Deep Learning and Reinforcement Learning: Lecture notes

Lectures 1–2 with Probability Appendix

Contents

Lecture 1

Introduction

1.1 A motivating regression example

Suppose we observe a sample

$$S = \{(X_i, Y_i)\}_{i=1}^n. \tag{1.1.1}$$

A useful motivational model is

$$Y = f^\dagger(X) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid X] = 0. \tag{1.1.2}$$

Here $f^\dagger$ is an unknown signal and ε represents intrinsic randomness or unobserved variation. Gaussian noise is one example, not a universal assumption. Most of the theory below does not require the additive model (1.1.2).

For a predictor $f : \mathcal{X} \to \mathbb{R}$, the empirical squared loss is

$$\widehat{L}_n(f) := \frac{1}{2n} \sum_{i=1}^n \big(Y_i - f(X_i)\big)^2. \tag{1.1.3}$$

The linear class is

$$\mathcal{F}_{\text{lin}} := \{f_\theta : x \mapsto \theta^\top x,\ \theta \in \mathbb{R}^d\}. \tag{1.1.4}$$

Choosing this class makes a representation assumption: the signal is being approximated by a linear function. An affine predictor $x \mapsto \theta^\top x + b$ includes an intercept b. This conventional *bias parameter* is unrelated to the statistical bias of an estimator defined later in this chapter.

Generalized linear classes. For a fixed response map $\psi : \mathbb{R} \to \mathbb{R}$, consider

$$\mathcal{F}_{\text{glm}} := \{f_\theta : x \mapsto \psi(\theta^\top x),\ \theta \in \mathbb{R}^d\}. \tag{1.1.5}$$

The logistic map $\psi(u) = (1 + e^{-u})^{-1}$ produces probabilities for a binary response, while $\psi(u) = e^u$ produces a positive conditional mean. A full generalized linear model pairs the response map with an observation model and a corresponding loss. The predictor is nonlinear in x after ψ, but it still depends on x only through the one-dimensional score $\theta^\top x$.

Neural-network classes. With the ReLU activation $\rho(u) := \max\{u, 0\}$, a one-hidden-layer class is

$$\mathcal{F}_{\text{NN}}(m) := \left\{ x \mapsto a_0 + \sum_{j=1}^m a_j \rho(w_j^\top x + b_j) : {\scriptstyle a_0, a_j, b_j \in \mathbb{R},\ w_j \in \mathbb{R}^d} \right\}. \tag{1.1.6}$$

Unlike the linear and generalized linear classes above, which use the single score $\theta^\top x$ with a fixed response map, a neural network learns the hidden features $x \mapsto \rho(w_j^\top x + b_j)$ together with their output weights. This flexibility usually makes the training objective nonconvex.

These examples separate several kinds of modeling flexibility. A linear predictor uses the identity response map; a generalized linear class adds a fixed nonlinear response map; and a neural network learns multiple hidden features. Moving to a richer class or increasing m can reduce approximation error, but it also changes the statistical and optimization problems. In every case the observations are the labeled sample S, and performance is judged by prediction loss on a fresh draw rather than only by training loss.

1.2 Population risk, empirical risk, and the empirical measure

Let $Z = (X, Y)$ have an unknown distribution P on $\mathcal{X} \times \mathcal{Y}$. A loss $\ell : \mathcal{U} \times \mathcal{Y} \to [0, \infty)$ evaluates a prediction.

Common choices of loss. The loss must match the prediction space and the response encoding. Common examples include

$$\begin{aligned} \ell_{\mathrm{sq}}(a, y) &:= \frac{1}{2}(a - y)^2, && a, y \in \mathbb{R}, \\ \ell_{0\text{-}1}(a, y) &:= \mathbf{1}\{ya \le 0\}, && a \in \mathbb{R}, \quad y \in \{-1, +1\}, \\ \ell_{\mathrm{bce}}(p, y) &:= -y \log p - (1 - y) \log(1 - p), && p \in (0, 1), \quad y \in \{0, 1\}. \end{aligned}$$

The squared loss is the choice used in (1.1.3); its unrestricted population target is the conditional mean. The 0-1 loss directly counts classification errors but is discontinuous. Binary cross-entropy assesses a predicted probability. For a real-valued score and a signed label, its logistic form is

$$\ell_{\log}(a, y) := \log(1 + e^{-ya}), \qquad a \in \mathbb{R}, \quad y \in \{-1, +1\}.$$

For regression with possible outliers, the absolute loss $\ell_{\mathrm{abs}}(a, y) := |a - y|$ is a robust choice whose unrestricted population target is a conditional median. A differentiable compromise is the Huber loss

$$\ell_{\mathrm{H},\delta}(a, y) := \begin{cases} \frac{1}{2}(a - y)^2, & |a - y| \le \delta, \\ \delta|a - y| - \frac{1}{2}\delta^2, & |a - y| > \delta, \end{cases} \qquad \delta > 0.$$

It is quadratic for small residuals and linear in the tails, so a large residual has less influence than under squared loss. Only the 0-1 loss above is automatically bounded by 1; the other losses generally require additional assumptions or different concentration arguments.

Definition 1.1 (Population and empirical risk)**.** The population risk of $f : \mathcal{X} \to \mathcal{U}$ is

$$L(f) := \mathbb{E}_{Z \sim P}[\ell(f(X), Y)]. \tag{1.2.1}$$

For independent observations $Z_1, \ldots, Z_n \sim P$, the empirical risk is

$$\widehat{L}_n(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i). \tag{1.2.2}$$

Define the empirical measure

$$P_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i}. \tag{1.2.3}$$

Then $L(f) = P\ell_f$ and $\widehat{L}_n(f) = P_n \ell_f$, where $\ell_f(x, y) := \ell(f(x), y)$. Learning replaces an expectation under the unknown P by an expectation under the random P_n.

