CME 307: Optimization CME 307 / MS&E 311 / OIT 676

Lecture 11: Gradient Descent

Fall 2025 *Prof. Udell*

1 Setup and conventions

We study gradient descent (GD) for unconstrained smooth optimization

$$\min_{x\in\mathbb{R}^n} f(x), \qquad f \text{ differentiable, with an attained optimal value } f^\star := \min_x f(x).$$

The basic iteration with constant step size $t > 0$ is

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

We also discuss line search strategies (e.g., Armijo backtracking) that choose t^k adaptively.

First-order optimality (recall). If $x^\star$ minimizes a differentiable f, then $\nabla f(x^\star) = 0$.

2 Quadratic upper bound: L-smoothness

Definition 2.1 (Smoothness)**.** A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is *L-smooth* if for all x, y,

$$f(y) \le f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2.$$

Equivalently (when $\nabla^2 f$ exists), $\|\nabla f(y) - \nabla f(x)\| \le L\|y-x\|$ and $\nabla^2 f(x) \preceq LI$ for all x in the domain.

Example 2.2 (Quadratic)**.** For $f(x) = \frac{1}{2}x^TAx$ with $A \succeq 0$, f is L-smooth with $L = \lambda_{\max}(A)$.

3 Quadratic lower bound: μ-strong convexity

Definition 3.1 (Strong convexity)**.** A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is *μ-strongly convex* if for all x, y,

$$f(y) \ge f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|^2.$$

Equivalently (when $\nabla^2 f$ exists), $\nabla^2 f(x) \succeq \mu I$; and the gradient is μ-coercive in the sense $\|\nabla f(y) - \nabla f(x)\| \ge \mu\|y-x\|$.

Example 3.2 (Quadratic). For $f(x) = \frac{1}{2}x^T Ax$ with $A \succeq 0$, f is μ-strongly convex with $\mu = \lambda_{\min}(A)$ if and only if $A \succ 0$.

4 Some important losses: smoothness and strong convexity

Example 4.1 (Least squares and logistic regression). Let $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$.

- Quadratic loss: $f(x) = \|Ax - b\|^2$ is smooth, and is strongly convex if A has full column rank ($\lambda_{\min}(A^T A) > 0$).
- Logistic loss: $f(x) = \sum_{i=1}^m \log\bigl(1 + \exp(b_i a_i^T x)\bigr)$ is smooth; it is strongly convex on any compact set when A has full column rank.

Worked details. For logistic loss, $\nabla^2 f(x) = A^T D(x)A$ with $D(x) = \operatorname{diag}\bigl(\sigma(s_i)(1 - \sigma(s_i))\bigr)$ and $s_i = b_i a_i^T x$, so $0 \preceq D(x) \preceq \frac{1}{4}I$, giving $L \le \frac{1}{4}\lambda_{\max}(A^T A)$. On bounded sets that keep $\sigma(s_i) \in [\delta, 1-\delta]$, $D(x) \succeq \delta(1-\delta)I$, giving $\mu \ge \delta(1-\delta)\lambda_{\min}(A^T A)$.

5 Choosing the next iterate by optimizing the upper bound

Minimizing the quadratic upper model at x^k yields

$$x^{k+1} = \operatorname*{argmin}_y \Bigl\{ f(x^k) + \nabla f(x^k)^T (y - x^k) + \tfrac{L}{2}\|y - x^k\|^2 \Bigr\} = x^k - \tfrac{1}{L}\nabla f(x^k).$$

Thus $t = 1/L$ is the natural stepsize when L is known. (We will prove it guarantees decrease.)

Remark 5.1 (Quadratic approximation viewpoint). Replacing the Hessian by $H = \frac{1}{t}I$ in the local quadratic model yields $x^+ = x - t\nabla f(x)$, i.e., gradient descent.

6 The Polyak-Łojasiewicz (PL) condition

Definition 6.1 (PL). A differentiable function f satisfies the *μ-PL inequality* if

$$\frac{1}{2}\|\nabla f(x)\|^2 \ge \mu\bigl(f(x) - f^\star\bigr) \qquad \text{for all } x.$$

PL does *not* require convexity and does *not* imply uniqueness of minimizers; under PL, objective convergence does not necessarily imply iterate convergence.

Proposition 6.2 (Strong convexity $\Rightarrow$ PL). *If f is μ-strongly convex, then f is μ-PL.*

Proof sketch. Minimize the strong convexity lower bound over y:

$$f^\star \ge \min_y \Bigl\{ f(x) + \nabla f(x)^T (y - x) + \tfrac{\mu}{2}\|y - x\|^2 \Bigr\} = f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2,$$

where the minimum is attained at $y = x - \nabla f(x)/\mu$. Rearranging gives the PL inequality. $\square$

Example 6.3 (Compositions that are PL). If $f(x) = g(Ax)$ with g strongly convex and A linear, then f satisfies a PL inequality (even when f is not strongly convex or convex) [**?**]. This covers least squares, and logistic regression on compact sets when A has full column rank.

