CMPUT 654: Mathematical Foundations of Modern AI Systems Fall 2026

Lecture 11: Global convergence from local gradients

October 8, 2026

Lecturer: Csaba Szepesvári *Note prepared by: Csaba Szepesvári*

Reading guide. Lectures 8–10 separated three questions about learning by gradient methods. Lecture 8 asked whether the available gradients contain information about the target. Lecture 9 studied what exact empirical fitting can imply for prediction. Lecture 10 asked which solution an optimization method selects when many parameters fit the training data. Even the max-margin analysis in Lecture 10 first needed to know that logistic loss tends to zero. That fact was a supporting step there; it is the central question here. We ask when an update computed from the gradient at the current parameter drives the training loss toward its best possible value.

The principal example is the same unregularized softmax log loss used for language-model outputs in Lectures 1 and 3 and for separable logistic classification in Lecture 10. Depending on the data, it admits a general convex comparison bound, a linear rate when the relevant parameters remain bounded and all fitted class probabilities stay uniformly away from zero, or an $O(1/t)$ rate under strict separation even though no finite minimizer exists. We then analyze stochastic gradient descent and define a Lyapunov function, which extends the proof pattern to root-finding and fixed-point updates that need not be gradients of an objective. Material marked with * may be skipped on a first reading. It develops nonlinear gradient dominance, nonsmooth convex optimization, nonlinear least squares, overparameterized ReLU networks, softmax policy gradient, and temporal-difference learning. The prerequisites are multivariable calculus, eigenvalues of symmetric matrices, Hoeffding's inequality, and conditional expectation.

11.1 The local-to-global question

Several kinds of convergence can be relevant in optimization. They should not be treated as interchangeable.

(a) *Objective-value convergence* means that the loss values approach the best possible loss.

(b) *Parameter convergence* means that the iterates themselves approach a finite vector, $\theta_t \to \theta^\star$. This is stronger and can fail when there are many minimizers or when the best loss is attained only in the limit as the parameter norm diverges.

(c) When the norm diverges, *directional convergence* may still hold: $\theta_t/\|\theta_t\|_2$ can approach a limiting direction. Lecture 10 used this conclusion to describe max-margin selection in separable logistic regression.

Objective-value convergence is the minimum optimization conclusion. Lecture 10 needed it before asking which direction gradient descent selected. This lecture concentrates on that conclusion and remarks on parameter or directional convergence when the example permits one. Its guiding question is

> When does local first-order information certify convergence of the objective value to its global infimum?

Let $L : \mathbb{R}^d \to \mathbb{R}$ be differentiable and bounded below, and define

$$L^\star = \inf_{\theta \in \mathbb{R}^d} L(\theta). \tag{11.1}$$

We do not assume that this infimum is attained: there may be no finite minimizer $\theta^\star$ satisfying $L(\theta^\star) = L^\star$. The main target throughout the lecture is objective-value convergence,

$$L(\theta_t) \longrightarrow L^\star.$$

Equivalently, the nonnegative *objective gap* $L(\theta_t) - L^\star$ tends to zero. For separable softmax models, this happens while the parameter norm diverges, so objective-value convergence cannot be replaced by convergence to a minimizer.

Gradient descent uses the update

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t), \tag{11.2}$$

where $\eta > 0$ is the stepsize. Differentiability gives, for a perturbation r,

$$L(\theta_t + r) = L(\theta_t) + \langle \nabla L(\theta_t), r \rangle + o(\|r\|_2).$$

Thus the gradient controls infinitesimal changes around θ_t; it does not directly compare $L(\theta_t)$ with values at distant parameters or with $L^\star$. To control the finite step in Eq. (11.2), we begin with smoothness.

Definition 11.1.1 (Smooth objective)**.** The function L is β-smooth if

$$\|\nabla L(u) - \nabla L(v)\|_2 \le \beta \|u - v\|_2 \qquad \text{for every } u, v \in \mathbb{R}^d.$$

Smoothness says that the gradient cannot change faster than linearly with the parameter. It does not imply convexity: a smooth objective may still have several stationary points, including suboptimal ones. Its role here is to control the error in the first-order approximation over a finite step. The following inequality makes that control explicit.

Lemma 11.1.2 (Descent lemma)**.** *If L is β-smooth, then*

$$L(v) \le L(u) + \langle \nabla L(u), v - u \rangle + \frac{\beta}{2} \|v - u\|_2^2 \tag{11.3}$$

for every $u, v \in \mathbb{R}^d$.

Proof. Put $r = v - u$ and $g(s) = L(u + sr)$ for $s \in [0, 1]$. The fundamental theorem of calculus and smoothness give

$$\begin{aligned} L(v) - L(u) - \langle \nabla L(u), r \rangle &= \int_0^1 \langle \nabla L(u + sr) - \nabla L(u), r \rangle \, ds \\ &\le \int_0^1 \beta s \|r\|_2^2 \, ds = \frac{\beta}{2} \|r\|_2^2. \end{aligned}$$

□

For twice continuously differentiable objectives, Hessian inequalities characterize the standard first- and second-order conditions.

Proposition 11.1.3 (Hessian characterizations of smoothness and convexity)**.** *Let $L : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable.*

(a) The quadratic upper bound Eq. (11.3) holds for every $u, v \in \mathbb{R}^d$ if and only if

$$\nabla^2 L(\theta) \preceq \beta I \qquad \text{for every } \theta \in \mathbb{R}^d.$$

(b) The function L is convex if and only if $\nabla^2 L(\theta) \succeq 0$ for every θ. The function L is μ-strongly convex if and only if $\nabla^2 L(\theta) \succeq \mu I$ for every θ.

(c) The function L is β-smooth if and only if

$$-\beta I \preceq \nabla^2 L(\theta) \preceq \beta I \qquad \text{for every } \theta \in \mathbb{R}^d.$$

Consequently, for a twice continuously differentiable convex objective, the quadratic upper bound and β-smoothness are equivalent; both are equivalent to

$$0 \preceq \nabla^2 L(\theta) \preceq \beta I \qquad \text{for every } \theta \in \mathbb{R}^d.$$

The converse from the quadratic upper bound to smoothness can fail without convexity. A nonconvex objective can have arbitrarily large negative curvature while remaining below every tangent plane. Exercise 1 asks you to prove all parts of Proposition 11.1.3 and verify a one-dimensional counterexample when convexity is removed.

We now use the descent lemma to analyze gradient descent. Substituting the gradient step into the descent lemma gives a decrease that depends only on the gradient at the current point.

A point θ is *stationary* if $\nabla L(\theta) = 0$ and *ϵ-stationary* if $\|\nabla L(\theta)\|_2 \le \epsilon$.

Proposition 11.1.4 (Smoothness gives descent and stationarity)**.** *Suppose that L is β-smooth and that the iterates follow Eq. (11.2) with $0 < \eta \le 1/\beta$. Then*

$$L(\theta_{t+1}) \le L(\theta_t) - \frac{\eta}{2}\|\nabla L(\theta_t)\|_2^2. \tag{11.4}$$

Consequently, for every integer $T \ge 1$,

$$\min_{0 \le t < T} \|\nabla L(\theta_t)\|_2^2 \le \frac{2\big(L(\theta_0) - L^\star\big)}{\eta T}. \tag{11.5}$$

Proof. Use $v - u = -\eta \nabla L(u)$ in Eq. (11.3):

$$L(\theta_{t+1}) \le L(\theta_t) - \eta\left(1 - \frac{\beta\eta}{2}\right)\|\nabla L(\theta_t)\|_2^2 \le L(\theta_t) - \frac{\eta}{2}\|\nabla L(\theta_t)\|_2^2.$$

Summing from $t = 0$ to $T - 1$ and using $L(\theta_T) \ge L^\star$ gives the complete bound

$$\frac{\eta T}{2} \min_{0 \le t < T} \|\nabla L(\theta_t)\|_2^2 \le \frac{\eta}{2} \sum_{t=0}^{T-1} \|\nabla L(\theta_t)\|_2^2 \le L(\theta_0) - L^\star,$$

which is Eq. (11.5). □

Equation (11.5) is a global statement about reaching an approximately stationary point. The minimum means that at least one of the first T iterates has a small gradient; the result does not claim this for the last iterate or for every iterate. More importantly, a small gradient need not mean a small objective gap.

Example 11.1.5 (A stationary point can be globally suboptimal)**.** Let

$$L(\theta) = 1 - \cos\theta.$$

This function is 1-smooth and has $L^\star = 0$. If $\theta_0 = \pi$, then $\nabla L(\theta_0) = \sin\pi = 0$, so gradient descent never moves. Its objective value remains 2. Thus smoothness and exact local descent do not connect stationarity to global optimality.

Exercise 2 asks you to verify every part of this failure and to identify exactly why the stationarity bound does not imply global objective convergence.

When smoothness fails. The smoothness assumption concerns the objective, not the rule used to choose an update direction. Lecture 10's signed gradient descent applies a discontinuous transformation to the gradient of a smooth objective. A ReLU network creates a different issue: its training objective can itself fail to be differentiable.

The function $L(x) = |x|$ gives the basic stepsize obstruction. Away from zero, use the available derivative in

$$x_{t+1} = x_t - \eta\,\mathrm{sign}(x_t).$$

If $0 < x_0 < \eta$, then $x_1 = x_0 - \eta < 0$ and $x_2 = x_0$. Thus a fixed stepsize can make the iterates alternate across the minimizer rather than converge to it. The smooth descent lemma is unavailable, and a locally decreasing direction does not by itself give a finite-step decrease bound.

For a convex nonsmooth objective, subgradient descent replaces the objective-descent calculation by a squared-distance calculation and obtains an $O(1/\sqrt{T})$ comparator bound for an averaged iterate; this is developed in Section 11.7.2. General ReLU training is also nonconvex, so that convex result does not apply. For example,

$$L(w) = \frac{1}{2}\big([w]_+ - 1\big)^2, \qquad [w]_+ = \max\{0, w\},$$

has global minimum zero at $w = 1$, but $L(w) = 1/2$ and $L'(w) = 0$ for every $w < 0$. A unit initialized there is *dead*: its output and parameter derivative both vanish. Section 11.7.4 gives an exact example in which width makes complete death unlikely and then relates the mechanism to the Jacobian condition used in this lecture.

11.2 From small gradients to small objective gaps

The previous section showed that gradient descent reaches iterates with small gradient norm. Our actual goal is to control the objective gap $L(\theta_t) - L^\star$. We therefore need a relation between these two quantities.

The most direct route is to impose exactly the comparison needed to complete the argument: for some $\mu > 0$, require

$$\frac{1}{2}\|\nabla L(\theta)\|_2^2 \geq \mu\big(L(\theta) - L^\star\big) \qquad \text{for every } \theta \in \mathbb{R}^d. \tag{11.6}$$

This condition is the *μ-Polyak–Łojasiewicz inequality*, abbreviated the *μ-PL inequality*. It says that a small gradient is possible only at a small objective gap, and it rules out globally suboptimal stationary points.

Combining the PL inequality with smooth descent gives the main reusable theorem.

Theorem 11.2.1 (Linear convergence under PL)**.** *Suppose that L is β-smooth and satisfies the μ-PL inequality. For gradient descent with $0 < \eta \le 1/\beta$,*

$$L(\theta_t) - L^\star \le (1 - \eta\mu)^t \big(L(\theta_0) - L^\star\big). \tag{11.7}$$

Proof. By Eqs. (11.4) and (11.6),

$$\begin{aligned} L(\theta_{t+1}) - L^\star &\le L(\theta_t) - L^\star - \frac{\eta}{2}\|\nabla L(\theta_t)\|_2^2 \\ &\le (1 - \eta\mu)\big(L(\theta_t) - L^\star\big). \end{aligned}$$

For a nonconstant β-smooth objective that is bounded below, apply the descent lemma to $v = u - \nabla L(u)/\beta$ to obtain

$$L^\star \le L(v) \le L(u) - \frac{1}{2\beta}\|\nabla L(u)\|_2^2.$$

Combining this upper bound on the gradient norm with PL gives $\mu \le \beta$. Hence $1 - \eta\mu \in [0, 1)$, and iterating the preceding recursion proves the claim. □

The term *linear convergence* refers to the geometric bound $q^t(L(\theta_0) - L^\star)$ with $q = 1 - \eta\mu < 1$. It does not mean that the error is a linear function of the iteration number.

The theorem separates two proof obligations. Smoothness proves that the local step decreases the objective. PL proves that the available decrease is large relative to the remaining objective gap. A convergence proof that checks only the first obligation reaches Eq. (11.5), not Eq. (11.7).

The PL inequality is linear in the squared gradient norm and the objective gap. Other powers give other rates. The exact calculation is developed in Section 11.7.1; the separated softmax model below will use the quadratic relation $\|\nabla L(\theta)\|_2^2 \gtrsim (L(\theta) - L^\star)^2$.

The definition leaves two questions. When can we establish a PL inequality? What can replace it when no uniform PL constant exists?

11.2.1 A sufficient condition for PL

At a point u, the affine function

$$v \longmapsto L(u) + \langle \nabla L(u), v - u \rangle$$

is the tangent-plane approximation to L.

Definition 11.2.2 (Convexity and strong convexity)**.** A differentiable function L is convex if

$$L(v) \ge L(u) + \langle \nabla L(u), v - u \rangle \tag{11.8}$$

for every $u, v \in \mathbb{R}^d$. It is μ-strongly convex, for $\mu > 0$, if

$$L(v) \ge L(u) + \langle \nabla L(u), v - u \rangle + \frac{\mu}{2}\|v - u\|_2^2 \tag{11.9}$$

for every $u, v \in \mathbb{R}^d$.

Strong convexity supplies one way to verify the PL inequality.

Proposition 11.2.3 (Strong convexity implies PL)**.** *If L is differentiable and μ-strongly convex, then it satisfies the μ-PL inequality.*

Proof. Fix u and compare it with an arbitrary v. Rearranging Eq. (11.9) gives

$$L(u) - L(v) \le \langle \nabla L(u), u - v \rangle - \frac{\mu}{2}\|u - v\|_2^2.$$

Thus, the excess loss at u relative to the loss at v is upper bounded by a function that combines the gradient at u and $u - v$. We would be essentially done if this upper bound were further upper bounded by just $\|\nabla L(u)\|_2$, scaled as needed. This is promising: the first term, which scales linearly with $\nabla L(u)$, is linear in $u - v$, while the second is quadratic with a fixed negative coefficient. For every $a \in \mathbb{R}$,

$$\sup_{r \in \mathbb{R}} \left\{ ar - \frac{\mu}{2} r^2 \right\} = \frac{a^2}{2\mu}.$$

Thus, by the Cauchy–Schwarz inequality,

$$\begin{aligned} L(u) - L(v) &\le \|\nabla L(u)\|_2 \|u - v\|_2 - \frac{\mu}{2}\|u - v\|_2^2 \\ &\le \sup_{r \in \mathbb{R}} \left\{ \|\nabla L(u)\|_2 r - \frac{\mu}{2} r^2 \right\} \\ &= \frac{\|\nabla L(u)\|_2^2}{2\mu}. \end{aligned}$$

Taking the supremum over v and using $L(u) - L^\star = \sup_v \{L(u) - L(v)\}$ gives Eq. (11.6). □

The converse is false. A PL objective need not be convex. The following example gives this distinction an elementary machine-learning interpretation.

Example 11.2.4 (A nonconvex residual model with a PL loss)**.** Consider regression on one training example with target zero. Suppress the fixed input and let a one-parameter residual model output

$$g(\theta) = \theta + 10 \tanh \theta, \qquad L(\theta) = \frac{1}{2} g(\theta)^2.$$

The additive linear term prevents the parameter-to-residual derivative from vanishing:

$$g'(\theta) = 1 + 10 \operatorname{sech}^2 \theta \ge 1.$$

Since g is strictly increasing and $g(0) = 0$, we have $L^\star = 0$ and

$$\frac{1}{2}|L'(\theta)|^2 = \frac{1}{2} g(\theta)^2 g'(\theta)^2 \ge L(\theta).$$

Thus L is 1-PL. Its second derivative is bounded because g' is bounded and $\theta \operatorname{sech}^2 \theta$ is bounded, so L is globally smooth. It is not convex. At $\theta = \log 3$, where $\tanh \theta = 4/5$ and $\operatorname{sech}^2 \theta = 9/25$,

$$L''(\log 3) = -\frac{623 + 144 \log 3}{25} < 0.$$

Thus a nonconvex parameterization can still have a gradient that certifies the global residual. The Jacobian condition in Section 11.7.3 gives the corresponding statement for a vector of training residuals.