Quantity	Data role	Interpretation
Training error $\widehat{L}_{\mathrm{tr}}(f)$	Fits ordinary model parameters	An empirical objective, usually optimistically small after fitting
Population or generalization error $L(f)$	Defined by the unknown distribution	Expected loss on a fresh draw; the target quantity
Validation error $\widehat{L}_{\mathrm{val}}(f)$	Selects hyperparameters or stopping time	A selection criterion, not a final unbiased report after repeated use
Test error $\widehat{L}_{\mathrm{test}}(f)$	Used after all choices are frozen	A finite-sample estimate of $L(f)$, with its own uncertainty

1.3 Exact ERM and the three-error decomposition

Given a model class $\mathcal{F}$, empirical risk minimization is the prescription

$$\widehat{f}_n \in \arg\min_{f\in\mathcal{F}} \widehat{L}_n(f). \tag{1.3.1}$$

The operator arg min hides computation. A practical optimizer returns an iterate $\widetilde{f}_n$, not an abstract exact minimizer.

Definition 1.2 (Empirical optimization gap)**.** The empirical optimization gap of $\widetilde{f}_n \in \mathcal{F}$ is

$$\varepsilon_{\mathrm{opt}} := \widehat{L}_n(\widetilde{f}_n) - \inf_{f\in\mathcal{F}} \widehat{L}_n(f). \tag{1.3.2}$$

Equivalently,

$$\widehat{L}_n(\widetilde{f}_n) \le \inf_{f\in\mathcal{F}} \widehat{L}_n(f) + \varepsilon_{\mathrm{opt}}. \tag{1.3.3}$$

The gap measures objective value, not distance between parameters. In an overparameterized problem many parameters may have the same empirical loss, so a small gap alone does not identify which solution the algorithm selects.

For the linear regression example, let $\mathbf{X} \in \mathbb{R}^{n\times d}$ have rows $X_i^\top$ and let $\mathbf{y} = (Y_1, \dots, Y_n)^\top$. Exact ERM satisfies the normal equation

$$\mathbf{X}^\top \mathbf{X}\widehat{\theta} = \mathbf{X}^\top \mathbf{y}. \tag{1.3.4}$$

If $\mathbf{X}^\top\mathbf{X}$ is invertible, this gives a unique closed form; otherwise minimizers may not be unique. Large neural-network objectives do not offer such a formula, which is why the algorithmic output must be kept separate from the abstract ERM prescription.

Let $\mathcal{F}_{\mathrm{all}}$ be a broad benchmark class and let $\mathcal{F} \subseteq \mathcal{F}_{\mathrm{all}}$ be the working model class. Set

$$L^\star := \inf_{f\in\mathcal{F}_{\mathrm{all}}} L(f), \qquad L^\star_{\mathcal{F}} := \inf_{f\in\mathcal{F}} L(f), \qquad \Delta_n(\mathcal{F}) := \sup_{f\in\mathcal{F}} |L(f) - \widehat{L}_n(f)|. \tag{1.3.5}$$

The exact split

$$L(\widetilde{f}_n) - L^\star = \underbrace{L^\star_{\mathcal{F}} - L^\star}_{\text{approximation}} + \underbrace{L(\widetilde{f}_n) - L^\star_{\mathcal{F}}}_{\text{within-class excess risk}} \tag{1.3.6}$$

separates the price of restricting the representation from everything that happens within the class. Uniform comparison of L and $\widehat{L}_n$ then separates sampling from computation.

Theorem 1.3 (Master excess-risk bound)**.** *If $\widetilde{f}_n$ satisfies (1.3.3), then*

$$\boxed{L(\widetilde{f}_n) - L^\star \le \underbrace{L^\star_{\mathcal{F}} - L^\star}_{\textit{approximation error}} + \underbrace{2\Delta_n(\mathcal{F})}_{\textit{statistical control}} + \underbrace{\varepsilon_{\mathrm{opt}}}_{\textit{optimization error}}.} \tag{1.3.7}$$

Proof. For every comparator $f \in \mathcal{F}$,

$$\begin{aligned} L(\widetilde{f}_n) &\le \widehat{L}_n(\widetilde{f}_n) + \Delta_n(\mathcal{F}) \\ &\le \inf_{g\in\mathcal{F}} \widehat{L}_n(g) + \varepsilon_{\mathrm{opt}} + \Delta_n(\mathcal{F}) \\ &\le \widehat{L}_n(f) + \varepsilon_{\mathrm{opt}} + \Delta_n(\mathcal{F}) \\ &\le L(f) + \varepsilon_{\mathrm{opt}} + 2\Delta_n(\mathcal{F}). \end{aligned}$$

Subtract $L^\star$ and take the infimum over $f \in \mathcal{F}$. □

1.4 Capacity, underfitting, and overfitting

Capacity describes the variety of functions that a learning procedure can effectively choose among [5, Section 5.2]. It is therefore not synonymous with parameter count: two parameterizations may describe the same functions, while a norm constraint or a training rule may make only a small part of a large parameterized family practically accessible.

Representational capacity and three different risks

Suppose that a complexity index q produces nested classes

$$\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots . \tag{1.4.1}$$

Depending on the application, q might count selected features, polynomial degree, network width or depth, or the radius of a norm constraint.

Concrete example: polynomial degree. For a scalar input, consider

$$\mathcal{F}_q^{\text{poly}} := \left\{ x \mapsto \sum_{j=0}^{q} \theta_j x^j : \theta \in \mathbb{R}^{q+1} \right\}. \tag{1.4.2}$$

If the signal is approximately quadratic, a line may miss systematic curvature, while a quadratic can capture the main pattern. At the other extreme, with n distinct training inputs, a polynomial of degree at most $n-1$ can interpolate arbitrary observed responses.