7 Types and rates of convergence

Definition 7.1 (Objective and iterate convergence). We say GD achieves *objective convergence* if $f(x^k) \to f^\star$ and *iterate convergence* if $x^k \to x^\star$. Under strong convexity, objective convergence implies iterate convergence; under PL, not necessarily (the minimizer set may be a manifold).

Definition 7.2 (Rates). We say $f(x^k) - f^\star \leq c^k(f(x^0) - f^\star)$ for some $c \in (0,1)$ is *linear* (geometric) convergence, which appears as a straight line on a semilog plot; rates like $O(1/k)$ are *sublinear* and curve upward in semilog.

8 Main theorem: GD under L-smoothness and PL

Theorem 8.1 (GD is linearly convergent under PL). *If f is L-smooth and μ-PL, and $x^\star$ exists, then GD with $t = 1/L$ satisfies*

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

Proof. By L-smoothness with $x = x^k$ and $y = x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k)$,

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

By PL, $\|\nabla f(x^k)\|^2 \geq 2\mu\big(f(x^k) - f^\star\big)$; combine to get

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

and iterate. $\square$

Remark 8.2 (What improves with exact line search). Exact line search always does at least as well as $t = 1/L$ in function decrease, so the same linear rate bound holds (and can be faster in practice).

9 Sublinear rate on smooth convex functions

For completeness, we include the standard $O(1/k)$ rate for convex L-smooth f (no PL).

Theorem 9.1 (GD on L-smooth convex f)**.** *If f is convex and L-smooth, GD with $t = 1/L$ satisfies*

$$f(x^k) - f^\star \le \frac{L}{2k}, \|x^0 - x^\star\|^2.$$

Proof sketch. Combine the descent lemma $f(x^{k+1}) \le f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2$ with convexity, $f(x^k) - f^\star \le \nabla f(x^k)^T(x^k - x^\star)$, and nonexpansiveness of the GD step, to telescope $\|x^{k+1} - x^\star\|^2 \le \|x^k - x^\star\|^2 - \frac{2}{L}\big(f(x^k) - f^\star\big)$. Summing over k yields the bound. $\square$

10 Line search and guaranteed decrease

Definition 10.1 (Armijo backtracking)**.** Given $c \in (0,1)$ and shrinkage factor $\beta \in (0,1)$, set $t \leftarrow 1$ and decrease $t \leftarrow \beta t$ until

$$f(x - t\nabla f(x)) \le f(x) - ct\|\nabla f(x)\|^2.$$

Proposition 10.2 (Armijo accepts small enough steps)**.** *If f is L-smooth, then Armijo with any $c \le \frac{1}{2}$ accepts any $t \le 1/L$. In particular, the procedure always terminates.*

Proof. By L-smoothness, $f(x - tg) \le f(x) - t\|g\|^2 + \frac{L}{2}t^2\|g\|^2$ with $g = \nabla f(x)$. If $t \le 1/L$, then $-t + \frac{L}{2}t^2 \le -\frac{1}{2}t$, hence $f(x - tg) \le f(x) - \frac{1}{2}t\|g\|^2$, which is Armijo with $c \le \frac{1}{2}$. $\square$

11 Quadratics: spectral viewpoint and exact line search

Consider $f(x) = \frac{1}{2}x^TAx - b^Tx$ with $A \succ 0$ (unique minimizer $x^\star = A^{-1}b$).

- With constant $t \in \big(0, \frac{2}{\lambda_{\max}(A)}\big)$,

$$x^{k+1} - x^\star = (I - tA)(x^k - x^\star), \quad \|x^k - x^\star\|_A \le \rho^k\|x^0 - x^\star\|_A, \quad \rho = \max_i |1 - t\lambda_i(A)|.$$

- With *exact line search*,

$$t_k = \operatorname*{argmin}_{\alpha \ge 0} f(x^k - \alpha\nabla f(x^k)) = \frac{\|\nabla f(x^k)\|^2}{\nabla f(x^k)^T A \nabla f(x^k)}.$$

These formulas make the role of the condition number $\kappa = \lambda_{\max}/\lambda_{\min}$ explicit and explain zig-zagging in elongated valleys.

12 Practical convergence and local vs. global

Remark 12.1 (Exact line search dominates fixed t). For $t = 1/L$, the exact-line-search iterate satisfies

$$f(x^{k+1}) = \min_{\alpha \geq 0} f\big(x^k - \alpha \nabla f(x^k)\big) \leq f\Big(x^k - \tfrac{1}{L}\nabla f(x^k)\Big),$$

so it never does worse (and is typically better) in function decrease.

Remark 12.2 (Local vs. global). Rates like Theorem **??** are global under PL. For general nonconvex f, PL may only hold in a neighborhood of a minimum (a local linear rate), even when iterates globally decrease.

13 Worked examples

Example 13.1 (Least squares step sizes)**.** Let $f(x) = \frac{1}{2}\|Ax - b\|^2$. Then $L = \lambda_{\max}(A^TA)$. If A has full column rank, $\mu = \lambda_{\min}(A^TA)$, so GD with $t = 1/L$ has linear rate $(1 - \mu/L)^k = (1 - 1/\kappa)^k$. (Compute L and μ from the spectrum of A^TA.)