Exercise 3 asks you to derive the PL certificate from strong convexity and then obtain the linear convergence recursion. The example above shows why PL, rather than strong convexity, is the property used by that recursion.

11.2.2 Convexity without PL

PL is not necessary for objective-value convergence. Rearranging the convexity inequality gives, for every $u \in \mathbb{R}^d$,

$$L(\theta) - L(u) \le \langle \nabla L(\theta), \theta - u \rangle \le \|\nabla L(\theta)\|_2 \|\theta - u\|_2. \tag{11.10}$$

The second inequality is Cauchy–Schwarz. Thus a small gradient gives a small loss difference when the distance to a near-optimal reference parameter is controlled; a bounded parameter set supplies such a bound. On an unbounded parameter space, this distance need not be bounded, and the infimum may not be attained. We call the arbitrary reference parameter u a *comparator*; it need not be a minimizer.

Gradient descent makes stronger use of the first inequality in Eq. (11.10). Tracking squared distance to a fixed comparator makes the distance terms telescope.

Theorem 11.2.5 (Gradient descent against every comparator)**.** *Let $L : \mathbb{R}^d \to \mathbb{R}$ be convex and β-smooth. Let gradient descent use $0 < \eta \le 1/\beta$. Then for every $u \in \mathbb{R}^d$ and every integer $T \ge 1$,*

$$L(\theta_T) - L(u) \le \frac{\|\theta_0 - u\|_2^2}{2\eta T}. \tag{11.11}$$

Proof. Convexity and the distance recursion give

$$\begin{aligned} \|\theta_{t+1} - u\|_2^2 &= \|\theta_t - u\|_2^2 - 2\eta \langle \nabla L(\theta_t), \theta_t - u \rangle + \eta^2 \|\nabla L(\theta_t)\|_2^2 \\ &\le \|\theta_t - u\|_2^2 - 2\eta \big(L(\theta_t) - L(u)\big) + \eta^2 \|\nabla L(\theta_t)\|_2^2. \end{aligned}$$

By Eq. (11.4),

$$\eta^2 \|\nabla L(\theta_t)\|_2^2 \le 2\eta \big(L(\theta_t) - L(\theta_{t+1})\big).$$

Combining the two displays gives

$$2\eta \big(L(\theta_{t+1}) - L(u)\big) \le \|\theta_t - u\|_2^2 - \|\theta_{t+1} - u\|_2^2.$$

Summing from $t = 0$ to $T - 1$ makes the squared distances telescope. Since the objective values are nonincreasing,

$$2\eta T \big(L(\theta_T) - L(u)\big) \le \|\theta_0 - u\|_2^2 - \|\theta_T - u\|_2^2 \le \|\theta_0 - u\|_2^2,$$

which proves the claim. □

If a finite minimizer $\theta^\star$ exists, choosing $u = \theta^\star$ gives the usual $O(1/T)$ rate. A minimizer is not required for convergence. For every $\epsilon > 0$, choose a finite u_ϵ satisfying $L(u_\epsilon) \le L^\star + \epsilon/2$. The theorem gives

$$0 \le L(\theta_T) - L^\star \le \frac{\epsilon}{2} + \frac{\|\theta_0 - u_\epsilon\|_2^2}{2\eta T}.$$

For all sufficiently large T, the second term is at most $\epsilon/2$. Therefore $L(\theta_T) \to L^\star$. Without a bound relating $\|\theta_0 - u_\epsilon\|_2$ to ϵ, this argument gives convergence but no explicit rate to the unattained infimum.

Exercise 5 asks you to reconstruct the telescoping argument and prove the last conclusion using an approximate minimizer.

The distinction between stationarity and objective convergence is also visible in iteration complexity. With the largest standard stepsize $\eta = 1/\beta$, smoothness alone guarantees an iterate with $\|\nabla L(\theta_t)\|_2 \le \epsilon$ among the first T iterates whenever

$$T \ge \frac{2\beta(L(\theta_0) - L^\star)}{\epsilon^2} \tag{11.12}$$

steps. Under PL, the inequality $(1-u)^t \le e^{-ut}$ gives

$$t \ge \frac{\beta}{\mu} \log \frac{L(\theta_0) - L^\star}{\epsilon} \quad \Longrightarrow \quad L(\theta_t) - L^\star \le \epsilon. \tag{11.13}$$

The conclusions differ as much as the rates: the first controls a local quantity, whereas the second controls the objective gap. The ratio $\kappa = \beta/\mu$ is the *condition number* and appears directly in Eq. (11.13).

11.3 The update need not equal the negative gradient

Exact gradient descent is one way to obtain local decrease, but it is not the only update used in the course. A *preconditioned* update multiplies the gradient by a matrix, thereby rescaling or mixing its coordinates. Numerical or sampling error can perturb a computed gradient, and Lecture 10 used signed gradient descent. To determine which parts of the PL proof survive, we separate two requirements: the proposed direction must decrease the first-order approximation, and its length must be controlled so that the quadratic smoothness term cannot overwhelm that decrease.

Theorem 11.3.1 (Aligned updates)**.** *Let L be β-smooth, and consider*

$$\theta_{t+1} = \theta_t + \eta d_t. \tag{11.14}$$

Suppose that constants $c, C > 0$ satisfy

$$-\langle \nabla L(\theta_t), d_t \rangle \ge c \|\nabla L(\theta_t)\|_2^2, \qquad \|d_t\|_2 \le C \|\nabla L(\theta_t)\|_2 \tag{11.15}$$

for every t. If $0 < \eta \le c/(\beta C^2)$, then

$$L(\theta_{t+1}) \le L(\theta_t) - \frac{\eta c}{2} \|\nabla L(\theta_t)\|_2^2. \tag{11.16}$$

If L also satisfies the μ-PL inequality, then

$$L(\theta_t) - L^\star \le (1 - \eta c \mu)^t \big(L(\theta_0) - L^\star\big). \tag{11.17}$$

Proof. The descent lemma and Eq. (11.15) give

$$\begin{aligned} L(\theta_{t+1}) &\le L(\theta_t) + \eta \langle \nabla L(\theta_t), d_t \rangle + \frac{\beta \eta^2}{2} \|d_t\|_2^2 \\ &\le L(\theta_t) - \left(\eta c - \frac{\beta \eta^2 C^2}{2} \right) \|\nabla L(\theta_t)\|_2^2 \\ &\le L(\theta_t) - \frac{\eta c}{2} \|\nabla L(\theta_t)\|_2^2. \end{aligned}$$

Under PL, the last term is at least $\eta c \mu (L(\theta_t) - L^\star)$. Cauchy–Schwarz gives $c \le C$; together with $\mu \le \beta$ and the stepsize bound, this implies $0 \le 1 - \eta c \mu < 1$. Iteration gives Eq. (11.17). □

For example, if $d_t = -P_t \nabla L(\theta_t)$ and the symmetric matrices P_t satisfy

$$mI \preceq P_t \preceq MI,$$

then Eq. (11.15) holds with $c = m$ and $C = M$. If $d_t = -\nabla L(\theta_t) + e_t$ with $\|e_t\|_2 \le \rho \|\nabla L(\theta_t)\|_2$ for some $\rho < 1$, it holds with $c = 1 - \rho$ and $C = 1 + \rho$.

Exercise 6 asks you to prove the theorem and specialize it to a uniformly positive-definite preconditioner.

The conditions are sufficient, not necessary. For the signed-gradient direction from Lecture 10,

$$d_t = -\operatorname{sign}(\nabla L(\theta_t)), \qquad -\langle \nabla L(\theta_t), d_t \rangle = \|\nabla L(\theta_t)\|_1.$$

The displayed inner product is positive whenever the gradient is nonzero, so the signed-gradient direction decreases the first-order approximation. Its Euclidean norm, however, is the square root of the number of nonzero gradient coordinates and need not approach zero with the gradient norm. It therefore fails the second condition in Eq. (11.15) near a zero gradient. The signed-gradient result in Lecture 10 instead uses a decreasing stepsize and the norm-specific steepest-descent calculation. The present theorem does not subsume that analysis.

A positive inner product with the negative gradient is not sufficient by itself for a fixed-step result. For $L(\theta) = \theta^2/2$, choose $d_t = -M\theta_t$ with $M > 0$. Then $-L'(\theta_t)d_t = M\theta_t^2 > 0$ whenever $\theta_t \neq 0$, but

$$\theta_{t+1} = (1 - \eta M)\theta_t.$$

The objective increases whenever $|1 - \eta M| > 1$. The norm condition in Eq. (11.15) controls the second-order cost of the proposed step. A *line search* instead tests candidate stepsizes against the objective and accepts one that gives sufficient decrease. Such a procedure, or a different Lyapunov argument, can replace the norm condition. The inner-product condition alone cannot.

11.3.1 What the certificates say about earlier examples

The results above recover part of the optimization story from the previous lectures, while keeping it separate from the statistical and solution-selection questions.

Earlier example	Loss-control condition	What this lecture adds
Interpolating least squares	PL on the data span, with constant $\lambda_{\min}(\mathbf{X}\mathbf{X}^\top)$	Gradient descent converges linearly in training loss. Its unchanged null-space component still determines which interpolant is selected, as in Lecture 10.
Nonseparable logistic regression	Convex comparison always applies; overlap can give PL after unchanged parameter directions are removed	The loss reaches its finite optimum, with a linear rate under the overlap condition developed below.
Separable logistic regression	A non-PL certificate has the form $\|\nabla L_n(w)\|_2 \geq a L_n(w)$	The loss decreases to zero at rate $O(1/t)$, while Lecture 10 describes the max-margin direction selected as the parameter norm diverges.
Signed gradient descent	Its inner product with the gradient is negative, but its length is not proportional to the gradient norm	The aligned-update theorem does not apply; the norm-specific argument from Lecture 10 is needed.

For the least-squares row, $L_n(w) = \|\mathbf{X}w - y\|_2^2/2$ and the displayed constant assumes the full-row-rank interpolation setting of Lecture 10. Parameter convergence and objective convergence answer different questions: PL controls the residual, whereas initialization controls the component in $\ker(\mathbf{X})$.

11.4 Unregularized softmax log loss

This example joins several parts of the course. Lectures 1 and 3 introduced softmax probabilities and token log loss. Lecture 10 studied the binary, separable case and asked which classifier direction gradient descent selects. We now ask the logically prior optimization question: does the loss itself approach its infimum, at what rate, and how does the answer change between overlap and separation?

In this section, p denotes the feature dimension; the generic optimization results above used d for the dimension of a vector-valued parameter. Fix feature vectors $x_1, \dots, x_n \in \mathbb{R}^p$ and labels $y_1, \dots, y_n \in [K]$. Let $W \in \mathbb{R}^{K \times p}$ have row k equal to $w_k^\top$. Define

$$q_W(k \mid x_i) = \frac{\exp(w_k^\top x_i)}{\sum_{j=1}^K \exp(w_j^\top x_i)}, \tag{11.18}$$

$$L_n(W) = \frac{1}{n} \sum_{i=1}^n \left(\log \sum_{k=1}^K \exp(w_k^\top x_i) - w_{y_i}^\top x_i \right). \tag{11.19}$$

For a frozen transformer representation, this is the objective for training the output head: x_i is a realized hidden vector of the form $h_{t_i}^L$, W is the output matrix W_U, and K is the vocabulary size. "Frozen" means that the vectors x_i are treated as fixed while W is updated. A bias can be included by appending a constant coordinate to every x_i. For $K = 2$, only the difference between the two rows matters, and the model reduces to the binary logistic regression of Lecture 10.

11.4.1 Gradient, Hessian, and unchanged parameter directions

Let $e_k \in \mathbb{R}^K$ be the kth standard basis vector. Differentiating one example gives

$$\nabla_W L_n(W) = \frac{1}{n} \sum_{i=1}^n \big(q_W(\cdot \mid x_i) - e_{y_i}\big) x_i^\top, \tag{11.20}$$

or, row by row,

$$\nabla_{w_k} L_n(W) = \frac{1}{n} \sum_{i=1}^n \big(q_W(k \mid x_i) - \mathbb{I}\{y_i = k\}\big) x_i. \tag{11.21}$$

For matrices, write $\langle A, B \rangle_F = \sum_{k,j} A_{kj} B_{kj}$ and $\|A\|_F^2 = \langle A, A \rangle_F$; these are the *Frobenius inner product and norm*. To state the Hessian, fix W and i, and let Y_i^W be a random class with mass function

$$\mathbb{P}(Y_i^W = k) = q_W(k \mid x_i), \qquad k \in [K].$$

The Hessian quadratic form below is the second derivative of $s \mapsto L_n(W + s\Delta)$ at $s = 0$. It measures the local *curvature* of the loss in the direction Δ. For a direction $\Delta \in \mathbb{R}^{K \times p}$ whose kth row is $\delta w_k^\top$, a second differentiation gives

$$\big\langle \Delta, \nabla^2 L_n(W) \Delta \big\rangle_F = \frac{1}{n} \sum_{i=1}^n \mathrm{Var}\big((\delta w_{Y_i^W})^\top x_i\big) \geq 0. \tag{11.22}$$

The Hessian is positive semidefinite, so L_n is convex.

Define the empirical feature second-moment matrix

$$G = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top. \tag{11.23}$$

For any categorical probability vector and any $a \in \mathbb{R}^K$, the elementary range bound for variance gives

$$\mathrm{Var}(a_Y) \le \frac{1}{4}(\max_k a_k - \min_k a_k)^2 \le \frac{1}{2}\|a\|_2^2.$$

Applying this bound inside Eq. (11.22) yields

$$\begin{aligned}\left\langle \Delta, \nabla^2 L_n(W)\Delta \right\rangle_F &\le \frac{1}{2n}\sum_{i=1}^{n} \|\Delta x_i\|_2^2 \\ &\le \frac{1}{2}\lambda_{\max}(G)\|\Delta\|_F^2.\end{aligned}$$

Thus L_n is β-smooth with

$$\beta = \frac{1}{2}\lambda_{\max}(G). \tag{11.24}$$

Exercise 8 asks you to derive the gradient and the variance form of the Hessian and then obtain this smoothness constant.

The positive-semidefinite Hessian is not generally positive definite. The reason is that different parameter matrices can give exactly the same probabilities. Such a parameter transformation is an *invariance* of the model. For any $a \in \mathbb{R}^p$,

$$W \longmapsto W + \mathbf{1}a^\top \tag{11.25}$$

adds the same scalar $a^\top x_i$ to all logits and leaves the probabilities and loss unchanged. Feature directions orthogonal to $\mathrm{span}\{x_1, \dots, x_n\}$ are also invisible on the training examples. To choose one representative from each family of common shifts, we work in the zero-sum subspace

$$\mathcal{H} = \left\{ W \in \mathbb{R}^{K\times p} : \sum_{k=1}^{K} w_k = 0 \right\}. \tag{11.26}$$

Equation Eq. (11.21) shows that the gradient rows sum to zero. Gradient descent therefore preserves $\mathcal{H}$. We may also restrict each row to the span of the features.

Exercise 9 asks you to identify both invisible parameter directions and to prove that gradient descent preserves the zero-sum subspace.

Example 11.4.1 (Binary overlap and separation in one dimension)**.** Consider a single scalar logit difference s and $q_s(1) = e^s/(1+e^s)$. If a fraction $r \in [0,1]$ of the labels equal one, the average log loss is

$$L_r(s) = -r\log q_s(1) - (1-r)\log(1-q_s(1)). \tag{11.27}$$

Its first two derivatives are

$$L_r'(s) = q_s(1) - r, \qquad L_r''(s) = q_s(1)(1-q_s(1)). \tag{11.28}$$

If $0 < r < 1$, the unique minimizer is the finite log-odds $s^\star = \log(r/(1-r))$. A *sublevel set* at height c is the set of parameters whose loss is at most c. Since $L_r(s) \to \infty$ as $|s| \to \infty$, every nonempty

sublevel set is bounded. On each such set, $q_s(1)(1-q_s(1))$ has a positive minimum. This is the *overlap regime*: both labels occur with positive frequency at the same feature value.

If $r = 1$, then

$$L_1(s) = \log(1+e^{-s}), \qquad L_1'(s) = -\frac{1}{1+e^s}.$$

The loss approaches its infimum zero only as $s \to \infty$. Both the second derivative and the gradient vanish as the fitted probability approaches one. This is the separated regime. The calculation displays all three notions from the opening: the loss can converge to its infimum, the parameter need not converge to a finite value, and in higher-dimensional versions its normalized direction can still converge.