The construction in Figure 1.1 follows the book's example: six noiseless observations lie exactly on a quadratic signal. A line cannot express its curvature. The quadratic class contains the signal. A degree-nine class contains the signal too, but its ten coefficients are underdetermined by six interpolation constraints, so many functions achieve zero training loss. A tie-breaking rule such as a Moore–Penrose solution can select an interpolant whose behavior between observations is very different. Thus sparse sampling alone can permit overfitting; label noise would make the problem still more visible.

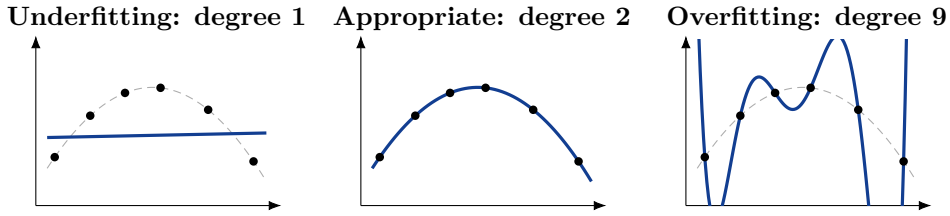

Figure 1.1: Three fits to the same sparse, noiseless quadratic sample. The gray dashed curve is the signal and the blue curve is the fitted predictor. The right panel shows one possible degree-nine interpolant: every observed point is fitted exactly, yet the behavior between points is poorly constrained. Adapted from [5, Figure 5.2].

The example makes capacity visible, but it does not make degree or parameter count a universal measure. Architecture, parameter sharing, norms, and the training algorithm all change which fitted functions are easy to obtain.

Returning to the general nested-class setup, three quantities must be kept separate:

$$\widehat{L}^{\star}_{n,q} := \inf_{f \in \mathcal{F}_q} \widehat{L}_n(f), \qquad L^{\star}_{\mathcal{F}_q} := \inf_{f \in \mathcal{F}_q} L(f), \qquad \widehat{f}_{q,S} := \mathsf{A}_q(S). \tag{1.4.3}$$

The first is the best training risk available in the class, the second is the best population risk available in the class, and the third is what a specified algorithm actually returns from sample S. Nesting gives the deterministic monotonicity relations

$$\widehat{L}^{\star}_{n,q+1} \le \widehat{L}^{\star}_{n,q}, \qquad L^{\star}_{\mathcal{F}_{q+1}} \le L^{\star}_{\mathcal{F}_q}. \tag{1.4.4}$$

In contrast, $L(\widehat{f}_{q,S})$ need not decrease with q. The fitted predictor depends on a finite sample, initialization, optimization, and any regularization or stopping rule. A larger class creates more opportunities to represent useful structure, but also more opportunities for the procedure to respond to sample-specific variation.

In particular, nesting does not produce a U-shaped population-risk curve: both class-oracle quantities in (1.4.4) decrease.

Bias–variance trade-off under squared loss

Let $S \sim P^n$ be the training sample and let $\widehat{f}_{q,S} := \mathsf{A}_q(S)$ be the predictor returned at capacity q. We will use the repeated-sample averages

$$T(q) := \mathbb{E}_S\left[\widehat{L}_n(\widehat{f}_{q,S})\right], \qquad R(q) := \mathbb{E}_S\left[L(\widehat{f}_{q,S})\right]. \tag{1.4.5}$$

For the exact additive explanation under squared loss, let $m(x) := \mathbb{E}[Y \mid X = x]$, let the evaluation pair (X, Y) be independent of S, and set $\overline{f}_q(x) := \mathbb{E}_S[\widehat{f}_{q,S}(x)]$. At a fixed input x, the estimator bias and variance are

$$\mathrm{Bias}_S\{\widehat{f}_{q,S}(x)\} := \overline{f}_q(x) - m(x), \qquad \mathrm{Var}_S\{\widehat{f}_{q,S}(x)\} := \mathbb{E}_S\left[\{\widehat{f}_{q,S}(x) - \overline{f}_q(x)\}^2\right].$$

They are properties of the learning procedure across hypothetical repetitions of the complete training sample, not quantities that can be read from one fitted model. Because this chapter uses $\ell(a, y) = \frac{1}{2}(a - y)^2$,

$$R(q) = \underbrace{\frac{1}{2}\mathbb{E}_X[\mathrm{Var}(Y \mid X)]}_{\text{irreducible noise}} + \underbrace{\frac{1}{2}\left\|\overline{f}_q - m\right\|^2_{L^2(P_X)}}_{\text{squared estimator bias}} + \underbrace{\frac{1}{2}\mathbb{E}_X\left[\mathrm{Var}_S\{\widehat{f}_{q,S}(X)\}\right]}_{\text{estimator variance}}. \tag{1.4.6}$$

To verify the identity, write

$$Y - \widehat{f}_{q,S}(X) = \{Y - m(X)\} + \{m(X) - \overline{f}_q(X)\} + \{\overline{f}_q(X) - \widehat{f}_{q,S}(X)\}.$$

The cross terms vanish after conditioning on X and averaging over the independent training sample. In the motivating additive model, $m = f^\dagger$.

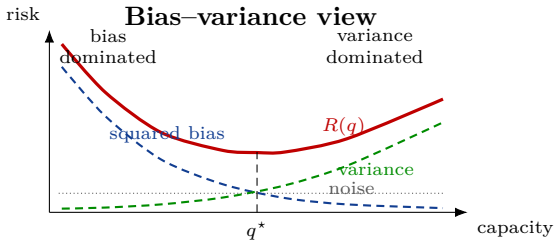

Figure 1.2: Bias–variance view of the finite-sample capacity trade-off under squared loss, adapted from [5, Figure 5.6]. The red curve is the repeated-sample population risk $R(q)$ in (1.4.5), decomposed by (1.4.6). The U shape is a typical schematic, not a universal law.

