. Then for any x and any $\eta > 0$,

$$f(x - \eta\nabla f(x)) \le f(x) - \eta\Big(1 - \frac{L\eta}{2}\Big)\|\nabla f(x)\|^2.$$

- If $\eta \le \frac{1}{L}$, then $1 - \frac{L\eta}{2} \ge \frac{1}{2}$, so

$$f(x - \eta\nabla f(x)) \le f(x) - \frac{\eta}{2}\|\nabla f(x)\|^2.$$

- Therefore Armijo holds automatically whenever

$$\frac{\eta}{2} \ \ge \ c\,\eta \quad \Longleftrightarrow \quad c \le \frac{1}{2},$$

and $\eta \le 1/L$.

Conclusion

Backtracking will stop after finitely many shrink steps and returns a step size η_t that guarantees descent.

What Guarantees Do We Get?

With Armijo backtracking on L-smooth f (using $d^t = -\nabla f(x^t)$):

- **Monotone decrease:** $f(x^{t+1}) \leq f(x^t)$.
- **Sufficient decrease:**

$$f(x^{t+1}) \leq f(x^t) - c\,\eta_t \|\nabla f(x^t)\|^2.$$

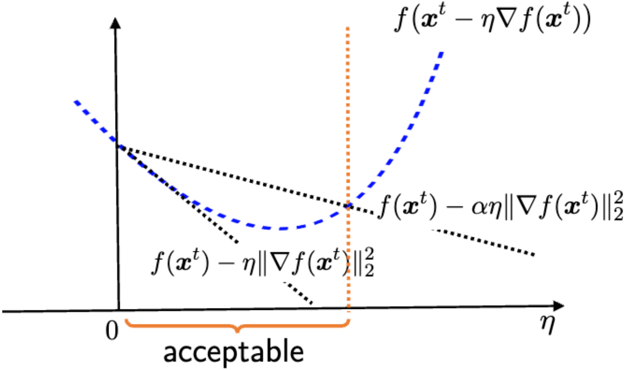
- Summing over t yields

$$\sum_{t \geq 0} \eta_t \|\nabla f(x^t)\|^2 \;<\; \infty \quad \Rightarrow \quad \|\nabla f(x^t)\| \to 0 \text{ (under mild conditions)}$$

Key message

Backtracking gives **automatic step-size selection** with **provable progress**, without knowing L.

Backtracking Line Search



Preconditioned Gradient Descent

Quadratic Optimization: Rate with Stepsize $1/L$

Consider

$$f(\boldsymbol{x}) = \frac{1}{2}x^\top Qx, \qquad Q = Q^\top \succ 0.$$

Let $0 < \mu = \lambda_{\min}(Q) \leq \lambda_{\max}(Q) = L$. Then gradient descent with $\eta = \frac{1}{L}$ satisfies

$$f(x^t) - f^\star \;\leq\; \left(1 - \frac{\mu}{L}\right)^t (f(x^0) - f^\star).$$

- Linear rate governed by condition number $\kappa = L/\mu$.
- Same form as PL rate (quadratics satisfy PL with $\mu = \lambda_{\min}(Q)$).

Exact Line Search

Idea: pick the stepsize that minimizes f along the descent direction.

Given x^t, take $d^t := -\nabla f(x^t)$ and choose

$$\eta_t \in \arg\min_{\eta \geq 0} \ f(x^t + \eta d^t) \quad \Longleftrightarrow \quad \eta_t \in \arg\min_{\eta \geq 0} \ f(x^t - \eta \nabla f(x^t)).$$

Update:

$$x^{t+1} = x^t - \eta_t \nabla f(x^t).$$

- Guarantees monotone decrease: $f(x^{t+1}) \leq f(x^t)$.
- Parameter-free stepsize; especially clean for quadratics.