11.4.2 The conclusion available from convexity alone

The softmax objective is convex and β-smooth, so Theorem 11.2.5 applies after stacking the entries of W into one vector. For every $U \in \mathbb{R}^{K\times p}$,

$$L_n(W_T) - L_n(U) \le \frac{\|W_0 - U\|_F^2}{2\eta T}. \tag{11.29}$$

If a finite minimizer $W^\star$ exists, choosing $U = W^\star$ gives an $O(1/T)$ rate. If the infimum is not attained, the approximate-comparator argument following Theorem 11.2.5 gives

$$L_n(W_T) \longrightarrow L_n^\star. \tag{11.30}$$

The comparison theorem does not require a regularizer.

11.4.3 Overlap gives strong convexity after removing invariances

Here, *overlap* means that, after removing parameter directions invisible on the sample, the relevant empirical-loss sublevel set is bounded. In the binary example above, this holds when both labels occur with positive frequency. In general it rules out both complete and quasi-complete separation. We now state this condition precisely and show that it gives a linear rate.

Restrict the feature space so that $x_1, \dots, x_n$ span $\mathbb{R}^p$, and suppose

$$G \succeq \lambda_X I \qquad \text{for some } \lambda_X > 0. \tag{11.31}$$

The matrix inequality means $v^\top G v \ge \lambda_X \|v\|_2^2$ for every $v \in \mathbb{R}^p$: the observed features cover every remaining feature direction. Initialize at $W_0 \in \mathcal{H}$ and define the initial sublevel set

$$\mathcal{S}_0 = \{W \in \mathcal{H} : L_n(W) \le L_n(W_0)\}. \tag{11.32}$$

This is precisely the set of zero-sum parameters whose loss is no larger than the initial loss. Assume that $\mathcal{S}_0$ is bounded. After the invisible directions have been removed, this assumption rules out parameter sequences that can escape to infinity without increasing the loss.

The set $\mathcal{S}_0$ is closed and bounded in a finite-dimensional space, hence compact. Every softmax probability is positive and continuous, so

$$\alpha = \min_{W\in\mathcal{S}_0} \min_{i\in[n]} \min_{k\in[K]} q_W(k \mid x_i) > 0. \tag{11.33}$$

For a probability vector q with $q_k \geq \alpha$ and any $a \in \mathbb{R}^K$,

$$\mathrm{Var}_q(a_Y) = \min_{c\in\mathbb{R}} \sum_{k=1}^{K} q_k(a_k - c)^2 \geq \alpha \min_{c\in\mathbb{R}} \sum_{k=1}^{K} (a_k - c)^2. \tag{11.34}$$

If $\Delta \in \mathcal{H}$, then $\sum_k (\delta w_k)^\top x_i = 0$, so the last minimum in Eq. (11.34) is attained at $c = 0$. Therefore, for $W \in \mathcal{S}_0$,

$$\begin{aligned} \big\langle \Delta, \nabla^2 L_n(W)\Delta \big\rangle_F &\geq \frac{\alpha}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \big((\delta w_k)^\top x_i\big)^2 \\ &= \alpha \sum_{k=1}^{K} (\delta w_k)^\top G \delta w_k \\ &\geq \alpha\lambda_X \|\Delta\|_F^2. \end{aligned} \tag{11.35}$$

Convexity of L_n makes $\mathcal{S}_0$ convex. Continuity implies that L_n attains a global minimum on $\mathcal{S}_0$: points outside $\mathcal{S}_0$ have loss greater than $L_n(W_0)$, while the minimum inside is no greater than $L_n(W_0)$. Moreover, Eq. (11.35) gives the strong-convexity inequality along every line segment in $\mathcal{S}_0$. Gradient descent remains in this sublevel set by Eq. (11.4). We may therefore apply the PL proof on the sublevel set with $\mu = \alpha\lambda_X$.

Corollary 11.4.2 (Linear convergence under overlap)**.** *Under Eqs. (11.31) and (11.32) and the bounded-sublevel assumption above, gradient descent with $0 < \eta \leq 1/\beta$ satisfies*

$$L_n(W_t) - L_n^\star \leq (1 - \eta\alpha\lambda_X)^t \big(L_n(W_0) - L_n^\star\big). \tag{11.36}$$

The minimizer is unique after restricting to the zero-sum subspace and the span of the observed features.

Exercise 10 asks you to prove the variance lower bound that turns probability overlap and feature coverage into the Hessian lower bound used by this corollary.

The constant depends on both feature coverage and the smallest class probability encountered in the relevant sublevel set. If fitted probabilities approach the boundary of the probability simplex—the set of class-probability vectors—this uniform Hessian lower bound disappears. Adding $\lambda\|W\|_F^2/2$ would give a global strong-convexity constant λ, but it would also change the objective and the fitted probabilities.[1]

11.4.4 Separation gives a different gradient-to-gap relation

Suppose there is a multiclass separating direction. To remove the common shift, define the normalized margin

$$\gamma = \max_{\substack{U\in\mathcal{H} \\ \|U\|_F\leq 1}} \min_{i\in[n]} \min_{k\neq y_i} (u_{y_i} - u_k)^\top x_i, \tag{11.37}$$

and assume $\gamma > 0$. Compactness gives a maximizer $\overline{W}$. For every W, Eq. (11.20) and the margin definition imply

$$\begin{aligned} -\langle \nabla L_n(W), \overline{W} \rangle_F &= \frac{1}{n} \sum_{i=1}^{n} \sum_{k\neq y_i} q_W(k \mid x_i)(\overline{w}_{y_i} - \overline{w}_k)^\top x_i \\ &\geq \frac{\gamma}{n} \sum_{i=1}^{n} \big(1 - q_W(y_i \mid x_i)\big). \end{aligned} \tag{11.38}$$

Let

$$\ell_i(W) = -\log q_W(y_i \mid x_i), \qquad L_0 = L_n(W_0).$$

Gradient descent makes $L_n(W_t) \le L_0$, hence $0 < \ell_i(W_t) \le nL_0$. The function $u \mapsto 1 - e^{-u}$ is concave and vanishes at zero. Its chord on $[0, nL_0]$ gives

$$1 - q_{W_t}(y_i \mid x_i) = 1 - e^{-\ell_i(W_t)} \ge \rho_0 \ell_i(W_t), \qquad \rho_0 = \frac{1 - e^{-nL_0}}{nL_0}. \tag{11.39}$$

Combining Eq. (11.38) with $\|\overline{W}\|_F \le 1$ gives

$$\|\nabla L_n(W_t)\|_F \ge a L_n(W_t), \qquad a = \gamma\rho_0. \tag{11.40}$$

This is not a PL inequality: the right side is the loss rather than its square root. It nevertheless converts the local gradient into a global certificate.

Exercise 11 asks you to derive this certificate from the multiclass margin.

Theorem 11.4.3 (Loss decay on separable softmax data)**.** *Assume $\gamma > 0$, and let β be given by Eq. (11.24). If*

$$0 < \eta \le \min\left\{\frac{1}{\beta}, \frac{1}{a^2 L_0}\right\}, \tag{11.41}$$

then gradient descent satisfies

$$L_n(W_t) \le \frac{L_0}{1 + (\eta a^2 L_0/2)t}. \tag{11.42}$$

Moreover, $L_n^\star = 0$ is not attained at any finite W, and $\|W_t\|_F \to \infty$.

Proof. Equations Eqs. (11.4) and (11.40) give

$$L_n(W_{t+1}) \le L_n(W_t) - \frac{\eta a^2}{2} L_n(W_t)^2.$$

Put $c = \eta a^2/2$. The stepsize assumption and monotonicity give $cL_n(W_t) \le 1/2$. Since every finite softmax loss is positive,

$$\begin{aligned}\frac{1}{L_n(W_{t+1})} &\ge \frac{1}{L_n(W_t)(1 - cL_n(W_t))} \\ &\ge \frac{1}{L_n(W_t)} + c.\end{aligned}$$

Summing this inequality and taking reciprocals proves Eq. (11.42).

Along the separating ray $r\overline{W}$,

$$\begin{aligned}L_n(r\overline{W}) &= \frac{1}{n}\sum_{i=1}^n \log\left(1 + \sum_{k \ne y_i} e^{-r(\overline{w}_{y_i} - \overline{w}_k)^\top x_i}\right) \\ &\le (K-1)e^{-r\gamma} \longrightarrow 0.\end{aligned}$$

Thus $L_n^\star = 0$. At every finite W, every softmax probability is strictly below one and $L_n(W) > 0$, so the infimum is not attained. Finally, Eq. (11.42) gives $L_n(W_t) \to 0$. If $\|W_t\|_F$ did not tend to infinity, it would have a bounded subsequence and hence a convergent subsequence. Continuity would then give a finite parameter with zero loss, a contradiction. □

The reciprocal step is the only new recursion in this proof: Exercise 12 asks you to isolate and prove it.

There cannot be a uniform PL constant in this regime. Let $R = \max_i \|x_i\|_2$. For one example,

$$\|q_W(\cdot \mid x_i) - e_{y_i}\|_2 \le \sqrt{2}\big(1 - q_W(y_i \mid x_i)\big) \le \sqrt{2}\ell_i(W).$$

The triangle inequality in Eq. (11.20) therefore gives

$$\|\nabla L_n(W)\|_F \le \sqrt{2}RL_n(W). \tag{11.43}$$

Along the separating ray, $L_n(r\overline{W}) \to 0$, so

$$\frac{\frac{1}{2}\|\nabla L_n(r\overline{W})\|_F^2}{L_n(r\overline{W}) - L_n^\star} \le R^2 L_n(r\overline{W}) \longrightarrow 0.$$

No positive μ can satisfy a global PL inequality. The $O(1/t)$ rate is therefore not a weakened presentation of a hidden linear rate; it reflects a different gradient–loss relation.

Exercise 13 asks you to establish the gradient upper bound and use the separating ray to rule out a uniform PL constant.

The separation inequality is the case $\alpha = 2$ of the nonlinear gradient-dominance condition in Eq. (11.58). The PL inequality has $\alpha = 1$ and gives geometric decay; separated softmax has $\alpha = 2$ and gives reciprocal decay. The softmax policy-gradient result in Section 11.7.5 has an additional nonuniformity: its coefficient depends on the current policy rather than being a fixed global constant.

The direction selected by gradient descent is a separate question from loss decay. Under additional conditions, the normalized iterates approach a multiclass max-margin solution, as discussed in Lecture 10. The present theorem does not need that directional limit.

11.4.5 The three regimes

The same unregularized objective now has three useful conclusions.

Regime	Certificate	Conclusion for gradient descent
Arbitrary fixed features	Convex comparison with every finite U	$L_n(W_t) \to L_n^\star$; $O(1/t)$ if a finite minimizer is available as a fixed comparator
Bounded sublevel after removing invariances	Restricted strong convexity, hence PL	Linear convergence in objective value
Strict separation	$\|\nabla L_n(W)\|_F \ge aL_n(W)$ on the initial sublevel	$L_n(W_t) = O(1/t)$, while $\|W_t\|_F \to \infty$

Convex comparison also supplies a direct, though slightly weaker, finite-time bound in the separated regime. With $W_0 = 0$, $\eta = 1/\beta$, and $U_T = (\log T/\gamma)\overline{W}$, Eq. (11.29) gives

$$L_n(W_T) \le \frac{K-1}{T} + \frac{\beta(\log T)^2}{2\gamma^2 T}. \tag{11.44}$$

Here the comparator changes with the time horizon. The bound illustrates how convex comparison proves convergence to an unattained infimum.

Strict separation is not the only way a minimizer can fail to exist. In binary logistic regression, write $z_i = y_i x_i$ for labels $y_i \in \{-1, 1\}$. Consider three transformed examples

$$z_1 = (1, 0)^\top, \qquad z_2 = (0, 1)^\top, \qquad z_3 = (0, -1)^\top.$$

The direction $(1, 0)^\top$ has nonnegative margin on every example and a strictly positive margin on the first, but no direction has a strictly positive margin on all three. The loss is

$$L(w) = \frac{1}{3}\left[\log(1 + e^{-w_1}) + \log(1 + e^{-w_2}) + \log(1 + e^{w_2})\right]. \tag{11.45}$$

Sending $w_1 \to \infty$ and setting $w_2 = 0$ approaches the positive infimum $2 \log 2/3$, which no finite w attains. This is a *quasi-separated* case. The strict margin in Eq. (11.37) is zero, so Theorem 11.4.3 does not apply. The comparator theorem still proves $L(w_t) \to 2 \log 2/3$. This example explains why the comparator result is the general conclusion and the explicit $O(1/t)$ theorem is a sharper result for strict separation.

11.5 Stochastic gradient descent under PL

Full-batch gradients are rarely used to train a language model. *Stochastic gradient descent* (SGD) replaces the exact gradient by a random estimate; a minibatch average is the standard example. The deterministic decrease then acquires a variance term. Lecture 10 distinguished persistent noise from minibatch noise that vanishes at interpolation. The calculation below shows where that distinction enters a global convergence proof.

Let $\Theta_0, \Theta_1, \ldots$ be random iterates, and let $G_t \in \mathbb{R}^d$ be a random gradient estimate used in

$$\Theta_{t+1} = \Theta_t - \eta G_t. \tag{11.46}$$

Assume that, conditional on Θ_t,

$$\mathbb{E}[G_t \mid \Theta_t] = \nabla L(\Theta_t), \qquad \mathbb{E}\big[\|G_t - \nabla L(\Theta_t)\|_2^2 \mid \Theta_t\big] \le \frac{v}{B}. \tag{11.47}$$

The parameter B represents the minibatch size. These assumptions specify the randomness needed in the calculation. The first says that G_t is unbiased after the current iterate is fixed. The second bounds its conditional mean-squared error by v/B. No independence beyond these stated conditional properties is required.

Theorem 11.5.1 (Constant-stepsize SGD under PL)**.** *Suppose that L is β-smooth and μ-PL, and that Eqs. (11.46) and (11.47) hold. If $0 < \eta \le 1/\beta$, then*

$$\mathbb{E}[L(\Theta_{t+1}) - L^\star \mid \Theta_t] \le (1 - \eta\mu)\big(L(\Theta_t) - L^\star\big) + \frac{\beta\eta^2 v}{2B}, \tag{11.48}$$

$$\mathbb{E}[L(\Theta_t) - L^\star] \le (1 - \eta\mu)^t \mathbb{E}[L(\Theta_0) - L^\star] + \frac{\beta\eta v}{2\mu B}. \tag{11.49}$$

Proof. Conditional on Θ_t, smoothness gives

$$L(\Theta_{t+1}) \le L(\Theta_t) - \eta\langle \nabla L(\Theta_t), G_t\rangle + \frac{\beta\eta^2}{2}\|G_t\|_2^2.$$

Write $G_t = \nabla L(\Theta_t) + E_t$. Conditional unbiasedness gives $\mathbb{E}[E_t \mid \Theta_t] = 0$, so the cross term vanishes and

$$\mathbb{E}[\|G_t\|_2^2 \mid \Theta_t] = \|\nabla L(\Theta_t)\|_2^2 + \mathbb{E}[\|E_t\|_2^2 \mid \Theta_t] \le \|\nabla L(\Theta_t)\|_2^2 + \frac{v}{B}.$$

Taking conditional expectations, using $\eta \le 1/\beta$, and then applying PL proves Eq. (11.48). Now set

$$\delta_t = \mathbb{E}[L(\Theta_t) - L^\star], \qquad q = 1 - \eta\mu, \qquad b = \frac{\beta\eta^2 v}{2B}.$$

Taking expectations gives $\delta_{t+1} \le q\delta_t + b$. Repeated substitution yields

$$\delta_t \le q^t\delta_0 + b\sum_{j=0}^{t-1} q^j \le q^t\delta_0 + \frac{b}{1-q},$$

which is Eq. (11.49). □

Exercise 14 asks you to derive the conditional recursion and iterate it, keeping the contraction and variance terms separate.

The second term in Eq. (11.49) is often called the *noise floor* of this bound: after the geometric term becomes small, the bound no longer decreases below a quantity of order η/B. It is an upper bound, not a claim that every minibatch method has a nonzero limiting error. At an interpolating solution, per-example gradients can vanish and the conditional variance can decrease with the objective gap, as discussed in Lecture 10. A decreasing stepsize or an estimator designed so that its variance decreases over time gives other ways to make the additive term vanish.