Example 13.2 (Logistic regression step sizes)**.** For $f(x) = \sum_i \log(1 + \exp(b_i a_i^T x))$, $\nabla^2 f(x) = A^T D(x) A$ with $0 \preceq D(x) \preceq \frac{1}{4}I$, hence $L \leq \frac{1}{4}\lambda_{\max}(A^TA)$. On bounded domains with A full column rank, $\mu > 0$ exists, giving linear convergence with GD. (Empirically, backtracking picks steps near $1/L$ early on.)

Gotcha 13.3 (Units and step size). Gradients live in the dual space and carry units; $x^{k+1} = x^k - t\nabla f(x^k)$ implies t has units of (variable units)^2. Mismatched units make t hard to tune; standardize features.

14 Summary: what to remember

- L-smooth $\Rightarrow$ quadratic upper bound; μ-strongly convex $\Rightarrow$ quadratic lower bound.
- PL strictly generalizes strong convexity in the sense of convergence proofs; it applies beyond convex functions.
- Under L-smooth + PL, GD with $t = 1/L$ converges linearly with rate $(1 - \mu/L)^k$.
- For convex L-smooth f without PL, GD achieves $O(1/k)$ sublinear rate.
- Backtracking Armijo guarantees sufficient decrease and terminates; exact line search often accelerates.

Appendix A. The descent lemma (proof and variations)

Lemma 14.1 (Descent lemma)**.** *If f is L-smooth, then for all x, y,*

$$f(y) \le f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|y - x\|^2.$$

Proof. Define $\phi(t) = f\big(x + t(y - x)\big)$. Then $\phi'(t) = (y - x)^T \nabla f(x + t(y - x))$ and

$$\phi(1) - \phi(0) = \int_0^1 \phi'(t), dt = \int_0^1 \big[\nabla f(x) + (\nabla f(x + t(y - x)) - \nabla f(x))\big]^T (y - x), dt.$$

Apply Cauchy-Schwarz and Lipschitz continuity of ∇f to bound the second term by $\frac{L}{2}\|y - x\|^2$. $\square$

Corollaries. (i) For GD with $t \le 1/L$, $f(x^{k+1}) \le f(x^k) - \big(t - \frac{L}{2}t^2\big)\|\nabla f(x^k)\|^2$. (ii) With $t = 1/L$, $f(x^{k+1}) \le f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2$ (used in Theorem **??**).

Appendix B. Equivalent smoothness characterizations

Under twice differentiability, the following are equivalent:

$$L\text{-smooth}; \Longleftrightarrow; \nabla^2 f(x) \preceq LI \ \forall x \Longleftrightarrow \ \|\nabla f(y) - \nabla f(x)\| \le L\|y - x\| \ \forall x, y.$$

(See Definition **??**.)

Appendix C. Quadratics in detail

For $f(x) = \frac{1}{2}x^T A x - b^T x$ with $A \succ 0$:

$$L = \lambda_{\max}(A), \quad \mu = \lambda_{\min}(A), \quad x^{k+1} - x^\star = (I - tA)(x^k - x^\star).$$

The optimal fixed t minimizes $\max_i \|1 - t\lambda_i(A)\|$, attained at $t = \frac{2}{\lambda_{\max} + \lambda_{\min}}$, with rate $\rho = \frac{\kappa - 1}{\kappa + 1}$ in the A-norm; exact line search uses $t_k = \frac{\|\nabla f(x^k)\|^2}{\nabla f(x^k)^T A \nabla f(x^k)}$.

Appendix D. Backtracking always terminates

From Appendix A, $f(x - t\nabla f(x)) \le f(x) - t\|\nabla f(x)\|^2 + \frac{L}{2}t^2\|\nabla f(x)\|^2$. For $t \le \min 1, L^{-1}$, the Armijo condition with $c \le 1/2$ holds. Hence halving will eventually find an acceptable t. (This formalizes the slide's "A: yes!" remark.)

Appendix E. PL without convexity

Example 14.2 (A nonconvex PL function)**.** Let $f(x) = \frac{1}{2}\mathrm{dist}(x, \mathcal{M})^2$ where $\mathcal{M}$ is a closed subspace; PL holds with $\mu = 1$ though f is flat along $\mathcal{M}$ and not strongly convex. Under PL, GD still decreases linearly in objective to $f^\star = 0$, but x^k may converge only to the set $\mathcal{M}$ (not to a unique point). (Compare the slides' "river valley" comment.)

Appendix F. When GD diverges

For $f(x) = \frac{1}{2}Lx^2$ in 1D, the GD map is $x^{k+1} = (1 - tL)x^k$. If $t > 2/L$, then $\|1 - tL\| > 1$ and iterates diverge even though f is convex and smooth. This illustrates the tight stability range $t \in (0, 2/L)$ for quadratics.

Appendix G. Units, scaling, and step-size choice

Gradients inhabit the dual space: if x has units "meters," ∇f can have units "1/meters," so t carries "meters^2." Poor scaling across coordinates makes a single global t awkward; standardizing features and rescaling variables can make L and μ more benign and GD more stable.