Exact Line Search for Quadratic Objectives

Consider

$$f(\boldsymbol{x}) = \frac{1}{2}x^\top Q x, \qquad Q = Q^\top \succeq 0.$$

Let $g^t := \nabla f(x^t) = Qx^t$. Exact line search solves

$$\eta_t \in \arg\min_{\eta \geq 0} \; f(x^t - \eta g^t).$$

Closed-form stepsize

$$\eta_t = \frac{\|g^t\|_2^2}{(g^t)^\top Q g^t}.$$

- A 1D convex quadratic in η.
- If $Q \succ 0$, the unique minimizer is $x^\star = 0$.

Exact Line Search: Convergence Rate (Quadratic)

Let $Q \succ 0$ with eigenvalues $0 < \lambda_n(Q) \le \cdots \le \lambda_1(Q)$. If

$$\eta_t = \arg\min_{\eta > 0} \; f(x^t - \eta \nabla f(x^t)), \qquad f(\boldsymbol{x}) = \frac{1}{2} x^\top Q x,$$

then

$$f(x^t) - f^\star \;\le\; \left(\frac{\lambda_1(Q) - \lambda_n(Q)}{\lambda_1(Q) + \lambda_n(Q)} \right)^{2t} (f(x^0) - f^\star).$$

- Objective-value rate; depends on $\kappa = \lambda_1/\lambda_n$ via $\frac{\kappa-1}{\kappa+1}$.
- Not faster (in worst case) than the constant stepsize rule.

Exact Line Search: Proof Sketch

Let $x^\star = 0$ and $g^t = \nabla f(x^t) = Qx^t$. Exact line search yields

$$\eta_t = \frac{(g^t)^\top g^t}{(g^t)^\top Q g^t}.$$

Compute

$$\begin{aligned} f(x^{t+1}) &= \frac{1}{2}(x^t - \eta_t g^t)^\top Q (x^t - \eta_t g^t) \\ &= f(x^t) - \eta_t \|g^t\|_2^2 + \frac{\eta_t^2}{2}(g^t)^\top Q g^t \\ &= f(x^t) - \frac{\|g^t\|_2^4}{2(g^t)^\top Q g^t} \\ &= \left(1 - \frac{\|g^t\|_2^4}{(g^t)^\top Q g^t \cdot (g^t)^\top Q^{-1} g^t}\right) f(x^t), \end{aligned}$$

using $f(x^t) = \frac{1}{2}(g^t)^\top Q^{-1} g^t$.

Exact Line Search: Proof Sketch (cont.)

Kantorovich inequality: for all $y \neq 0$,

$$\frac{\|y\|_2^4}{(y^\top Q y)(y^\top Q^{-1} y)} \;\geq\; \frac{4\lambda_1(Q)\lambda_n(Q)}{(\lambda_1(Q)+\lambda_n(Q))^2}.$$

Apply it with $y = g^t$:

$$f(x^{t+1}) \leq \left(1 - \frac{4\lambda_1(Q)\lambda_n(Q)}{(\lambda_1(Q)+\lambda_n(Q))^2}\right) f(x^t) = \left(\frac{\lambda_1(Q)-\lambda_n(Q)}{\lambda_1(Q)+\lambda_n(Q)}\right)^2 f(x^t).$$

Since $f^\star = 0$, iterating gives the stated rate. $\square$

Preconditioning via Linear Transformations

Ill-conditioned problems can slow down gradient methods.

A common remedy is to scale variables via a linear change of variables.

Consider $\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$ and let S be nonsingular. Define $x = Sy$ and $g(y) := f(Sy)$.

$$\min_x f(\boldsymbol{x}) \quad \Longleftrightarrow \quad \min_y g(y) = f(Sy).$$

Preconditioning via Linear Transformations

By the chain rule,

$$\nabla g(y) = S^\top \nabla f(Sy).$$

Applying gradient descent to g:

$$y_{k+1} = y_k - t_k S^\top \nabla f(Sy_k).$$

Multiplying by S and letting $x_k = Sy_k$ gives

$$x_{k+1} = x_k - t_k SS^\top \nabla f(x_k).$$

Scaled Gradient Method

Define the scaling (preconditioning) matrix

$$D := SS^\top \succ 0.$$

Then the update becomes

$$x_{k+1} = x_k - t_k D \nabla f(x_k).$$

Scaled gradient method

$$x_{k+1} = x_k - t_k D \nabla f(x_k), \qquad D \succ 0.$$

- Standard GD corresponds to $D = I$.
- Choosing D well can dramatically improve conditioning.