The update direction need not be a stochastic gradient. Let

$$\Theta_{t+1} = \Theta_t + \eta D_t, \tag{11.50}$$

and suppose that, conditional on Θ_t, constants $c, C > 0$ and $\sigma^2 \ge 0$ satisfy

$$-\mathbb{E}[\langle \nabla L(\Theta_t), D_t\rangle \mid \Theta_t] \ge c\|\nabla L(\Theta_t)\|_2^2, \qquad \mathbb{E}[\|D_t\|_2^2 \mid \Theta_t] \le C^2\|\nabla L(\Theta_t)\|_2^2 + \frac{\sigma^2}{B}. \tag{11.51}$$

The first condition gives expected first-order decrease; the second controls both the part of the step that scales with the gradient and its remaining noise.

Proposition 11.5.2 (Stochastic aligned directions)**.** *Suppose that L is β-smooth and μ-PL and that Eqs. (11.50) and (11.51) hold. If*

$$0 < \eta \le \min\left\{\frac{c}{\beta C^2}, \frac{1}{c\mu}\right\},$$

then, with $\delta_t = \mathbb{E}[L(\Theta_t) - L^\star]$,

$$\mathbb{E}[L(\Theta_{t+1}) - L^\star \mid \Theta_t] \le (1 - \eta c\mu)(L(\Theta_t) - L^\star) + \frac{\beta\eta^2\sigma^2}{2B}, \tag{11.52}$$

$$\delta_t \le (1 - \eta c\mu)^t\delta_0 + \frac{\beta\eta\sigma^2}{2Bc\mu}. \tag{11.53}$$

For SGD, take $D_t = -G_t$, $c = C = 1$, and $\sigma^2 = v$.

Exercise 15 asks you to prove the proposition. It is the conditional-expectation counterpart of the aligned-update theorem in Section 11.3.

For a time-varying stepsize $\eta_t \le 1/\beta$, the same calculation gives

$$\delta_{t+1} \le (1 - \mu\eta_t)\delta_t + \frac{\beta v}{2B}\eta_t^2, \qquad \delta_t = \mathbb{E}[L(\Theta_t) - L^\star].$$

Proposition 11.5.3 (A decreasing-stepsize rate)**.** *Under the assumptions of Theorem 11.5.1, let*

$$\eta_t = \frac{2}{\mu(t+t_0)}, \qquad t_0 \geq \frac{2\beta}{\mu}.$$

Then

$$\delta_t \leq \frac{1}{t+t_0} \max\left\{ t_0\delta_0, \frac{2\beta v}{B\mu^2} \right\}. \tag{11.54}$$

In particular, the expected objective gap converges to zero at rate $O(1/t)$.

Here η_t is of order $1/t$: the cumulative steps are unbounded, while the squared steps are summable. This is a basic *stochastic approximation* schedule. Unlike the constant-stepsize bound, the resulting bound has no fixed noise floor. Exercise 16 asks you to prove Eq. (11.54) directly by induction.

11.6 The Lyapunov form of the argument

The preceding proofs use the objective gap to measure progress. The same argument can use another nonnegative function of the parameter. Consider a deterministic parameter update

$$\theta_{t+1} = \mathcal{T}(\theta_t), \qquad \theta_t \in \mathbb{R}^d. \tag{11.55}$$

For the rate arguments in this lecture, a *Lyapunov function* for a stated convergence target is a function

$$V : \mathbb{R}^d \longrightarrow [0, \infty)$$

whose infimum is zero and whose value does not increase under the update. The zero value need not be attained by a finite parameter. A useful convergence proof strengthens nonincrease to a quantitative decrease that forces $V(\theta_t) \to 0$. This is the minimal optimization use of the term needed here.[2]

For example, let $U : \mathbb{R}^d \to [0, \infty)$ measure the progress available at one parameter. If constants $c, \kappa > 0$ with $c\kappa \leq 1$ satisfy

$$V(\theta_{t+1}) \leq V(\theta_t) - cU(\theta_t), \qquad U(\theta_t) \geq \kappa V(\theta_t), \tag{11.56}$$

then the sequence $V_t = V(\theta_t)$ satisfies

$$V_t \leq (1 - c\kappa)^t V_0. \tag{11.57}$$

For gradient descent under PL,

$$V(\theta) = L(\theta) - L^\star, \qquad U(\theta) = \|\nabla L(\theta)\|_2^2.$$

Smoothness proves the first inequality in Eq. (11.56); PL proves the second. For separable softmax, $V(W) = L_n(W)$ is still a Lyapunov function even though its infimum is attained only as $\|W\|_F \to \infty$. The available progress instead satisfies $U(W) \geq a^2 V(W)^2$, which produces reciprocal decay. For a random parameter Θ_t, the corresponding statement is a *conditional drift inequality*: a bound on the expected one-step change after the current parameter is fixed. It controls $\mathbb{E}[V(\Theta_{t+1}) \mid \Theta_t]$, possibly with an additive noise term.

The aligned-update theorem is one way to prove a Lyapunov decrease. In that theorem, $V(\theta) = L(\theta) - L^\star$, and alignment with $-\nabla L(\theta)$ makes V decrease. The same proof applies to any

smooth V: if an update direction is aligned with $-\nabla V(\theta)$ and has controlled norm, smoothness converts that alignment into a decrease of V. A direct Lyapunov calculation can also establish decrease without first representing the update through the gradient of V.

This broader viewpoint matters for methods that do not begin with a loss. A *root-finding problem* asks for a parameter $\theta^\star$ satisfying $g(\theta^\star) = 0$ for a given map $g : \mathbb{R}^d \to \mathbb{R}^d$. One may design an update from the computable residual $g(\theta_t)$ and only afterward construct a Lyapunov function for its analysis. That function may depend on the unknown solution: for example, $V(\theta) = \|\theta - \theta^\star\|_2^2$ can be useful in a proof even though the algorithm cannot evaluate it. Calling V a surrogate loss would therefore be misleading here: it certifies the update but does not define it.

A *fixed point* of a map $\mathcal{T} : \mathbb{R}^d \to \mathbb{R}^d$ satisfies $\mathcal{T}(\theta^\star) = \theta^\star$. This is the root-finding problem $g(\theta) = \theta - \mathcal{T}(\theta) = 0$. The optional material gives precise root-finding and fixed-point results. It then treats temporal-difference learning, whose mean update is designed from a linear equation rather than from a training loss.

For a time-varying or random update, the same role is played by a per-step decrease inequality or a conditional drift inequality. The update map itself need not be the same at every iteration.

11.7 Optional reading*

11.7.1 Beyond PL: nonlinear gradient dominance

The PL inequality compares the squared gradient norm linearly with the objective gap. A more general comparison is

$$\frac{1}{2}\|\nabla L(\theta)\|_2^2 \geq \mu\big(L(\theta) - L^\star\big)^\alpha. \tag{11.58}$$

The exponent determines the rate obtained from smooth descent.

Proposition 11.7.1 (Rates from nonlinear gradient dominance)**.** *Let L be differentiable and bounded below, and suppose a gradient-descent sequence satisfies*

$$L(\theta_{t+1}) \leq L(\theta_t) - \frac{\eta}{2}\|\nabla L(\theta_t)\|_2^2 \qquad (t \geq 0)$$

for some $\eta > 0$. This descent bound holds, for example, when L is β-smooth and $0 < \eta \leq 1/\beta$. Suppose Eq. (11.58) holds with $\mu > 0$ and $\alpha \geq 1$ throughout the initial sublevel set

$$\mathcal{S}_0 = \{\theta : L(\theta) \leq L(\theta_0)\}.$$

Write $\Delta_t = L(\theta_t) - L^\star$. Then

$$\Delta_{t+1} \leq \Delta_t - \eta\mu\Delta_t^\alpha. \tag{11.59}$$

For $\alpha = 1$,

$$\Delta_t \leq (1 - \eta\mu)^t \Delta_0.$$

For $\alpha > 1$ and $\Delta_0 > 0$,

$$\Delta_t \leq \left(\Delta_0^{1-\alpha} + \eta\mu(\alpha - 1)t\right)^{-1/(\alpha-1)}. \tag{11.60}$$

If $\Delta_0 = 0$, every subsequent gap is zero.

Proof. Smooth descent keeps every iterate in $\mathcal{S}_0$. Combining Eqs. (11.4) and (11.58) gives Eq. (11.59). The case $\alpha = 1$ is the PL recursion.

Suppose $\alpha > 1$ and put $s = \alpha - 1$. If an iterate reaches zero gap, the claim follows. Otherwise, Eq. (11.59) gives

$$\Delta_{t+1} \leq \Delta_t(1 - \eta\mu\Delta_t^s).$$

If the factor is zero, then $\Delta_{t+1} = 0$. Otherwise it is positive. Using $(1-z)^{-s} \geq 1 + sz$ for $z \in [0, 1)$ gives

$$\Delta_{t+1}^{-s} \geq \Delta_t^{-s}(1 - \eta\mu\Delta_t^s)^{-s} \geq \Delta_t^{-s} + s\eta\mu.$$

Summing this inequality and raising both sides to the power $-1/s$ proves Eq. (11.60). □

Two scalar examples show how the exponent changes the conclusion. For $L(x) = x^4/4$,

$$\frac{1}{2}|L'(x)|^2 = 4L(x)^{3/2}.$$

If $x_0 \neq 0$ and $0 < \eta \leq 1/(3x_0^2)$, then $x_{t+1} = x_t(1 - \eta x_t^2)$ remains in $[-|x_0|, |x_0|]$, where L is $3x_0^2$-smooth. The proposition gives

$$L(x_t) \leq \Big(L(x_0)^{-1/2} + 2\eta t\Big)^{-2}.$$

For $L(x) = e^{-x}$,

$$\frac{1}{2}|L'(x)|^2 = \frac{1}{2}L(x)^2.$$

Its infimum zero is not attained. If $0 < \eta \leq 1/L(x_0)$, gradient descent remains in the initial sublevel set, where the objective is $L(x_0)$-smooth, and

$$L(x_t) \leq \Big(L(x_0)^{-1} + \frac{\eta}{2}t\Big)^{-1}.$$

Separated softmax has the same exponent $\alpha = 2$.

Exponents below one cannot hold near a smooth infimum along points with positive gaps tending to zero. Indeed, the descent lemma applied to $v = \theta - \nabla L(\theta)/\beta$ gives

$$\frac{1}{2}\|\nabla L(\theta)\|_2^2 \leq \beta\big(L(\theta) - L^\star\big).$$

Combining this upper bound with Eq. (11.58) would give $\mu\Delta^{\alpha-1} \leq \beta$. If $\alpha < 1$, the left side diverges as a positive gap Δ tends to zero.

Exercise 4 asks you to prove the polynomial-rate recursion, apply it to the quartic objective, and establish the obstruction for $\alpha < 1$.

11.7.2 Nonsmooth convex objectives

An objective is *nonsmooth* if it is not differentiable everywhere. Convexity still supplies a global first-order comparison, but the first-order object is now a set rather than a single gradient. For a convex function $L : \mathbb{R}^d \to \mathbb{R}$, a vector $s \in \mathbb{R}^d$ is a *subgradient* of L at θ if

$$L(u) \geq L(\theta) + \langle s, u - \theta\rangle \qquad \text{for every } u \in \mathbb{R}^d. \tag{11.61}$$

The set of all such vectors is the *subdifferential* $\partial L(\theta)$. At a differentiable point it contains only $\nabla L(\theta)$. For $L(x) = |x|$,

$$\partial L(x) = \begin{cases} \{1\}, & x > 0, \\ [-1, 1], & x = 0, \\ \{-1\}, & x < 0. \end{cases}$$

Subgradient descent chooses $s_t \in \partial L(\theta_t)$ and updates

$$\theta_{t+1} = \theta_t - \eta_t s_t. \tag{11.62}$$

The objective need not decrease at every step, as the fixed-stepsize behavior of $|x|$ already shows. Instead, the following result uses squared distance to an arbitrary comparator as a Lyapunov function.

Theorem 11.7.2 (Convex subgradient comparator bound)**.** *Let $L : \mathbb{R}^d \to \mathbb{R}$ be convex, let the iterates follow Eq. (11.62), and suppose $\|s_t\|_2 \le G$ for every t. For positive stepsizes, define*

$$S_T = \sum_{t=0}^{T-1} \eta_t, \qquad \bar{\theta}_T = \frac{1}{S_T} \sum_{t=0}^{T-1} \eta_t \theta_t.$$

Then every comparator $u \in \mathbb{R}^d$ satisfies

$$L(\bar{\theta}_T) - L(u) \le \frac{\|\theta_0 - u\|_2^2 + G^2 \sum_{t=0}^{T-1} \eta_t^2}{2S_T}. \tag{11.63}$$

In particular, if $\|\theta_0 - u\|_2 \le R$ and $R, G > 0$, the constant choice $\eta_t = R/(G\sqrt{T})$ gives

$$L(\bar{\theta}_T) - L(u) \le \frac{RG}{\sqrt{T}}. \tag{11.64}$$

Proof. Expanding one update and using the subgradient inequality gives

$$\begin{aligned} \|\theta_{t+1} - u\|_2^2 &= \|\theta_t - u\|_2^2 - 2\eta_t \langle s_t, \theta_t - u \rangle + \eta_t^2 \|s_t\|_2^2 \\ &\le \|\theta_t - u\|_2^2 - 2\eta_t \big(L(\theta_t) - L(u)\big) + \eta_t^2 G^2. \end{aligned}$$

Rearrange and sum from $t = 0$ to $T - 1$. The squared-distance terms telescope, and the last one is nonnegative, so

$$2 \sum_{t=0}^{T-1} \eta_t \big(L(\theta_t) - L(u)\big) \le \|\theta_0 - u\|_2^2 + G^2 \sum_{t=0}^{T-1} \eta_t^2.$$

Convexity and the definition of $\bar{\theta}_T$ imply

$$S_T L(\bar{\theta}_T) \le \sum_{t=0}^{T-1} \eta_t L(\theta_t).$$

Combining the last two displays proves Eq. (11.63). For the stated constant stepsize, its numerator is at most $2R^2$ and its denominator is $2R\sqrt{T}/G$, which proves Eq. (11.64). □

The theorem does not require the infimum to be attained because u is an arbitrary finite comparator. It controls an averaged iterate rather than every last iterate. With a fixed stepsize η, its right-hand side becomes

$$\frac{\|\theta_0 - u\|_2^2}{2\eta T} + \frac{\eta G^2}{2},$$

which displays the same fixed-stepsize residual seen for $|x|$. Decreasing stepsizes give convergence through this bound when $S_T \to \infty$ and $\sum_{t<T} \eta_t^2 / S_T \to 0$. Under the usual existence and bounded-moment conditions, stochastic subgradient arguments commonly use the stronger conditions $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$ so that progress continues while the cumulative squared-stepsize noise remains finite. For a strongly convex nonsmooth objective with bounded selected subgradients, stepsizes of order $1/t$ with suitable averaging give an $O(1/T)$ objective gap. Strong convexity alone does not recover the linear rate obtained from smoothness and PL.

11.7.3 A Jacobian condition for nonlinear least squares

Let f_θ be a nonlinear predictor and let $F(\theta) \in \mathbb{R}^r$ collect its residuals on r training constraints. For squared-loss training, after any desired normalization, the empirical objective has the form

$$L(\theta) = \frac{1}{2}\|F(\theta)\|_2^2.$$

Let $J_F(\theta) \in \mathbb{R}^{r\times d}$ be the *Jacobian*, with residual coordinates along rows and parameter coordinates along columns; explicitly, $[J_F(\theta)]_{j,k} = \partial F_j(\theta)/\partial\theta_k$. The chain rule gives

$$\nabla L(\theta) = J_F(\theta)^\top F(\theta).$$

Suppose $L^\star = 0$ and, on the relevant sublevel set,

$$J_F(\theta)J_F(\theta)^\top \succeq \sigma^2 I_r. \tag{11.65}$$

It follows that

$$\|\nabla L(\theta)\|_2^2 = F(\theta)^\top J_F(\theta)J_F(\theta)^\top F(\theta) \geq \sigma^2\|F(\theta)\|_2^2 = 2\sigma^2 L(\theta).$$

Thus L satisfies a σ^2-PL inequality on that sublevel set. The matrix inequality says that changing the parameters can affect every residual direction with strength at least σ. It can hold when $d > r$ even though the parameter itself is not uniquely determined, and it does not require L to be convex. The residual model in Example 11.2.4 is the scalar case. Interpolating linear least squares is the constant-Jacobian case: $F(w) = \mathbf{X}w - y$ and full row rank give $\mathbf{X}\mathbf{X}^\top \succeq \lambda_{\min}(\mathbf{X}\mathbf{X}^\top)I$. Thus the same calculation explains why overparameterized least squares can converge linearly in objective value even though Lecture 10 found many possible minimizing parameters.