Diagnosing underfitting and overfitting

Underfitting means that the fitted predictor fails to capture useful, transferable structure. High training and independent-validation losses are a common symptom. The cause may be a restrictive class, incomplete optimization, excessive regularization, inappropriate features, or a poorly matched loss. *Overfitting* means that the fitted predictor uses sample-specific variation in a way that does not transfer. Low training loss together with a substantially larger independent-validation loss is a common symptom. Low training loss by itself is not evidence of overfitting. Likewise, the mnemonics *underfitting means high bias* and *overfitting means high variance* describe common mechanisms, not definitions. Bias and variance average over hypothetical repetitions of the whole training sample; one train–validation split can reveal symptoms but cannot identify either quantity.

Observed pattern	Plausible diagnosis	Useful next comparison
Training and validation losses are both high	The representation or optimization may be too restrictive	Increase class size, train longer, or weaken regularization, changing one factor at a time.
Training loss is low, but validation loss is much higher	The procedure may be sensitive to the sample; leakage or excessive tuning is also possible	Add data or regularization, reduce effective capacity, and audit the split and preprocessing pipeline.
Training and validation losses are both low and close	No obvious underfitting or overfitting signal	Freeze modeling decisions and use the untouched test set once for final estimation.

Learning curves add an important axis to this diagnosis. If the training– validation gap shrinks as the sample size grows, finite-sample variability is a plausible limitation. If both curves plateau at an unsatisfactory level, representation, optimization, regularization, or the chosen loss may be the bottleneck. These patterns are clues, not logical equivalences: distribution shift, label noise, leakage, and a mismatched evaluation metric can imitate the same symptoms.

Hyperparameters, validation, and regularization

Ordinary parameters are fitted inside a chosen training procedure. *Hyperparameters* specify that procedure: examples include class size q, network architecture, a regularization level, learning rate, batch size, and stopping time. Capacity-controlling hyperparameters cannot be chosen by minimizing training error alone, because that criterion systematically favors more fit and less regularization.

A clean protocol is nested:

1. use the training set to produce a candidate $\widehat{f}_\lambda$ for each $\lambda \in \Lambda$;
2. use independent validation data to choose $\widehat{\lambda}$;
3. freeze every choice and use an untouched test set once to estimate $L(\widehat{f}_{\widehat{\lambda}})$.

Validation is itself a model-selection problem, so its reported minimum is optimistic after search. If the test set is inspected repeatedly, it has quietly become another validation set. There is no universal 80/20 split: the allocation balances fitting accuracy against selection and evaluation uncertainty. A standard finite-candidate concentration bound can quantify this selection effect.

Regularization expresses a preference among solutions. A constrained version uses

$$\mathcal{F}_R := \{f \in \mathcal{F} : \Omega(f) \le R\},$$

whereas a penalized procedure minimizes

$$\widehat{L}_n(f) + \lambda\Omega(f).$$

The second display defines a different empirical objective for every λ. Its optimization gap must be measured relative to that penalized objective, not silently substituted for ε_{opt} in the unregularized master bound. Regularization can change approximation bias, sampling variability, and optimization geometry at the same time; it is not a fourth independent error term.

Ridge regression as a capacity-control example. For clarity, omit the intercept; in practice it is usually left unpenalized, and all centering constants must be learned from the training set only. For $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\mathbf{y} \in \mathbb{R}^n$, and $\lambda \ge 0$, ridge regression solves

$$\widehat{\theta}_\lambda \in \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\theta\|^2 + \frac{\lambda}{2} \|\theta\|^2 \right\}.$$

For $\lambda > 0$, the first-order condition gives

$$\widehat{\theta}_\lambda = (\mathbf{X}^\top\mathbf{X} + n\lambda I_d)^{-1}\mathbf{X}^\top\mathbf{y}.$$

The factor $n\lambda$ appears because the empirical squared loss is divided by n. At $\lambda = 0$, take the minimum-norm least-squares solution $\widehat{\theta}_0 = \mathbf{X}^\dagger\mathbf{y}$; it is the limit of $\widehat{\theta}_\lambda$ as $\lambda \downarrow 0$.

Let $\mathbf{X} = \mathbf{U}_r \operatorname{diag}(s_1, \dots, s_r)\mathbf{V}_r^\top$ be a thin singular-value decomposition and define

$$a_j(\lambda) := \frac{s_j^2}{s_j^2 + n\lambda}.$$

Then

$$\widehat{\theta}_\lambda = \sum_{j=1}^{r} \frac{s_j}{s_j^2 + n\lambda}(\mathbf{u}_j^\top \mathbf{y})\mathbf{v}_j,$$

$$\widehat{\mathbf{y}}_\lambda := \mathbf{X}\widehat{\theta}_\lambda = \sum_{j=1}^{r} a_j(\lambda)(\mathbf{u}_j^\top \mathbf{y})\mathbf{u}_j.$$

Thus ridge shrinks the unregularized fit in direction j by $a_j(\lambda)$. Small-singular-value directions are damped most strongly, and the effective degrees of freedom

$$\operatorname{tr}\big[\mathbf{X}(\mathbf{X}^\top\mathbf{X} + n\lambda I_d)^{-1}\mathbf{X}^\top\big] = \sum_{j=1}^{r} a_j(\lambda)$$

decrease with λ.