Scaled Gradient is a Descent Method

If $D \succ 0$ and $\nabla f(x_k) \neq 0$, then $-D\nabla f(x_k)$ is a descent direction:

$$f'(x_k; -D\nabla f(x_k)) = -\nabla f(x_k)^\top D\nabla f(x_k) < 0.$$

- Strict inequality uses positive definiteness of D.
- Any standard stepsize rule applies (constant, exact, backtracking).

Interpretation: Scaling Changes Geometry

Let $x = D^{1/2}y$ and define $g(y) = f(D^{1/2}y)$. Then

$$\nabla g(y) = D^{1/2}\nabla f(\boldsymbol{x}), \qquad \nabla^2 g(y) = D^{1/2}\nabla^2 f(\boldsymbol{x})\, D^{1/2}.$$

- The curvature is transformed to the scaled Hessian $D^{1/2}\nabla^2 f(\boldsymbol{x})D^{1/2}$.
- Choose D so the scaled Hessian is closer to I.

Practical Remarks

- Stepsize t_k can be chosen by:
  - constant stepsize,
  - exact line search,
  - backtracking line search.
- In large-scale problems, D is often chosen diagonal (cheap to store/apply).
- Allowing $D = D_k$ to change over time motivates adaptive scaling methods (AdaGrad / RMSProp / Adam) and quasi-Newton ideas.

Scaled Gradient Method: Template

Input

Tolerance $\varepsilon > 0$.

Initialization

Choose $x_0 \in \mathbb{R}^n$.

Iteration

For $k = 0, 1, 2, \ldots$:

1. Pick $D_k \succ 0$.
2. Choose t_k by line search on $g(t) = f(x_k - tD_k \nabla f(x_k))$.
3. Update $x_{k+1} = x_k - t_k D_k \nabla f(x_k)$.
4. Stop if $\|\nabla f(x_{k+1})\| \leq \varepsilon$.

Why Scaling Helps

The convergence rate depends on the conditioning of the scaled Hessian

$$D_k^{1/2}\nabla^2 f(x_k) D_k^{1/2}.$$

- Goal: make the scaled Hessian closer to I.
- When $\nabla^2 f(x_k) \succ 0$, the ideal choice is

$$D_k = (\nabla^2 f(x_k))^{-1},$$

yielding $D_k^{1/2}\nabla^2 f(x_k) D_k^{1/2} = I$.

Connection to Newton's Method

If $D_k = (\nabla^2 f(x_k))^{-1}$, then

$$x_{k+1} = x_k - t_k(\nabla^2 f(x_k))^{-1}\nabla f(x_k).$$

Newton step

With $t_k = 1$,

$$x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1}\nabla f(x_k).$$

Computationally, this requires solving the linear system

$$\nabla^2 f(x_k)\, d_k = \nabla f(x_k), \qquad x_{k+1} = x_k - d_k.$$

Diagonal Scaling: Motivation

Full Hessian information may be too expensive.

A simple alternative is diagonal scaling:

$$D_k = \operatorname{diag}(d_{k,1}, \ldots, d_{k,n}).$$

- Cheap to store and apply.
- Helps when variables have very different magnitudes/units.

Diagonal Scaling: A Natural Rule

A natural choice uses the diagonal of the Hessian:

$$(D_k)_{ii} = \left(\nabla^2 f(x_k)\right)_{ii}^{-1},$$

when $(\nabla^2 f(x_k))_{ii} > 0$.

With this choice, the scaled Hessian has unit diagonal:

$$\left(D_k^{1/2} \nabla^2 f(x_k) D_k^{1/2}\right)_{ii} = 1.$$

- Captures curvature coordinate-wise.
- Approximates Newton scaling using only diagonal information.