11.7.4 Overparameterized ReLU networks

The one-unit example in Section 11.1 fails because its only gate is inactive. Width supplies several gates, so individual dead units need not make the whole network dead. The following scalar model isolates this effect before we consider several training examples.

For $w = (w_1, \dots, w_m) \in \mathbb{R}^m$, define

$$f_m(w) = \frac{1}{\sqrt{m}}\sum_{r=1}^m [w_r]_+, \qquad L_m(w) = \frac{1}{2}\big(f_m(w) - 1\big)^2. \tag{11.66}$$

Let $S_1, \dots, S_m$ be independent random signs satisfying $\mathbb{P}(S_r = 1) = \mathbb{P}(S_r = -1) = 1/2$. For a fixed $a \in (0, 1)$, initialize

$$W_{r,0} = \frac{aS_r}{\sqrt{m}}, \qquad r \in [m],$$

and run gradient descent with $0 < \eta \leq 1$. Put

$$Q = \sum_{r=1}^m \mathbb{I}\{S_r = 1\},$$

the number of initially active units.

Proposition 11.7.3 (Width makes complete ReLU death unlikely)**.** *On the event $\{Q = 0\}$, gradient descent never moves and $L_m(W_t) = 1/2$ for every t. This event has probability 2^{-m}. On the event $\{Q > 0\}$, the activation pattern remains fixed and*

$$1 - f_m(W_t) = \left(1 - \eta\frac{Q}{m}\right)^t \big(1 - f_m(W_0)\big), \tag{11.67}$$

$$L_m(W_t) = \left(1 - \eta\frac{Q}{m}\right)^{2t} L_m(W_0). \tag{11.68}$$

Moreover, for every $\delta \in (0, 1/2)$, with probability at least $1 - e^{-2m\delta^2}$,

$$L_m(W_t) \le (1 - \eta(1/2 - \delta))^{2t}\, L_m(W_0) \qquad \text{for every } t \ge 0. \tag{11.69}$$

Proof. The event $\{Q = 0\}$ is $\bigcap_{r=1}^m \{S_r = -1\}$. Independence gives probability 2^{-m}. Every coordinate is then negative, so every ReLU output and every parameter derivative is zero.

Now condition on $Q = q > 0$. Initially $f_m(W_0) = aq/m < 1$. If the initial activation pattern has persisted through time t and $f_m(W_t) < 1$, then an initially active coordinate satisfies

$$W_{r,t+1} = W_{r,t} + \frac{\eta}{\sqrt{m}}\big(1 - f_m(W_t)\big),$$

so it remains positive; an initially inactive coordinate has zero derivative and remains negative. Consequently,

$$f_m(W_{t+1}) = f_m(W_t) + \eta\frac{q}{m}\big(1 - f_m(W_t)\big),$$

or equivalently

$$1 - f_m(W_{t+1}) = \left(1 - \eta\frac{q}{m}\right)\big(1 - f_m(W_t)\big).$$

Because $0 < \eta q/m \le 1$, induction shows that the activation pattern persists, $f_m(W_t) \le 1$, and Eqs. (11.67) and (11.68) hold for every t. Finally, Hoeffding's inequality gives

$$\mathbb{P}\left(\frac{Q}{m} < \frac{1}{2} - \delta\right) \le e^{-2m\delta^2}.$$

On the complementary event, the factor in Eq. (11.68) is at most $1 - \eta(1/2 - \delta)$, proving Eq. (11.69). □

This example does not make the ReLU objective smooth. It makes the bad event "every available gate is inactive" exponentially unlikely. For several training examples, the corresponding requirement is stronger: the active gates must collectively cover every residual direction.

To state that requirement, fix training pairs $(x_i, y_i)_{i=1}^n$ with $x_i \in \mathbb{R}^p$, and consider

$$f_W(x) = \frac{1}{\sqrt{m}}\sum_{r=1}^m a_r [w_r^\top x]_+, \qquad a_r \in \{-1, 1\},$$

where the output weights a_r are fixed and the hidden weights $W = (w_1, \dots, w_m)$ are trained. Let

$$F_i(W) = f_W(x_i) - y_i, \qquad L(W) = \frac{1}{2}\|F(W)\|_2^2.$$

At a parameter with no zero training preactivation, define the *empirical tangent Gram matrix*

$$K(W) = J_F(W) J_F(W)^\top. \tag{11.70}$$

Its entries are

$$K_{ij}(W) = \frac{x_i^\top x_j}{m} \sum_{r=1}^{m} \mathbb{I}\{w_r^\top x_i > 0\}\mathbb{I}\{w_r^\top x_j > 0\}. \tag{11.71}$$

Thus $K(W) \succeq \lambda I_n$ says that the active parameter derivatives can change every training-residual direction with strength at least $\sqrt{\lambda}$.

Proposition 11.7.4 (Persistent tangent coverage gives loss convergence)**.** *Suppose an absolutely continuous parameter trajectory follows gradient flow at almost every time at which L is differentiable, and suppose*

$$K(W(t)) \succeq \lambda I_n \qquad \text{for almost every } t \geq 0$$

for some $\lambda > 0$. Then

$$L(W(t)) \leq e^{-2\lambda t} L(W(0)). \tag{11.72}$$

Proof. At every time covered by the assumptions,

$$\dot{W}(t) = -\nabla L(W(t)) = -J_F(W(t))^\top F(W(t)),$$

and hence

$$\dot{F}(W(t)) = -K(W(t))F(W(t)).$$

Therefore, at almost every time,

$$\frac{d}{dt} L(W(t)) = -F(W(t))^\top K(W(t))F(W(t)) \leq -2\lambda L(W(t)).$$

Integrating this differential inequality proves Eq. (11.72). □

At random initialization, Eq. (11.71) is an average of independent random gate matrices. With sufficiently large width, it can concentrate near a positive-definite limiting matrix. The second part of a finite-width argument must then prove that the weights move little enough for $K(W(t))$ to remain well conditioned. This gives a bootstrap: well-conditioned tangent features decrease the loss, the resulting movement is small, and the small movement preserves tangent conditioning. Such results give global convergence from the stated random initialization, not a claim that the objective has favorable geometry at every parameter.

At infinite width under the standard $1/\sqrt{m}$ scaling, the tangent Gram matrix can converge to a deterministic matrix that remains constant during training. This is the *neural tangent kernel (NTK) regime*.[3]

Exercise 7 asks you to derive the exact scalar recursion and, as an optional extension, solve the fixed-kernel residual dynamics. The finite-width concentration and stability proof is substantially longer and is not included here.

11.7.5 Softmax policy gradient

Post-training by reinforcement learning replaces fixed supervised targets by rewards obtained from actions. The softmax parameterization remains, but the distribution of visited states now depends on the current policy. Even a softmax bandit, the one-state special case, already seeks an optimal policy on the boundary of the probability simplex.

Policy gradient applies gradient ascent to expected discounted return, viewed as a function of the policy parameters. We discuss ascent here, whereas the rest of the lecture states descent results for losses; changing the sign of the objective converts one convention into the other.

A finite discounted *Markov decision process* specifies finite state and action sets, transition probabilities, rewards, and a discount factor. A tabular softmax policy can be written

$$\pi_\theta(a \mid s) = \frac{e^{\theta(s,a)}}{\sum_b e^{\theta(s,b)}}.$$

Here *tabular* means that every state–action pair has its own parameter $\theta(s, a)$. The performance objective is nonconvex in θ, and the objective's gradient depends on the state distribution induced by the current policy. Nevertheless, exact softmax policy gradient has a global $O(1/t)$ convergence guarantee under standard finite tabular assumptions. The proof uses a nonuniform PL-type gradient certificate whose coefficient depends on the probabilities that the current policy assigns to optimal actions. There is no single positive PL constant valid for every policy.

Entropy regularization adds a term that rewards policies for spreading probability across actions. This prevents the target policy from lying on the boundary of the probability simplex, where some action probabilities are zero. Under the corresponding assumptions, the regularized objective admits a linear convergence result. The same distinction appears in supervised softmax: when the target lies on the probability-simplex boundary, the certificate can degenerate and the rate is sublinear; a uniform positive lower bound on the relevant probabilities can give a linear rate. The policy-gradient proof has additional terms involving the *discounted occupancy measure*. This measure records how often the policy visits each state, with later visits downweighted by the discount factor. It depends on the current policy, so the proof cannot treat visited states as a fixed supervised dataset.

11.7.6 Root finding and fixed points

A root-finding update can converge without being derived from an objective. One useful condition on the residual map is *strong monotonicity*: for some $\mu > 0$,

$$\langle g(\theta) - g(\theta'), \theta - \theta' \rangle \geq \mu \|\theta - \theta'\|_2^2 \qquad \text{for all } \theta, \theta' \in \mathbb{R}^d. \tag{11.73}$$

Suppose also that g is C-Lipschitz: $\|g(\theta) - g(\theta')\|_2 \leq C\|\theta - \theta'\|_2$.

Proposition 11.7.5 (Strongly monotone root finding)**.** *Suppose that $g : \mathbb{R}^d \to \mathbb{R}^d$ satisfies Eq. (11.73), is C-Lipschitz, and has a root $\theta^\star$. Then the root is unique. For the iteration*

$$\theta_{t+1} = \theta_t - \eta g(\theta_t), \tag{11.74}$$

if $0 < \eta < 2\mu/C^2$, then

$$\|\theta_t - \theta^\star\|_2^2 \leq \left(1 - 2\eta\mu + \eta^2 C^2\right)^t \|\theta_0 - \theta^\star\|_2^2. \tag{11.75}$$

Proof. Two roots $\theta^\star$ and $\widetilde{\theta}^\star$ would make the left side of Eq. (11.73) zero, so they must be equal. Put $e_t = \theta_t - \theta^\star$. Since $g(\theta^\star) = 0$,

$$\begin{aligned} \|e_{t+1}\|_2^2 &= \|e_t - \eta(g(\theta_t) - g(\theta^\star))\|_2^2 \\ &\leq (1 - 2\eta\mu + \eta^2 C^2)\|e_t\|_2^2. \end{aligned}$$

Strong monotonicity, Cauchy–Schwarz, and Lipschitz continuity imply $\mu \leq C$. Hence the factor is nonnegative, and the stepsize condition makes it strictly smaller than one. Iteration proves Eq. (11.75). □

This is also an aligned-update proof with a Lyapunov function. For

$$V(\theta) = \frac{1}{2}\|\theta - \theta^\star\|_2^2, \qquad d(\theta) = -g(\theta),$$

we have $\nabla V(\theta) = \theta - \theta^\star$ and

$$-\langle \nabla V(\theta), d(\theta)\rangle \geq \mu \|\nabla V(\theta)\|_2^2, \qquad \|d(\theta)\|_2 \leq C\|\nabla V(\theta)\|_2.$$

Thus the alignment is with the gradient of the analysis certificate V, not with the gradient of an objective used to construct the update.

A map $\mathcal{T}$ is a *contraction* if $\|\mathcal{T}(\theta) - \mathcal{T}(\theta')\|_2 \leq q\|\theta - \theta'\|_2$ for some $q < 1$. The contraction mapping theorem gives a unique fixed point $\theta^\star$, and the direct iteration $\theta_{t+1} = \mathcal{T}(\theta_t)$ satisfies

$$\|\theta_t - \theta^\star\|_2^2 \leq q^{2t}\|\theta_0 - \theta^\star\|_2^2.$$

Although the distance to $\theta^\star$ is not computable without knowing the solution, the fixed-point residual $g(\theta) = \theta - \mathcal{T}(\theta)$ is computable and satisfies

$$(1-q)\|\theta - \theta^\star\|_2 \leq \|g(\theta)\|_2 \leq (1+q)\|\theta - \theta^\star\|_2.$$

Exercises 17 and 18 ask you to develop these two arguments and to distinguish the update from the Lyapunov function used to analyze it.

11.7.7 Linear temporal-difference learning

Policy evaluation supplies a concrete incremental root-finding example. Fix a policy in a finite discounted Markov decision process. Let P_π be its state-transition matrix, let d be a stationary probability vector with positive entries, meaning $d^\top P_\pi = d^\top$. Put $D = \mathrm{diag}(d)$, and let r be the vector of expected one-step rewards. Let $\phi(s) \in \mathbb{R}^p$ be a feature vector and let Φ have row s equal to $\phi(s)^\top$. The linear value approximation is $v_w(s) = \phi(s)^\top w$.

To identify the mean update, consider a transition with

$$\mathbb{P}(S_t = s) = d_s, \qquad \mathbb{P}(S_{t+1} = s' \mid S_t = s) = P_\pi(s, s'), \qquad \mathbb{E}[R_{t+1} \mid S_t = s] = r_s.$$

With discount factor $\gamma_{\mathrm{disc}} \in [0, 1)$, the TD(0) increment at a fixed parameter w is

$$\big(R_{t+1} + \gamma_{\mathrm{disc}}\phi(S_{t+1})^\top w - \phi(S_t)^\top w\big)\phi(S_t).$$

Its expectation is $b - Aw$, where

$$A = \Phi^\top D(I - \gamma_{\mathrm{disc}} P_\pi)\Phi, \qquad b = \Phi^\top D r. \tag{11.76}$$

The corresponding mean iteration is therefore

$$w_{t+1} = w_t + \eta(b - Aw_t). \tag{11.77}$$

This calculation connects TD to the main lecture. The algorithm starts from the linear root equation $g(w) = Aw - b = 0$ and uses the computable direction $-g(w) = b - Aw$. A *stationary point of the mean iteration* is a parameter at which its expected increment vanishes. Thus it must satisfy

$$Aw^\star = b.$$

This is also a fixed point of Eq. (11.77). Because A is generally nonsymmetric, $b - Aw$ is not usually the negative gradient of a scalar objective. This is clearly also an example of root finding. As such, in the proof below we check the conditions of Proposition 11.7.5 to get convergence.

Theorem 11.7.6 (Linear convergence of the mean TD iteration)**.** *Suppose that $d_s > 0$ for every state, $d^\top P_\pi = d^\top$, $\gamma_{\text{disc}} < 1$, and Φ has full column rank. Then A is invertible, so the mean TD iteration has the unique stationary point $w^\star = A^{-1}b$. Define*

$$H = \frac{A + A^\top}{2}, \qquad a = \lambda_{\min}(H), \qquad C = \|A\|_{\text{op}}. \tag{11.78}$$

Then $a > 0$. If $0 < \eta < 2a/C^2$, the Lyapunov function $V(w) = \|w - w^\star\|_2^2/2$ satisfies

$$V(w_t) \le \left(1 - 2\eta a + \eta^2 C^2\right)^t V(w_0). \tag{11.79}$$

Consequently, $w_t \to w^\star$.

Proof. For vectors indexed by states, write $\langle u, v\rangle_D = u^\top D v$ and $\|u\|_D^2 = u^\top D u$. Jensen's inequality and $d^\top P_\pi = d^\top$ give

$$\|P_\pi u\|_D^2 \le \sum_s d_s \sum_{s'} P_\pi(s, s') u(s')^2 = \|u\|_D^2.$$

Thus Cauchy–Schwarz gives $\langle u, P_\pi u\rangle_D \le \|u\|_D^2$. For $u = \Phi z$,

$$z^\top A z = \|u\|_D^2 - \gamma_{\text{disc}} \langle u, P_\pi u\rangle_D \ge (1 - \gamma_{\text{disc}})\|u\|_D^2 > 0 \qquad (z \ne 0).$$

Thus the symmetric part $(A + A^\top)/2$ is positive definite. It follows that A is *positive stable*: all of its eigenvalues have positive real parts. In particular, A is invertible and $w^\star = A^{-1}b$ is the unique stationary point.

For $g(w) = Aw - b$ and any $w, w' \in \mathbb{R}^p$,

$$\langle g(w) - g(w'), w - w'\rangle = (w - w')^\top H (w - w') \ge a\|w - w'\|_2^2,$$

while $\|g(w) - g(w')\|_2 \le C\|w - w'\|_2$. Thus g is a-strongly monotone and C-Lipschitz. Applying Proposition 11.7.5 proves Eq. (11.79). □

This proof also makes the relation to aligned updates explicit. If $e = w - w^\star$ and $d(w) = b - Aw = -Ae$, then $\nabla V(w) = e$ and

$$-\langle \nabla V(w), d(w)\rangle = e^\top A e \ge a\|\nabla V(w)\|_2^2, \qquad \|d(w)\|_2 \le C\|\nabla V(w)\|_2. \tag{11.80}$$

The TD direction is therefore aligned with the negative gradient of V even though it is not itself the negative gradient of a scalar objective. The function V is 1-smooth and 1-PL, so Theorem 11.3.1 gives geometric convergence for $0 < \eta \le a/C^2$. The direct root-finding calculation in Theorem 11.7.6 gives the larger sufficient range $0 < \eta < 2a/C^2$. The algorithm evaluates $b - Aw$; it does not know $w^\star$ and does not evaluate V. The Lyapunov function is an analysis certificate constructed after the update has been specified.