The training residual makes clear why training loss cannot choose the amount of regularization:

$$\|\mathbf{y} - \widehat{\mathbf{y}}_\lambda\|^2 = \big\|(I_n - \mathbf{U}_r\mathbf{U}_r^\top)\mathbf{y}\big\|^2 + \sum_{j=1}^{r}\left(\frac{n\lambda}{s_j^2 + n\lambda}\right)^2 (\mathbf{u}_j^\top \mathbf{y})^2.$$

Every λ-dependent factor on the right is nondecreasing, so training squared error favors $\lambda = 0$, up to ties. Minimized penalized-objective values should not be used to select λ: changing λ changes the criterion, so their ordering does not estimate a common prediction risk. Instead, fit each candidate on the training set and choose

$$\widehat{\lambda} \in \arg\min_{\lambda \in \Lambda} \frac{1}{2n_{\text{val}}} \left\|\mathbf{y}_{\text{val}} - \mathbf{X}_{\text{val}}\widehat{\theta}_\lambda^{\text{tr}}\right\|^2.$$

Ridge also makes the coefficient bias–variance mechanism explicit. Conditional on the design, suppose $\mathbf{y} = \mathbf{X}\theta^\star + \boldsymbol{\varepsilon}$, with $\mathbb{E}[\boldsymbol{\varepsilon} \mid \mathbf{X}] = 0$ and $\operatorname{Cov}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \sigma^2 I_n$. For $\alpha_j := \mathbf{v}_j^\top \theta^\star$,

$$\mathbf{v}_j^\top \widehat{\theta}_\lambda = a_j(\lambda)\alpha_j + \frac{s_j}{s_j^2 + n\lambda}\mathbf{u}_j^\top \boldsymbol{\varepsilon},$$

so the coefficient in direction j has

$$\operatorname{Bias}_j^2(\lambda) = \{1 - a_j(\lambda)\}^2\alpha_j^2, \qquad \operatorname{Var}_j(\lambda) = \frac{\sigma^2 s_j^2}{(s_j^2 + n\lambda)^2}.$$

As λ increases, squared shrinkage bias is nondecreasing and sampling variance is nonincreasing in each observable direction. For isotropic future covariates, summing these terms, together with any unidentifiable null-space bias, gives the prediction bias and variance; a general test covariance weights and couples directions differently. The best trade-off therefore depends on the unknown signal, noise, and future covariate distribution, which is precisely why λ is selected by validation. It also previews why ridge can smooth the interpolation peak discussed next.

Remark 1.4 (k-fold cross-validation)**.** When data are scarce, split the sample into k folds, train k times, and evaluate each observation using a model that did not train on its fold. This uses data efficiently but costs k fits. Because the fitted training sets overlap, fold errors are dependent, so a naive independent-fold standard error is not justified. If cross-validation both selects hyperparameters and reports final performance, an outer evaluation layer or untouched test set is needed.

1.5 Double descent

Figure 1.2 shows the classical picture obtained when capacity is increased through an underfitting regime and into an overfitting regime. If the horizontal axis is continued until the learning procedure first achieves zero training loss, and then continued farther, a second pattern can appear. Write q_{int} for this *interpolation threshold*. The expected population risk may rise sharply near q_{int} and then decrease again in a more strongly overparameterized regime. The resulting curve is called *double descent* [1].

Throughout this subsection, $n := |S|$ is the number of training examples. We hold n fixed while increasing the general capacity index q. In the linear model below we set $q = d$, where d is the number of input coordinates retained by the learner and, because no intercept is fitted, also the number of unknown regression coefficients. The aspect ratio $\gamma := d/n$ distinguishes the underparameterized regime $d < n$, the critical regime $d \approx n$, and the overparameterized regime $d > n$.

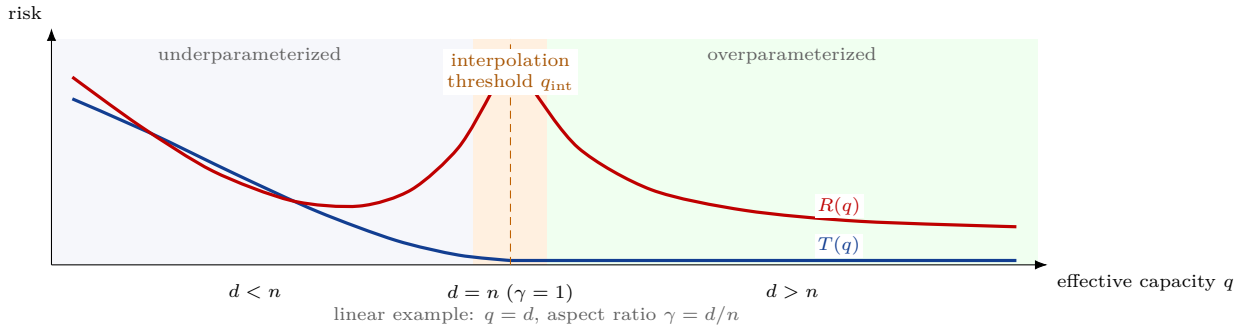

Figure 1.3: Schematic model-wise double descent, adapted from [1]. The quantities $T(q)$ and $R(q)$ are the repeated-sample averages defined in (1.4.5). Training risk first reaches zero at the interpolation threshold. Population risk can peak there and then fall as the procedure moves farther into an overparameterized regime. The shape is qualitative, not a theorem for every nested class; the labels involving d specialize the generic capacity axis to the linear model below.

A Gaussian linear model where the peak is explicit. Let $X = (X_1, X_2, \ldots)$ have independent standard Gaussian coordinates and consider the fixed regression function

$$m(X) = \sum_{j \geq 1} \beta_j X_j, \qquad Y = m(X) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2), \qquad \varepsilon \perp\!\!\!\perp X,$$

with $\sum_{j \geq 1} \beta_j^2 < \infty$. At capacity d, the learner uses only $X_1, \ldots, X_d$ and minimizes squared loss. For a training sample $S = \{(X_i, Y_i)\}_{i=1}^n$, let $\mathbf{X}_d \in \mathbb{R}^{n \times d}$ have row $X_{i,1:d}^\top$, and let $\mathbf{y} = (Y_1, \ldots, Y_n)^\top$. Thus $\mathbf{X}_d$ has one row per training example and one column per retained feature.

Solving the three least-squares regimes. For a coefficient vector $b \in \mathbb{R}^d$, the empirical objective, its gradient, and the normal equation are

$$\widehat{L}_{n,d}(b) = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}_d b\|_2^2, \qquad \nabla \widehat{L}_{n,d}(b) = \frac{1}{n} \mathbf{X}_d^\top (\mathbf{X}_d b - \mathbf{y}), \qquad \mathbf{X}_d^\top \mathbf{X}_d b = \mathbf{X}_d^\top \mathbf{y}. \tag{1.5.1}$$

A Gaussian n-by-d design has rank $\min\{n, d\}$ almost surely. Consequently, the Moore–Penrose solution has the three explicit forms

$$\widehat{\beta}_d := \mathbf{X}_d^\dagger \mathbf{y} = \begin{cases} (\mathbf{X}_d^\top \mathbf{X}_d)^{-1} \mathbf{X}_d^\top \mathbf{y}, & d < n, \\ \mathbf{X}_d^{-1} \mathbf{y}, & d = n, \\ \mathbf{X}_d^\top (\mathbf{X}_d \mathbf{X}_d^\top)^{-1} \mathbf{y}, & d > n, \end{cases} \qquad \widehat{f}_d(x) = x_{1:d}^\top \widehat{\beta}_d. \tag{1.5.2}$$

For $d < n$, $\mathbf{X}_d^\top \mathbf{X}_d$ is invertible, so the normal equation has the unique ordinary-least-squares solution in the first line. At $d = n$, the square matrix $\mathbf{X}_d$ is invertible and the unique solution fits $\mathbf{X}_d \widehat{\beta}_d = \mathbf{y}$. For $d > n$, the matrix has full row rank. Set

$$\beta_0 := \mathbf{X}_d^\top (\mathbf{X}_d \mathbf{X}_d^\top)^{-1} \mathbf{y} = \mathbf{X}_d^\dagger \mathbf{y}.$$

This is one interpolator because

$$\mathbf{X}_d \beta_0 = \mathbf{X}_d \mathbf{X}_d^\top (\mathbf{X}_d \mathbf{X}_d^\top)^{-1} \mathbf{y} = \mathbf{y}.$$

Consequently, if $v \in \ker(\mathbf{X}_d)$, then $\mathbf{X}_d(\beta_0 + v) = \mathbf{y}$, so every $\beta_0 + v$ is an interpolator. Conversely, if b is any interpolator, then $\mathbf{X}_d(b - \beta_0) = \mathbf{y} - \mathbf{y} = 0$, so $b - \beta_0 \in \ker(\mathbf{X}_d)$. The two inclusions prove that the complete solution set is

$$\{b \in \mathbb{R}^d : \mathbf{X}_d b = \mathbf{y}\} = \{\beta_0 + v : v \in \ker(\mathbf{X}_d)\} = \left\{\mathbf{X}_d^\top (\mathbf{X}_d \mathbf{X}_d^\top)^{-1} \mathbf{y} + v : v \in \ker(\mathbf{X}_d)\right\}. \tag{1.5.3}$$

Moreover, $\beta_0 \in \text{range}(\mathbf{X}_d^\top)$, while the fundamental theorem of linear algebra gives $\text{range}(\mathbf{X}_d^\top) = \ker(\mathbf{X}_d)^\perp$. Hence $\beta_0 \perp v$ and

$$\|\beta_0 + v\|_2^2 = \|\beta_0\|_2^2 + \|v\|_2^2.$$

Thus the norm is minimized uniquely at $v = 0$, proving that $\mathbf{X}_d^\dagger \mathbf{y}$ is the minimum-norm interpolator. It is also the solution selected by gradient descent from zero for a suitable fixed step size; the zero initialization is essential for this implicit selection.

The distinction is visible directly in the fitted values and training loss. For $d < n$, define the hat matrix $\mathbf{H}_d := \mathbf{X}_d(\mathbf{X}_d^\top \mathbf{X}_d)^{-1} \mathbf{X}_d^\top$. Then

$$\mathbf{X}_d \widehat{\beta}_d = \begin{cases} \mathbf{H}_d \mathbf{y}, & d < n, \\ \mathbf{y}, & d \geq n, \end{cases} \qquad \widehat{L}_{n,d}(\widehat{\beta}_d) = \begin{cases} \dfrac{1}{2n} \|(\mathbf{I}_n - \mathbf{H}_d)\mathbf{y}\|_2^2, & d < n, \\ 0, & d \geq n. \end{cases} \tag{1.5.4}$$

With nonzero continuous observation or omitted-coordinate noise, the residual in the first line is nonzero almost surely. Thus $d = n$ is the first dimension at which arbitrary training responses can be interpolated. For this specialization $q = d$, the interpolation threshold is $q_{\text{int}} = n$, equivalently $\gamma = 1$.