Weighted extension. The TD assumptions above make the symmetric part of A positive definite, so the ordinary squared distance is enough. A weighted distance is useful for a more general linear root-finding update in which A is positive stable but its symmetric part need not be positive definite.

Proposition 11.7.7 (Weighted contraction for a positive-stable update)**.** *Let $A \in \mathbb{R}^{p\times p}$ be positive stable, let $b \in \mathbb{R}^p$, and consider $w_{t+1} = w_t + \eta(b - Aw_t)$. For any $Q \succ 0$, let $M \succ 0$ be the unique solution of the continuous* Lyapunov equation

$$A^\top M + M A = Q. \tag{11.81}$$

Put $w^\star = A^{-1}b$, $e_t = w_t - w^\star$, *and* $V_M(w) = (w - w^\star)^\top M(w - w^\star)$. *Let* $q_Q = \lambda_{\min}(Q)$ *and* $c_A = \lambda_{\max}(A^\top MA)$. *If*

$$0 < \eta \le \min\left\{\frac{q_Q}{2c_A}, \frac{\lambda_{\max}(M)}{q_Q}\right\},$$

then

$$V_M(w_t) \le \left(1 - \frac{\eta q_Q}{2\lambda_{\max}(M)}\right)^t V_M(w_0). \tag{11.82}$$

Proof. The equation constructs the matrix M; the Lyapunov function is V_M. Since $e_{t+1} = (I - \eta A)e_t$,

$$\begin{aligned} V_M(w_{t+1}) - V_M(w_t) &= -\eta e_t^\top (A^\top M + MA)e_t + \eta^2 e_t^\top A^\top MAe_t \\ &= -\eta e_t^\top Qe_t + \eta^2 e_t^\top A^\top MAe_t. \end{aligned} \tag{11.83}$$

The stated stepsize bound gives

$$V_M(w_{t+1}) - V_M(w_t) \le -\frac{\eta q_Q}{2}\|e_t\|_2^2 \le -\frac{\eta q_Q}{2\lambda_{\max}(M)} V_M(w_t).$$

The contraction factor belongs to $[1/2, 1)$, and iteration proves Eq. (11.82). □

Exercise 19 asks you to derive the Lyapunov contraction and to relate it to alignment with $-\nabla V_M$. It also asks you to prove why nonsymmetry rules out a gradient interpretation. Actual TD uses dependent sampled transitions rather than the mean update Eq. (11.77); its convergence theorem additionally controls sampling noise and Markov dependence.

Optional value-function interpretation. The convergence proof uses the stationary equation $Aw^\star = b$ and the matrix properties established above; it does not require a Bellman interpretation. To connect this equation to policy evaluation, define the D-orthogonal projection onto the column space of Φ by

$$\Pi_D = \Phi(\Phi^\top D\Phi)^{-1}\Phi^\top D.$$

Then $Aw^\star = b$ is equivalent to

$$\Phi w^\star = \Pi_D\big(r + \gamma_{\text{disc}} P_\pi \Phi w^\star\big). \tag{11.84}$$

This is the *projected Bellman equation*: the Bellman update $r + \gamma_{\text{disc}} P_\pi \Phi w^\star$ need not lie in the feature space, so Π_D projects it back to that space. Thus $\Phi w^\star$ is the projected Bellman fixed point.

11.8 Take-home messages

- Objective-value convergence is the minimum optimization conclusion used by the solution-selection analyses in Lecture 10. It does not by itself imply finite parameter convergence; when parameters diverge, their normalized directions may still converge.
- Smoothness converts a local update into objective decrease. Convex comparison, PL, or another gradient-to-gap relation can convert that decrease into convergence to the global infimum. Without smoothness, even $|x|$ can make a fixed-step method oscillate; convex subgradient descent instead controls distance to a comparator and averages the iterates.
- Unregularized fixed-feature softmax does not have one universal rate. Convexity gives convergence against finite comparators, overlap on a bounded reduced parameter set gives a linear rate, and separation gives $O(1/t)$ loss decay toward an unattained infimum without a uniform PL inequality.

- Under PL, constant-stepsize SGD converges to a variance-dependent neighborhood in the stated bound, while a $1/t$ stepsize gives an $O(1/t)$ expected objective gap. The same calculation applies to stochastic directions with expected alignment and a conditional second-moment bound.
- The negative gradient is not essential. Aligned directions suffice for the same PL argument. More generally, an algorithm can start from a root or fixed-point equation rather than a loss; a Lyapunov function can then be constructed as an analysis certificate for its parameter update.

Endnotes

1. The regularized calculation is still useful as a comparison. The Hessian becomes $\nabla^2 L_n(W) + \lambda I$, so the objective is λ-strongly convex and satisfies a λ-PL inequality. The results in the main text do not require this modification.

2. In dynamical-systems and control texts, a Lyapunov function is usually tied to a fixed point $\theta^\star$ and is required to be continuous and *positive definite* about it: $V(\theta^\star) = 0$ and $V(\theta) > 0$ for $\theta \neq \theta^\star$. Comparison bounds relating $V(\theta)$ to $\|\theta - \theta^\star\|_2$ turn decrease of V into control of parameter distance. Together with such bounds, nonincrease gives Lyapunov stability, while a lower bound on $V(\theta) - V(\mathcal{T}(\theta))$ by a positive function of $\|\theta - \theta^\star\|_2$ gives asymptotic convergence. A *radially unbounded* function, meaning $V(\theta) \to \infty$ as $\|\theta\|_2 \to \infty$, has bounded sublevel sets and can support global rather than only local conclusions. We do not impose these properties here. In separable softmax there is no finite target parameter and the parameter norm diverges. Hence no radially unbounded function can both decrease along those iterates and serve as a Lyapunov function for their limit.

3. If the limiting matrix is $K_\infty = U \operatorname{diag}(\lambda_1, \ldots, \lambda_n) U^\top$, then the fixed-kernel residual dynamics have the explicit solution $F(t) = U \operatorname{diag}(e^{-\lambda_1 t}, \ldots, e^{-\lambda_n t}) U^\top F(0)$. The spectrum therefore determines the optimization rates, while the fixed kernel together with the initialization determines the selected interpolant. This converts nonlinear parameter training into an exactly analyzable linear system. It also shifts the substantive questions to the conditioning and suitability of the initial tangent features. Because those features remain essentially fixed, the regime does not describe substantial feature learning. It is a useful optimization model and comparison point, rather than a general description of finite-network training.

Bibliographical notes

The PL condition was introduced by Polyak and is closely related to an inequality of Łojasiewicz. The modern optimization treatment, including relations among PL, error bounds, convexity, and stochastic-gradient rates, is developed by Karimi et al. [2016]. Standard smooth convex comparison bounds and the nonsmooth subgradient method appear in many convex-optimization texts; see, for example, Bubeck [2015].

The scalar wide-ReLU calculation in Proposition 11.7.3 is included to isolate one elementary effect of redundant gates. The finite-sample tangent-Gram mechanism is the basis of the two-layer ReLU convergence theorem of Du et al. [2019]: under random initialization, nondegenerate inputs, sufficient width, and a suitably small stepsize, gradient descent reaches zero squared training loss at a linear rate with high probability. Deep finite-width extensions use related near-initialization and semismoothness arguments; see Allen-Zhu et al. [2019]. The infinite-width neural tangent kernel formulation is due to Jacot et al. [2018]. These results motivate the mechanism summarized in Section 11.7.4; the concentration and activation-stability arguments needed for their full finite-width conclusions are not reproduced here.

Existence of finite logistic-regression maximum-likelihood estimates and the distinction among overlap, complete separation, and quasi-complete separation were characterized by Albert and Anderson [1984]. The bounded sublevel condition after removing invariant parameter directions is a direct sufficient condition for the uniform probability and Hessian bounds needed by our proof. In quasi-separated cases the infimum may also fail to be attained, while the strict margin γ in Eq. (11.37) can equal zero. The general comparator argument still proves convergence in objective

value, but the explicit separable rate requires a positive margin. The max-margin directional limit for gradient descent on separable logistic and multiclass cross-entropy losses is due to Soudry et al. [2018].

The tabular softmax policy-gradient results summarized in Section 11.7.5 are due to Mei et al. [2020]. They prove a global $O(1/t)$ rate for the unregularized exact-gradient method and linear convergence for an entropy-regularized variant, under their stated finite-MDP conditions. The linear TD discussion abstracts the Lyapunov mechanism from the on-policy stochastic convergence analysis of Tsitsiklis and Van Roy [1997].

Glossary

Aligned update
An update direction that has a uniformly positive inner product with the negative gradient and whose norm is controlled by the gradient norm.

Averaged iterate
A weighted average of earlier parameters, such as $\bar{\theta}_T = (\sum_{t<T} \eta_t)^{-1} \sum_{t<T} \eta_t \theta_t$. For nonsmooth convex optimization, its objective value can converge even when the individual iterates oscillate.

Comparator bound
A guarantee that compares an iterate's objective value with the value at any fixed reference parameter, whether or not a minimizer exists.

Conditional drift inequality
A bound on the expected one-step change of a Lyapunov function after the current random parameter is fixed. It is the stochastic counterpart of a deterministic decrease inequality.

Contraction
A map $\mathcal{T}$ satisfying $\|\mathcal{T}(\theta) - \mathcal{T}(\theta')\|_2 \le q\|\theta - \theta'\|_2$ for some $q < 1$. Repeated application converges to its unique fixed point.

Condition number
The ratio $\kappa = \beta/\mu$ between the smoothness constant and the PL constant. It controls the linear-convergence iteration bound proved here.

Dead ReLU
A ReLU unit whose preactivation is negative on the relevant inputs, so both its output and its backpropagated parameter derivative vanish. Width can make it unnecessary for every individual unit to be active.

Directional convergence
Convergence of normalized iterates $\theta_t/\|\theta_t\|_2$ when the parameter norm may diverge. It can describe solution selection even when finite parameter convergence fails.

Discounted occupancy measure
The discounted frequency with which a policy visits each state. It changes with the policy and enters policy-gradient formulas because the policy affects both actions and future state visitation.

Entropy regularization
Adding an entropy term that rewards a policy for spreading probability across actions. It discourages action probabilities from approaching zero.

Fixed point
A parameter $\theta^\star$ satisfying $\mathcal{T}(\theta^\star) = \theta^\star$. Equivalently, it is a root of $g(\theta) = \theta - \mathcal{T}(\theta)$.

Gradient certificate
An inequality connecting a locally available quantity, such as the gradient norm, to a global objective gap.

Gradient dominance
A gradient certificate of the form $\|\nabla L(\theta)\|_2^2/2 \geq \mu(L(\theta) - L^\star)^\alpha$. Under smooth descent, $\alpha = 1$ gives geometric decay and $\alpha > 1$ gives polynomial decay.

Invariant parameter direction
A parameter direction along which the model predictions and objective do not change. Common shifts of all softmax rows and feature-null directions are examples in this lecture.

Line search
A procedure that tests candidate stepsizes and accepts one satisfying a stated objective-decrease condition.

Linear convergence
Geometric decrease of an error measure: $V_t \leq q^t V_0$ for some $q \in [0, 1)$.

Lyapunov equation
For a positive stable matrix A and $Q \succ 0$, the matrix equation $A^\top M + MA = Q$. Its positive-definite solution M can be used to construct a Lyapunov function; the equation itself is not that function.

Lyapunov function
A nonnegative function of the parameter whose infimum is zero and whose value decreases along the update. This is the minimal optimization sense used in these notes. Classical stability theory often also requires continuity, positive definiteness about a fixed point, and conditions that relate the function to distance from that point. The function may be constructed only for the convergence proof and need not be evaluated by the algorithm. Quantitative decrease can force its values to zero even when no finite parameter attains zero.

Markov decision process (MDP)
A sequential decision model specified by states, actions, transition probabilities, rewards, and, here, a discount factor. A policy assigns an action distribution to each state.

Neural tangent kernel (NTK) regime
An infinite-width training regime in which the Gram matrix of parameter derivatives approaches a deterministic limit and remains approximately fixed. The prediction dynamics then reduce to kernel gradient descent with fixed features.

Noise floor
In a stochastic bound with a fixed stepsize, the additive term that remains after the term depending on initialization has decayed. It is a property of the bound, not automatically a lower bound on the algorithm's limiting error.

Nonsmooth objective
An objective that is not differentiable everywhere. A ReLU training objective can be differentiable within one activation region but fail to be differentiable where an activation changes sign.

Objective gap
The difference $L(\theta) - L^\star$ between the current objective value and its infimum over all parameters.

Objective-value convergence
Convergence $L(\theta_t) \to L^\star$ of the achieved objective values to their infimum. It does not by itself imply convergence of the parameters.

Overlap regime
Here, a regime in which the relevant reduced softmax sublevel set is bounded, so every fitted class probability on that set has a uniform positive lower bound. This supplies the Hessian lower bound used for linear convergence.

Parameter convergence
Convergence of the iterates themselves to a finite parameter vector. It is stronger than objective-value convergence.

Polyak–Łojasiewicz inequality
The gradient certificate $\|\nabla L(\theta)\|_2^2/2 \geq \mu(L(\theta) - L^\star)$. It does not require convexity.

Policy gradient
Gradient ascent on expected return with respect to policy parameters. The gradient includes the state-visitation distribution induced by the policy.

Positive stable matrix
A square matrix whose eigenvalues all have positive real parts.

Preconditioned update
An update that multiplies the gradient by a matrix, thereby rescaling or mixing parameter coordinates before taking a step.

Projected Bellman equation
The equation $v = \Pi_D(r + \gamma_{\text{disc}} P_\pi v)$ for an approximate value function v in the feature space. The projection returns the Bellman update to that feature space.

Quasi-separated sample
A sample with a direction having nonnegative margins everywhere and a strict margin somewhere, but no direction with a strict margin everywhere. Its logistic-loss infimum can fail to be attained even when that infimum is positive.

Root-finding problem
The problem of finding a parameter $\theta^\star$ satisfying $g(\theta^\star) = 0$ for a given map g. The update can be designed from the computable residual $g(\theta)$ without being the gradient of an objective.

Separable sample
In the multiclass softmax setting, a sample admitting a score matrix with a strictly positive correct-class score gap against every incorrect class.

Smoothness
The Lipschitz-gradient condition $\|\nabla L(u) - \nabla L(v)\|_2 \leq \beta\|u - v\|_2$. It controls the error of a first-order approximation over a finite step.

Stationary distribution
A state distribution d satisfying $d^\top P_\pi = d^\top$. If the state has distribution d before one transition under P_π, it also has distribution d afterward.

Stationary point
For a gradient method, a parameter θ at which $\nabla L(\theta) = 0$; it is ϵ-stationary when $\|\nabla L(\theta)\|_2 \le \epsilon$. For a general iteration, a stationary point is a parameter at which the update vanishes; the mean TD condition is $Aw = b$.

Stochastic approximation
The analysis of iterative updates driven by noisy estimates. Decreasing stepsizes are commonly used so that cumulative progress is unbounded while cumulative squared-stepsize noise remains finite.

Strong convexity
Convexity with a uniform positive quadratic lower bound on departure from every tangent plane.

Strong monotonicity
For a map $g : \mathbb{R}^d \to \mathbb{R}^d$, the condition $\langle g(\theta) - g(\theta'), \theta - \theta' \rangle \ge \mu \|\theta - \theta'\|_2^2$ for some $\mu > 0$. Together with Lipschitz continuity, it gives a contraction rate for the root-finding update studied here.

Subgradient and subdifferential
For a convex function, a subgradient s at θ supplies the global affine lower bound $L(u) \ge L(\theta) + \langle s, u - \theta \rangle$ for every u. The subdifferential $\partial L(\theta)$ is the set of all such vectors.

Sublevel set
For a number c, the set of parameters satisfying $L(\theta) \le c$.

Tabular policy
A policy parameterization with a separate parameter for every state–action pair, rather than parameters shared through features or a neural network.

Tangent Gram matrix
For a vector of training residuals $F(\theta)$ with Jacobian $J_F(\theta)$, the matrix $J_F(\theta) J_F(\theta)^\top$. A positive lower eigenvalue means that parameter changes can affect every residual direction.