Define the omitted-signal energy, included-signal energy, and effective noise by

$$a_d := \sum_{j>d} \beta_j^2, \qquad B_d^2 := \sum_{j \leq d} \beta_j^2, \qquad \tau_d^2 := \sigma^2 + a_d.$$

The omitted coordinates are independent of the fitted coordinates and act like additional Gaussian training noise. More precisely,

$$\mathbf{y} = \mathbf{X}_d \beta_{1:d} + \boldsymbol{\eta}_d, \qquad \boldsymbol{\eta}_d \sim \mathcal{N}(0, \tau_d^2 \mathbf{I}_n), \qquad \boldsymbol{\eta}_d \perp\!\!\!\perp \mathbf{X}_d.$$

For $d < n$, equation (1.5.4) and $(\mathbf{I}_n - \mathbf{H}_d)\mathbf{X}_d = 0$ give the following calculation:

$$\mathbb{E}\left[\|(\mathbf{I}_n - \mathbf{H}_d)\mathbf{y}\|_2^2 \mid \mathbf{X}_d\right] = \tau_d^2 (n - d).$$

This follows because $\mathbf{H}_d$ is a rank-d orthogonal projector; the residual projector has trace $n - d$. Therefore the repeated-sample training curve is exactly

$$T(d) = \mathbb{E}_S\left[\widehat{L}_{n,d}(\widehat{\beta}_d)\right] = \begin{cases} \dfrac{\tau_d^2}{2}\left(1 - \dfrac{d}{n}\right), & d < n, \\ 0, & d \geq n. \end{cases} \tag{1.5.5}$$

Deriving the test curve. Set $\theta_d := \beta_{1:d} \in \mathbb{R}^d$, so $\|\theta_d\|_2^2 = B_d^2$. For an independent test observation, write

$$Y' = (X'_{1:d})^\top \theta_d + \eta'_d, \qquad \eta'_d := \sum_{j>d} \beta_j X'_j + \varepsilon'.$$

Then $X'_{1:d} \sim \mathcal{N}(0, \mathbf{I}_d)$, $\eta'_d \sim \mathcal{N}(0, \tau_d^2)$, and these variables are independent of each other and of the training sample. With $\Delta_d := \widehat{\beta}_d - \theta_d$, conditioning on the realized training sample gives

$$\mathbb{E}\left[(Y' - \widehat{f}_d(X'))^2 \mid S\right] = \mathbb{E}\left[\left(\eta'_d - (X'_{1:d})^\top \Delta_d\right)^2 \mid S\right]$$

$$= \tau_d^2 + \Delta_d^\top \mathbb{E}[X'_{1:d}(X'_{1:d})^\top]\Delta_d$$
$$= \tau_d^2 + \|\Delta_d\|_2^2.$$

The cross term in the second line is zero because η'_d and $X'_{1:d}$ are independent and centered. Thus it remains to calculate the mean squared coefficient error.

For $d < n$, ordinary least squares gives

$$\Delta_d = (\mathbf{X}_d^\top \mathbf{X}_d)^{-1}\mathbf{X}_d^\top \boldsymbol{\eta}_d.$$

Consequently,

$$\begin{aligned}\mathbb{E}[\|\Delta_d\|_2^2 \mid \mathbf{X}_d] &= \operatorname{tr}\big(\mathbb{E}[\Delta_d\Delta_d^\top \mid \mathbf{X}_d]\big)\\ &= \tau_d^2 \operatorname{tr}\big((\mathbf{X}_d^\top \mathbf{X}_d)^{-1}\mathbf{X}_d^\top \mathbf{X}_d(\mathbf{X}_d^\top \mathbf{X}_d)^{-1}\big)\\ &= \tau_d^2 \operatorname{tr}\big((\mathbf{X}_d^\top \mathbf{X}_d)^{-1}\big).\end{aligned}$$

Here $\mathbf{X}_d^\top \mathbf{X}_d$ is a d-dimensional Wishart matrix with n degrees of freedom. Its inverse first moment is

$$\mathbb{E}[(\mathbf{X}_d^\top \mathbf{X}_d)^{-1}] = \frac{\mathbf{I}_d}{n-d-1}, \qquad n > d+1.$$

It follows that

$$\mathbb{E}[\|\Delta_d\|_2^2] = \tau_d^2 \frac{d}{n-d-1}, \qquad d < n-1.$$

For $d > n$, introduce the d-by-d row-space projector

$$\boldsymbol{\Pi}_d := \mathbf{X}_d^\top (\mathbf{X}_d\mathbf{X}_d^\top)^{-1}\mathbf{X}_d$$

and $\mathbf{A}_d := \mathbf{X}_d^\top (\mathbf{X}_d\mathbf{X}_d^\top)^{-1}$. The minimum-norm solution satisfies

$$\widehat{\beta}_d = \boldsymbol{\Pi}_d\theta_d + \mathbf{A}_d\boldsymbol{\eta}_d, \qquad \Delta_d = -(\mathbf{I}_d - \boldsymbol{\Pi}_d)\theta_d + \mathbf{A}_d\boldsymbol{\eta}_d.$$

The first term lies in $\ker(\mathbf{X}_d)$, whereas the second lies in $\operatorname{range}(\mathbf{X}_d^\top) = \ker(\mathbf{X}_d)^\perp$. They are therefore orthogonal for every realization. Moreover, $\mathbf{A}_d^\top \mathbf{A}_d = (\mathbf{X}_d\mathbf{X}_d^\top)^{-1}$, so

$$\mathbb{E}[\|\Delta_d\|_2^2 \mid \mathbf{X}_d] = \|(\mathbf{I}_d - \boldsymbol{\Pi}_d)\theta_d\|_2^2 + \tau_d^2 \operatorname{tr}\big((\mathbf{X}_d\mathbf{X}_d^\top)^{-1}\big).$$

The Gaussian row space is a uniformly random n-dimensional subspace of $\mathbb{R}^d$. Rotational invariance therefore gives $\mathbb{E}[\boldsymbol{\Pi}_d] = c\mathbf{I}_d$; taking traces and using $\operatorname{tr}(\boldsymbol{\Pi}_d) = n$ gives $c = n/d$. Since $\mathbf{I}_d - \boldsymbol{\Pi}_d$ is itself an orthogonal projector,

$$\mathbb{E}\big[\|(\mathbf{I}_d - \boldsymbol{\Pi}_d)\theta_d\|_2^2\big] = \theta_d^\top\big(\mathbf{I}_d - \mathbb{E}[\boldsymbol{\Pi}_d]\big)\theta_d = \Big(1 - \frac{n}{d}\Big)B_d^2.$$