Temporal-difference learning
A value-learning method that updates from the discrepancy between a current value estimate and a one-step bootstrapped target: the observed reward plus a discounted next-state estimate. Its mean update need not be a gradient of an objective.

11.9 Exercises

General instructions. Justify every answer, including yes/no answers, even when the question does not explicitly ask for a justification.

Exercise profiles. Filled mountains show difficulty on a four-mountain scale: one is a direct application, two uses a familiar method, three combines ideas, and four is a substantial challenge. Filled cogs show technical load on a three-cog scale: one is light, two is sustained, and three is dense. Filled hourglasses show estimated time: one is under 20 minutes, two is 20–45 minutes, three is 45–90 minutes, and four is more than 90 minutes. Pale pictograms show the unused part of

each scale. The work row uses the canonical order proof, calculation, construction, computation, and interpretation; pale icons mark work types not used. marks the intended learning outcome. Time estimates cover the non-optional parts. Ratings are not prerequisites or grading weights. A superscript asterisk marks an exercise based on supplementary lecture material; it is not a difficulty rating and does not make the exercise optional when assigned.

Local descent and global certificates

Exercise 1: Equivalent smoothness and Hessian conditions.

Relate quadratic first-order bounds, Hessian inequalities, convexity, and Lipschitz continuity of the gradient.

Let $L : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable. For $\beta \geq 0$, consider the quadratic upper bound

$$L(v) \leq L(u) + \langle \nabla L(u), v - u \rangle + \frac{\beta}{2}\|v - u\|_2^2 \qquad \text{for every } u, v \in \mathbb{R}^d. \tag{U_β}$$

(a) Prove that Eq. (U_β) holds if and only if

$$\nabla^2 L(\theta) \preceq \beta I \qquad \text{for every } \theta \in \mathbb{R}^d.$$

(b) Prove both Hessian characterizations

$$\begin{aligned} L \text{ is convex} &\iff \nabla^2 L(\theta) \succeq 0 \quad \text{for every } \theta, \\ L \text{ is } \mu\text{-strongly convex} &\iff \nabla^2 L(\theta) \succeq \mu I \quad \text{for every } \theta. \end{aligned}$$

(c) Prove that L is β-smooth if and only if

$$-\beta I \preceq \nabla^2 L(\theta) \preceq \beta I \qquad \text{for every } \theta \in \mathbb{R}^d.$$

(d) Conclude that, when L is convex, β-smoothness is equivalent to Eq. (U_β). Verify that $L(x) = -x^4$ satisfies Eq. (U_β) with $\beta = 0$ but is not smooth with any finite constant. Explain why this example prevents the same conclusion without convexity.

Hints For the Hessian implications, restrict L to the line $t \mapsto u + t(v - u)$. For the reverse implications, use

$$\phi(1) = \phi(0) + \phi'(0) + \int_0^1 (1 - t)\phi''(t)\, dt.$$

For part (c), apply the fundamental theorem of calculus to the gradient along the line segment from u to v.

Exercise 2: Smooth descent can stop at the wrong value.

Distinguish convergence to stationarity from convergence to the global infimum.

Let $L(\theta) = 1 - \cos\theta$ for $\theta \in \mathbb{R}$.

(a) Prove that L is 1-smooth, determine $L^\star$, and identify all global minimizers.

(b) For arbitrary $\eta > 0$, run gradient descent from $\theta_0 = \pi$. Determine θ_t, $L(\theta_t)$, and $|L'(\theta_t)|$ for every t.

(c) Explain precisely why the smooth nonconvex stationarity bound

$$\min_{0\le t<T} |L'(\theta_t)|^2 \le \frac{2(L(\theta_0)-L^\star)}{\eta T}$$

does not imply that the objective value approaches $L^\star$ in this example.

HINTS Use $L'(\theta)=\sin\theta$ and the mean value theorem for the derivative.

Exercise 3: From strong convexity to a linear rate.

Derive the PL certificate supplied by strong convexity and use it in a descent recursion.

Let $L:\mathbb{R}^d\to\mathbb{R}$ be differentiable, μ-strongly convex, and β-smooth. Thus, for every $u,v\in\mathbb{R}^d$,

$$L(v)\ge L(u)+\langle \nabla L(u), v-u\rangle+\frac{\mu}{2}\|v-u\|_2^2.$$

(a) Compare $L(u)$ with $L(v)$ for an arbitrary v and prove

$$\frac{1}{2}\|\nabla L(u)\|_2^2 \ge \mu\big(L(u)-L^\star\big).$$

(b) Suppose gradient descent with $0<\eta\le 1/\beta$ satisfies

$$L(\theta_{t+1})\le L(\theta_t)-\frac{\eta}{2}\|\nabla L(\theta_t)\|_2^2.$$

Derive an explicit upper bound on $L(\theta_t)-L^\star$.

HINTS For (a), rearrange the strong-convexity inequality, apply Cauchy–Schwarz, and maximize the resulting bound over $r=\|u-v\|_2$. For (b), first obtain a one-step recursion for the objective gap.

Exercise 4: Rates beyond PL.[*]

Derive the polynomial rate produced by nonlinear gradient dominance and identify the exponent of a quartic objective.

Let $L:\mathbb{R}^d\to\mathbb{R}$ be differentiable and bounded below. Suppose gradient descent satisfies

$$L(\theta_{t+1})\le L(\theta_t)-\frac{\eta}{2}\|\nabla L(\theta_t)\|_2^2$$

for some $\eta>0$, and define

$$\Delta_t = L(\theta_t)-L^\star.$$

Suppose that, throughout the initial sublevel set, constants $\mu>0$ and $\alpha\ge 1$ satisfy

$$\frac{1}{2}\|\nabla L(\theta)\|_2^2 \ge \mu\big(L(\theta)-L^\star\big)^\alpha.$$

(a) Prove

$$\Delta_{t+1}\le \Delta_t-\eta\mu\Delta_t^\alpha.$$

(b) Suppose $\alpha>1$ and $\Delta_0>0$. Prove that

$$\Delta_t \le \left(\Delta_0^{1-\alpha}+\eta\mu(\alpha-1)t\right)^{-1/(\alpha-1)}.$$

State the resulting power of t in the asymptotic upper bound.

(c) Let $L(x) = x^4/4$, $x_0 \neq 0$, and $0 < \eta \leq 1/(3x_0^2)$. Show that the gradient-descent iterates remain in $[-|x_0|, |x_0|]$, that L is $3x_0^2$-smooth on this interval, and that

$$\frac{1}{2}|L'(x)|^2 = 4L(x)^{3/2}.$$

Deduce the explicit bound

$$L(x_t) \leq \Big(L(x_0)^{-1/2} + 2\eta t\Big)^{-2}.$$

(d) *Optional extension.* Suppose a β-smooth objective has a sequence of parameters whose positive objective gaps tend to zero. Prove that no constants $\mu > 0$ and $\alpha < 1$ can satisfy the displayed gradient-dominance inequality along this sequence.

Hints For (b), put $s = \alpha - 1$ and apply $(1 - z)^{-s} \geq 1 + sz$ to the recurrence from part (a). For the optional extension, apply the descent lemma to $\theta - \nabla L(\theta)/\beta$ to obtain an upper bound on the squared gradient norm in terms of the objective gap.



Exercise 5: Convergence without an attained minimum.

Use a convex comparator bound to prove convergence to an infimum that need not be attained.

Let $L : \mathbb{R}^d \to \mathbb{R}$ be convex and β-smooth. Gradient descent uses $0 < \eta \leq 1/\beta$.

(a) Starting from the squared-distance recursion and the descent inequality, prove that every $u \in \mathbb{R}^d$ satisfies

$$2\eta\big(L(\theta_{t+1}) - L(u)\big) \leq \|\theta_t - u\|_2^2 - \|\theta_{t+1} - u\|_2^2.$$

(b) Deduce

$$L(\theta_T) - L(u) \leq \frac{\|\theta_0 - u\|_2^2}{2\eta T}.$$

(c) Suppose only that $L^\star = \inf_u L(u) > -\infty$. For every $\epsilon > 0$, choose u_ϵ with $L(u_\epsilon) \leq L^\star + \epsilon$. Prove $L(\theta_T) \to L^\star$ without assuming that a minimizer exists.

Hints Use convexity to bound $\langle \nabla L(\theta_t), \theta_t - u\rangle$. In (b), use monotonicity of the objective values before telescoping.



Exercise 6: Preconditioned directions under PL.

Identify the two controls on an approximate descent direction that preserve a global PL rate.

Let $L : \mathbb{R}^d \to \mathbb{R}$ be β-smooth and μ-PL. Consider

$$\theta_{t+1} = \theta_t + \eta d_t,$$

where, for constants $c, C > 0$,

$$-\langle \nabla L(\theta_t), d_t\rangle \geq c\|\nabla L(\theta_t)\|_2^2, \qquad \|d_t\|_2 \leq C\|\nabla L(\theta_t)\|_2.$$

(a) Prove that $0 < \eta \leq c/(\beta C^2)$ implies

$$L(\theta_{t+1}) \leq L(\theta_t) - \frac{\eta c}{2}\|\nabla L(\theta_t)\|_2^2.$$

Deduce a geometric bound on $L(\theta_t) - L^\star$.

(b) Let $d_t = -P_t \nabla L(\theta_t)$, where P_t is symmetric and $mI \preceq P_t \preceq MI$. Give valid values of c and C, and state the resulting sufficient upper bound on η and convergence factor.

HINTS Apply the descent lemma to ηd_t. For (b), use the Rayleigh quotient and the operator norm of P_t.

Exercise 7: Width, dead ReLUs, and tangent coverage.[*]

◎ Show how independent ReLU gates make complete gradient failure unlikely and relate the resulting calculation to fixed-kernel convergence.

For $w = (w_1, \dots, w_m) \in \mathbb{R}^m$, let

$$f_m(w) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} [w_r]_+, \qquad L_m(w) = \frac{1}{2}\big(f_m(w) - 1\big)^2.$$

Let $S_1, \dots, S_m$ be independent random signs with $\mathbb{P}(S_r = 1) = \mathbb{P}(S_r = -1) = 1/2$, fix $a \in (0, 1)$, and initialize

$$W_{r,0} = \frac{a S_r}{\sqrt{m}}.$$

Gradient descent uses $W_{t+1} = W_t - \eta \nabla L_m(W_t)$ with $0 < \eta \le 1$. Define

$$Q = \sum_{r=1}^{m} \mathbb{I}\{S_r = 1\}.$$

(a) Show that

$$\mathbb{P}(Q = 0) = 2^{-m},$$

and that on the event $\{Q = 0\}$ the iterates never move and $L_m(W_t) = 1/2$ for every $t \ge 0$.

(b) Condition on $Q = q > 0$. Prove that the initially active coordinates remain positive, the initially inactive coordinates remain negative, and

$$1 - f_m(W_t) = \Big(1 - \eta \frac{q}{m}\Big)^t \big(1 - f_m(W_0)\big),$$
$$L_m(W_t) = \Big(1 - \eta \frac{q}{m}\Big)^{2t} L_m(W_0).$$

(c) You may use Hoeffding's inequality in the form

$$\mathbb{P}\left(\frac{Q}{m} < \frac{1}{2} - \delta\right) \le e^{-2m\delta^2}, \qquad 0 < \delta < \frac{1}{2}.$$

Prove that, with probability at least $1 - e^{-2m\delta^2}$,

$$L_m(W_t) \le \big(1 - \eta(1/2 - \delta)\big)^{2t} L_m(W_0) \qquad \text{for every } t \ge 0.$$

Explain which failure of the single-ReLU model is removed by width and which fact about the individual inactive units remains unchanged.

(d) *Optional extension.* Let $K_\infty \in \mathbb{R}^{n\times n}$ be symmetric and positive semidefinite, and suppose a residual vector follows

$$\dot r(t) = -K_\infty r(t).$$

If $K_\infty = U\,\mathrm{diag}(\lambda_1,\dots,\lambda_n)U^\top$ with U orthogonal, prove that

$$r(t) = U\,\mathrm{diag}(e^{-\lambda_1 t},\dots,e^{-\lambda_n t})U^\top r(0).$$

Deduce that $\|r(t)\|_2^2/2 \le e^{-2\lambda t}\|r(0)\|_2^2/2$ when $K_\infty \succeq \lambda I_n$. Explain both what the fixed-kernel assumption makes exactly analyzable and what kind of network behavior it excludes.

Hints For (b), first compute $f_m(W_0)$ and the derivative with respect to one active coordinate. Use induction to justify that the activation pattern does not change. For the optional extension, express the differential equation in the eigenbasis of K_∞.

Unregularized softmax

Exercise 8: Softmax curvature as a variance.

Express softmax curvature as categorical variance and obtain a global smoothness constant from the feature matrix.

Fix $x_i \in \mathbb{R}^p$, $y_i \in [K]$, and $W \in \mathbb{R}^{K\times p}$ with row k equal to $w_k^\top$. Define

$$q_W(k \mid x_i) = \frac{e^{w_k^\top x_i}}{\sum_{j=1}^K e^{w_j^\top x_i}}, \qquad L_n(W) = \frac{1}{n}\sum_{i=1}^n -\log q_W(y_i \mid x_i).$$

Let $\Delta \in \mathbb{R}^{K\times p}$ have row k equal to $\delta w_k^\top$.

(a) Derive

$$\nabla_{w_k} L_n(W) = \frac{1}{n}\sum_{i=1}^n \big(q_W(k \mid x_i) - \mathbb{I}\{y_i = k\}\big)x_i.$$

(b) For fixed W and i, let Y_i^W have mass function $\mathbb{P}(Y_i^W = k) = q_W(k \mid x_i)$. Prove

$$\langle \Delta, \nabla^2 L_n(W)\Delta\rangle_F = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}\big((\delta w_{Y_i^W})^\top x_i\big).$$

(c) Using $\mathrm{Var}(a_Y) \le \|a\|_2^2/2$, prove that L_n is $\lambda_{\max}(G)/2$-smooth in Frobenius norm, where $G = n^{-1}\sum_i x_i x_i^\top$.

Hints For (b), differentiate along the scalar path $W + s\Delta$. For (c), write $n^{-1}\sum_i \|\Delta x_i\|_2^2$ as a trace.

Exercise 9: The identified softmax subspace.

Remove the parameter directions that cannot change fitted softmax probabilities.

Fix $x_i \in \mathbb{R}^p$, $y_i \in [K]$, and $W \in \mathbb{R}^{K\times p}$ with row k equal to $w_k^\top$. Define

$$q_W(k \mid x_i) = \frac{e^{w_k^\top x_i}}{\sum_{j=1}^K e^{w_j^\top x_i}}, \qquad L_n(W) = \frac{1}{n}\sum_{i=1}^n -\log q_W(y_i \mid x_i).$$

Let $\mathbf{1} \in \mathbb{R}^K$ be the all-ones vector.

(a) Prove that for every $a \in \mathbb{R}^p$,

$$q_{W+\mathbf{1}a^\top}(k \mid x_i) = q_W(k \mid x_i)$$

for every i, k, and hence the loss is invariant under this transformation.

(b) Prove that the rows of $\nabla_W L_n(W)$ sum to zero. Deduce that if $\sum_k w_{0,k} = 0$, then gradient descent preserves this equality.

(c) Let $v \in \mathbb{R}^p$ satisfy $v^\top x_i = 0$ for every i. Explain why adding an arbitrary multiple of v to any row of W cannot change the fitted probabilities.

Hints In (a), cancel the common factor $e^{a^\top x_i}$ in the softmax ratio. In (b), sum the row-gradient formula over k.

Exercise 10: Restricted curvature from overlap.

Combine probability overlap with feature coverage to obtain strong curvature on the identified softmax subspace.

Let $q \in \mathbb{R}^K$ be a probability vector with $q_k \geq \alpha > 0$ for every k. Let $G = n^{-1} \sum_i x_i x_i^\top \succeq \lambda_X I$. Suppose $\Delta \in \mathbb{R}^{K \times p}$ has rows $\delta w_k^\top$ satisfying $\sum_k \delta w_k = 0$.

(a) For a class-valued random variable Y with mass function $\mathbb{P}(Y = k) = q_k$, prove

$$\mathrm{Var}(a_Y) = \min_{c \in \mathbb{R}} \sum_{k=1}^K q_k (a_k - c)^2 \geq \alpha \min_{c \in \mathbb{R}} \sum_{k=1}^K (a_k - c)^2.$$

(b) Apply part (a) with $a_k = (\delta w_k)^\top x_i$ and prove

$$\frac{1}{n} \sum_{i=1}^n \mathrm{Var}(a_{Y_i}) \geq \alpha \lambda_X \|\Delta\|_F^2$$

whenever every Y_i has all class probabilities at least α.

(c) State the PL constant and the gradient-descent convergence factor that follow on a sublevel set where these probability bounds hold.

Hints The minimizing c in the weighted expression is $\mathbb{E}[a_Y]$. In (b), the unweighted minimizing c is the average of the a_k, which is zero here.

Exercise 11: A gradient certificate from multiclass separation.

Convert a positive multiclass margin into a lower bound on the softmax gradient norm.

For fixed $x_i \in \mathbb{R}^p$, $y_i \in [K]$, and $W \in \mathbb{R}^{K \times p}$, define

$$q_W(k \mid x_i) = \frac{e^{w_k^\top x_i}}{\sum_{j=1}^K e^{w_j^\top x_i}}, \qquad L_n(W) = \frac{1}{n} \sum_{i=1}^n -\log q_W(y_i \mid x_i).$$

Suppose $\overline{W} \in \mathbb{R}^{K \times p}$ satisfies $\|\overline{W}\|_F \leq 1$ and

$$(\overline{w}_{y_i} - \overline{w}_k)^\top x_i \geq \gamma > 0 \qquad (i \in [n],\ k \neq y_i).$$

Let $\ell_i(W) = -\log q_W(y_i \mid x_i)$, and suppose $L_n(W) \le L_0$.

(a) Prove

$$-\langle \nabla L_n(W), \overline{W} \rangle_F \ge \frac{\gamma}{n} \sum_{i=1}^{n} \big(1 - q_W(y_i \mid x_i)\big).$$

(b) Prove that $0 < \ell_i(W) \le nL_0$ and, with $\rho_0 = (1 - e^{-nL_0})/(nL_0)$, that

$$1 - q_W(y_i \mid x_i) \ge \rho_0 \ell_i(W).$$

(c) Deduce the gradient certificate

$$\|\nabla L_n(W)\|_F \ge \gamma \rho_0 L_n(W).$$

HINTS Expand the inner product using the row-gradient formula. For (b), use the chord of the concave function $u \mapsto 1 - e^{-u}$ on $[0, nL_0]$.

Exercise 12: Reciprocal decay from a quadratic decrease.

Solve the nonlinear recursion that produces the $O(1/t)$ separable-loss rate.

Let $r_0, r_1, \dots$ be positive numbers satisfying

$$r_{t+1} \le r_t - c r_t^2$$

for some $c > 0$, and suppose $c r_0 \le 1/2$.

(a) Prove by induction that r_t is nonincreasing and $c r_t \le 1/2$.

(b) Use $(1-u)^{-1} \ge 1 + u$ for $u \in [0, 1)$ to prove

$$\frac{1}{r_{t+1}} \ge \frac{1}{r_t} + c.$$

(c) Deduce the explicit bound

$$r_t \le \frac{r_0}{1 + c r_0 t}.$$

State the values of r_t and c obtained by combining smooth descent with $\|\nabla L_n(W_t)\|_F \ge a L_n(W_t)$.

HINTS Factor the recursion as $r_{t+1} \le r_t(1 - c r_t)$ before taking reciprocals.

Exercise 13: Why separated softmax has no uniform PL constant.

Rule out uniform PL by comparing the softmax gradient with the vanishing loss.

Fix $x_i \in \mathbb{R}^p$, $y_i \in [K]$, and $W \in \mathbb{R}^{K \times p}$, and define

$$q_W(k \mid x_i) = \frac{e^{w_k^\top x_i}}{\sum_{j=1}^{K} e^{w_j^\top x_i}}, \qquad L_n(W) = \frac{1}{n} \sum_{i=1}^{n} -\log q_W(y_i \mid x_i).$$

Let $R = \max_i \|x_i\|_2$, $q_i = q_W(\cdot \mid x_i)$, and $\ell_i = -\log q_W(y_i \mid x_i)$. You may use

$$\nabla L_n(W) = \frac{1}{n} \sum_{i=1}^{n} (q_i - e_{y_i}) x_i^\top.$$

(a) Prove

$$\|q_i - e_{y_i}\|_2 \le \sqrt{2}\big(1 - q_W(y_i \mid x_i)\big) \le \sqrt{2}\ell_i.$$

(b) Deduce

$$\|\nabla L_n(W)\|_F \le \sqrt{2} R L_n(W).$$

(c) Suppose $W_r = r\overline{W}$ is a separating ray, so $L_n(W_r) \to 0 = L_n^\star$. Prove that no $\mu > 0$ can satisfy

$$\frac{1}{2}\|\nabla L_n(W)\|_F^2 \ge \mu\big(L_n(W) - L_n^\star\big)$$

for every W.

HINTS If $s = 1 - q_W(y_i \mid x_i)$, the incorrect-class probabilities are nonnegative and sum to s. For (c), divide by the positive loss along the ray.

Stochastic and incremental updates

Exercise 14: A minibatch recursion under PL.

Separate contraction from the additive variance term in stochastic gradient descent.

Let $L : \mathbb{R}^d \to \mathbb{R}$ be β-smooth and μ-PL. Random iterates satisfy $\Theta_{t+1} = \Theta_t - \eta G_t$, where

$$\mathbb{E}[G_t \mid \Theta_t] = \nabla L(\Theta_t), \qquad \mathbb{E}[\|G_t - \nabla L(\Theta_t)\|_2^2 \mid \Theta_t] \le \frac{v}{B}.$$

Assume $0 < \eta \le 1/\beta$.

(a) Prove

$$\mathbb{E}[\|G_t\|_2^2 \mid \Theta_t] \le \|\nabla L(\Theta_t)\|_2^2 + \frac{v}{B}.$$

(b) Apply smoothness and PL to derive

$$\mathbb{E}[L(\Theta_{t+1}) - L^\star \mid \Theta_t] \le (1 - \eta\mu)(L(\Theta_t) - L^\star) + \frac{\beta\eta^2 v}{2B}.$$

(c) Unroll the recursion and give an upper bound valid for every t. Explain how its nondecaying term scales with η and B.

HINTS For (a), expand $G_t = \nabla L(\Theta_t) + (G_t - \nabla L(\Theta_t))$ and use the conditional mean. For (c), sum a finite geometric series.

Exercise 15: Stochastic aligned directions.

Extend the PL recursion from unbiased stochastic gradients to conditionally aligned random directions.

Let $L : \mathbb{R}^d \to \mathbb{R}$ be β-smooth and μ-PL, and consider

$$\Theta_{t+1} = \Theta_t + \eta D_t.$$

Suppose that, conditional on Θ_t, constants $c, C > 0$ and $\sigma^2 \ge 0$ satisfy

$$-\mathbb{E}[\langle \nabla L(\Theta_t), D_t\rangle \mid \Theta_t] \ge c\|\nabla L(\Theta_t)\|_2^2, \qquad \mathbb{E}[\|D_t\|_2^2 \mid \Theta_t] \le C^2\|\nabla L(\Theta_t)\|_2^2 + \frac{\sigma^2}{B}.$$

Assume

$$0 < \eta \le \min\left\{\frac{c}{\beta C^2}, \frac{1}{c\mu}\right\}.$$

(a) Prove the conditional recursion

$$\mathbb{E}[L(\Theta_{t+1}) - L^\star \mid \Theta_t] \le (1 - \eta c\mu)(L(\Theta_t) - L^\star) + \frac{\beta\eta^2\sigma^2}{2B}.$$

(b) With $\delta_t = \mathbb{E}[L(\Theta_t) - L^\star]$, prove

$$\delta_t \le (1 - \eta c\mu)^t \delta_0 + \frac{\beta\eta\sigma^2}{2Bc\mu}.$$

(c) Suppose $D_t = -G_t$, where

$$\mathbb{E}[G_t \mid \Theta_t] = \nabla L(\Theta_t), \qquad \mathbb{E}[\|G_t - \nabla L(\Theta_t)\|_2^2 \mid \Theta_t] \le \frac{v}{B}.$$

Identify valid values of c, C, σ^2 and show that part (a) becomes

$$\mathbb{E}[L(\Theta_{t+1}) - L^\star \mid \Theta_t] \le (1 - \eta\mu)(L(\Theta_t) - L^\star) + \frac{\beta\eta^2 v}{2B}.$$

HINTS For (a), apply smoothness before taking the conditional expectation. For (b), sum a finite geometric series. For (c), use the conditional second-moment identity obtained by expanding $G_t = \nabla L(\Theta_t) + (G_t - \nabla L(\Theta_t))$.

Exercise 16: Removing the stochastic noise floor.

Turn the stochastic PL recursion into an explicit $O(1/t)$ bound using a decreasing stepsize.

Let $\delta_t \ge 0$ satisfy

$$\delta_{t+1} \le (1 - \mu\eta_t)\delta_t + \frac{\beta v}{2B}\eta_t^2.$$

Choose

$$\eta_t = \frac{2}{\mu(t + t_0)}, \qquad t_0 \ge \frac{2\beta}{\mu},$$

and define

$$C_0 = \max\left\{t_0\delta_0, \frac{2\beta v}{B\mu^2}\right\}.$$

(a) Verify that $\eta_t \le 1/\beta$ and rewrite the recursion as

$$\delta_{t+1} \le \left(1 - \frac{2}{t + t_0}\right)\delta_t + \frac{2\beta v}{B\mu^2(t + t_0)^2}.$$

(b) Prove by induction that

$$\delta_t \le \frac{C_0}{t + t_0} \qquad \text{for every } t \ge 0.$$

(c) State why this bound has no fixed noise floor, in contrast to the constant-stepsize bound, whose nondecaying term is $\beta\eta v/(2\mu B)$.

HINTS For (b), put $s = t + t_0$ and use $2C_0 - 2\beta v/(B\mu^2) \ge C_0$.

Root finding and Lyapunov functions

Exercise 17: Root finding without an objective.[*]

Separate the residual used by an algorithm from the Lyapunov function used to prove its convergence.

Let $g : \mathbb{R}^d \to \mathbb{R}^d$ be μ-strongly monotone and C-Lipschitz:

$$\langle g(\theta) - g(\theta'), \theta - \theta'\rangle \ge \mu\|\theta - \theta'\|_2^2, \qquad \|g(\theta) - g(\theta')\|_2 \le C\|\theta - \theta'\|_2.$$

Suppose $g(\theta^\star) = 0$, and consider

$$\theta_{t+1} = \theta_t - \eta g(\theta_t).$$

(a) Prove that the root $\theta^\star$ is unique and that

$$\|\theta_{t+1} - \theta^\star\|_2^2 \le (1 - 2\eta\mu + \eta^2 C^2)\|\theta_t - \theta^\star\|_2^2.$$

(b) Prove that $\mu \le C$ and that $0 < \eta < 2\mu/C^2$ makes the factor in part (a) belong to $[0, 1)$. State the resulting geometric bound.

(c) Define $V(\theta) = \|\theta - \theta^\star\|_2^2/2$ and $d(\theta) = -g(\theta)$. Verify the two aligned-direction conditions for V, with constants $c = \mu$ and C. Using the facts that V is 1-smooth and 1-PL, state the stepsize and rate supplied by the aligned-update theorem. Compare its stepsize range with part (b).

(d) Explain why the update can be run without evaluating either $\theta^\star$ or $V(\theta_t)$, and why V is an analysis certificate rather than an objective from which the update was derived.

Hints For (a), subtract $\theta^\star$ from the update and expand the squared norm. For (b), combine strong monotonicity with Cauchy–Schwarz and Lipschitz continuity. For (c), compute $\nabla V(\theta)$.

Exercise 18: A fixed-point residual and its Lyapunov function. *

Relate a computable fixed-point residual to distance from the unknown fixed point.

Let $\mathcal{T} : \mathbb{R}^d \to \mathbb{R}^d$ satisfy

$$\|\mathcal{T}(\theta) - \mathcal{T}(\theta')\|_2 \le q\|\theta - \theta'\|_2 \qquad \text{for all } \theta, \theta' \in \mathbb{R}^d,$$

where $q \in [0, 1)$. Starting from any θ_0, define $\theta_{t+1} = \mathcal{T}(\theta_t)$.

(a) Prove that, for every $t \ge 0$ and $k \ge 1$,

$$\|\theta_{t+k} - \theta_t\|_2 \le \frac{q^t}{1-q}\|\theta_1 - \theta_0\|_2.$$

Deduce that the iterates converge to a fixed point $\theta^\star$, and prove that this fixed point is unique.

(b) Define the fixed-point residual $g(\theta) = \theta - \mathcal{T}(\theta)$. Prove that its roots are exactly the fixed points of $\mathcal{T}$ and that

$$(1-q)\|\theta - \theta^\star\|_2 \le \|g(\theta)\|_2 \le (1+q)\|\theta - \theta^\star\|_2.$$

(c) For $V(\theta) = \|\theta - \theta^\star\|_2^2/2$, prove

$$V(\theta_t) \le q^{2t} V(\theta_0), \qquad \|\theta_t - \theta^\star\|_2 \le \frac{\|g(\theta_t)\|_2}{1-q}.$$

(d) State which of $g(\theta_t)$ and $V(\theta_t)$ can be evaluated without knowing $\theta^\star$. Explain why the second display in part (c) makes the computable residual a stopping certificate for parameter error.

Hints For (a), first bound consecutive differences and then sum a geometric series. The contraction property implies continuity, so it can be applied after taking the limit. For (b), write $g(\theta) = (\theta - \theta^\star) - (\mathcal{T}(\theta) - \mathcal{T}(\theta^\star))$.

Exercise 19: A weighted Lyapunov function for a nonsymmetric update.[*]

Construct an analysis certificate for a linear root-finding update that is not gradient descent.

Let $A \in \mathbb{R}^{d\times d}$ be positive stable and nonsymmetric, let $b \in \mathbb{R}^d$, and consider

$$w_{t+1} = w_t + \eta(b - Aw_t).$$

Fix $Q \succ 0$, and suppose $M \succ 0$ solves $A^\top M + MA = Q$. Define $w^\star = A^{-1}b$ and the candidate Lyapunov function

$$V(w) = (w - w^\star)^\top M(w - w^\star).$$

Thus $V(w) \geq 0$, with equality exactly at $w = w^\star$. Write $e_t = w_t - w^\star$ and $V_t = V(w_t) = e_t^\top M e_t$.

(a) Compute $\nabla V(w)$ and prove that the update direction $d(w) = b - Aw$ satisfies

$$-\langle \nabla V(w), d(w)\rangle = (w - w^\star)^\top Q(w - w^\star).$$

Thus the update is aligned with $-\nabla V(w)$ whenever $w \neq w^\star$.

(b) Prove

$$V_{t+1} - V_t = -\eta e_t^\top Q e_t + \eta^2 e_t^\top A^\top MA e_t.$$

Let $q = \lambda_{\min}(Q)$ and $c_A = \lambda_{\max}(A^\top MA)$. Show that the condition

$$0 < \eta \leq \min\left\{\frac{q}{2c_A}, \frac{\lambda_{\max}(M)}{q}\right\}$$

implies

$$V_{t+1} \leq V_t - \frac{\eta q}{2}\|e_t\|_2^2,$$

and use $V_t \leq \lambda_{\max}(M)\|e_t\|_2^2$ to derive a geometric bound with contraction factor in $[0, 1)$.

(c) Explain why nonsymmetry of A prevents the vector field $b - Aw$ from being the negative gradient of a scalar objective. Also explain why the algorithm can use $b - Aw_t$ without knowing $w^\star$ or evaluating $V(w_t)$, even though both appear in its convergence proof.

Hints For (a), use $\nabla V(w) = 2M(w - w^\star)$ and the Lyapunov equation. For (b), use $e_{t+1} = (I - \eta A)e_t$ and upper-bound the quadratic form involving $A^\top MA$. For (c), recall that the Jacobian of a gradient field is a symmetric Hessian.

References

Adelin Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984. https://doi.org/10.1093/biomet/71.1.1.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning*, pages 242–252, 2019. https://proceedings.mlr.press/v97/allen-zhu19a.html.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3–4):231–357, 2015. https://doi.org/10.1561/2200000050.

Simon S. Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019. https://arxiv.org/abs/1810.02054.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31*, 2018. https://papers.nips.cc/paper_files/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, pages 795–811, 2016. https://arxiv.org/abs/1608.04636.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. Proceedings of the 37th International Conference on Machine Learning, 2020. https://proceedings.mlr.press/v119/mei20b.html.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018. https://www.jmlr.org/papers/v19/18-188.html.

John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997. https://doi.org/10.1109/9.580874.