Also $\mathbf{X}_d\mathbf{X}_d^\top$ is an n-dimensional Wishart matrix with d degrees of freedom, and hence

$$\mathbb{E}[(\mathbf{X}_d\mathbf{X}_d^\top)^{-1}] = \frac{\mathbf{I}_n}{d-n-1}, \qquad d > n+1.$$

Therefore

$$\mathbb{E}[\|\Delta_d\|_2^2] = \Big(1 - \frac{n}{d}\Big)B_d^2 + \tau_d^2 \frac{n}{d-n-1}, \qquad d > n+1.$$

Both Wishart identities follow from one inverse-Gram calculation. If $\mathbf{Z} \in \mathbb{R}^{m\times p}$ has independent standard Gaussian entries and $m > p+1$, then

$$\mathbb{E}[(\mathbf{Z}^\top\mathbf{Z})^{-1}] = \frac{\mathbf{I}_p}{m-p-1}.$$

Applying it to $\mathbf{Z} = \mathbf{X}_d$ and $\mathbf{Z} = \mathbf{X}_d^\top$ yields the two moments used above and explains the strict dimension conditions.

Combining the coefficient errors with the independent test-noise term τ_d^2, the expected test mean-squared error is, away from the critical dimensions,

$$\mathcal{E}_d := \mathbb{E}\Big[(Y' - \widehat{f}_d(X'))^2\Big] = \begin{cases} \tau_d^2\left(1 + \dfrac{d}{n-d-1}\right), & d < n-1, \\ B_d^2\left(1 - \dfrac{n}{d}\right) + \tau_d^2\left(1 + \dfrac{n}{d-n-1}\right), & d > n+1. \end{cases} \tag{1.5.6}$$

Under the loss convention of this chapter, $R(d) = \frac{1}{2}\mathcal{E}_d$. The expectation in (1.5.6) is not finite at $d \in \{n-1, n, n+1\}$ when $\tau_d^2 > 0$: individual fitted risks are finite almost surely, but nearly singular designs give the average a heavy tail [6].

When does this formula actually give double descent? Equation (1.5.6) permits double descent, but does not force it. On the underparameterized branch,

$$\mathcal{E}_d = \tau_d^2 \frac{n-1}{n-d-1}, \qquad \tau_{d+1}^2 = \tau_d^2 - \beta_{d+1}^2.$$

For $d < n-2$, so that both adjacent expectations are finite, direct subtraction gives

$$\begin{aligned} \mathcal{E}_{d+1} - \mathcal{E}_d &= \frac{n-1}{n-d-2}(\tau_d^2 - \beta_{d+1}^2) - \frac{n-1}{n-d-1}\tau_d^2 \\ &= \frac{n-1}{(n-d-2)(n-d-1)}\left[\tau_d^2 - (n-d-1)\beta_{d+1}^2\right]. \end{aligned}$$

Consequently,

$$\mathcal{E}_{d+1} < \mathcal{E}_d \quad \Longleftrightarrow \quad \beta_{d+1}^2 > \frac{\tau_d^2}{n-d-1}.$$

The reverse inequality gives an ascent, and equality gives a flat step. Thus the decrease in τ_d^2 and the increase in $(n-1)/(n-d-1)$ compete; monotonicity of the two factors does not determine their product. For example, if $\beta_{d+1} = 0$ and $\tau_d^2 > 0$, that step strictly increases risk. If all added coordinates carry no signal, the entire finite underparameterized branch increases, so there is no first descent. Irregularly ordered coefficient magnitudes can even create several local turns.

The comparison from $d = n-2$ to $d = n-1$ is excluded from the finite difference above. When $\tau_{n-1}^2 > 0$, the expected risk at $d = n-1$ is infinite. This produces the singular peak around interpolation, but does not imply that the finite branch $d < n-1$ contains both a descent and an ascent.

On the overparameterized branch, write

$$\mathcal{E}_d = B_d^2\left(1 - \frac{n}{d}\right) + \tau_d^2 v_d, \qquad v_d := \frac{d-1}{d-n-1}, \qquad d > n+1.$$

The noise multiplier decreases because

$$v_{d+1} - v_d = \frac{d}{d-n} - \frac{d-1}{d-n-1} = -\frac{n}{(d-n)(d-n-1)} < 0.$$

Since τ_d^2 is also nonincreasing, the effective-noise contribution $\tau_d^2 v_d$ decreases beyond the threshold. In contrast, $B_d^2(1 - n/d)$ is nondecreasing and can offset that reduction. The familiar double-descent shape therefore requires three comparisons to go the right way: informative early coordinates create a first descent, positive effective noise creates the interpolation peak, and post-threshold noise reduction dominates the growing null-space signal term. The formula displays this mechanism; it is not a universal shape theorem.

The three parts of the curve now have distinct causes.

1. As informative coordinates are added, a_d decreases. If that approximation gain initially dominates estimation variability, the first descent appears.
2. Near $d = n$, the smallest singular values of $\mathbf{X}_d$ can be tiny. Least squares divides by those singular values, strongly amplifying label noise and omitted-feature noise. This produces the interpolation peak.
3. For $d > n$, training loss is zero and there are infinitely many interpolators. The minimum-norm selector has noise-amplification term $\tau_d^2 n/(d-n-1)$, which decreases beyond the threshold and can produce the second descent.

Connection to bias and variance. Equation (1.4.6) remains exact. For example, in the overparameterized Gaussian model above, the squared estimator bias and variance contributions to excess prediction MSE are

$$\mathrm{Bias}_d^2 = a_d + \left(1 - \frac{n}{d}\right)^2 B_d^2,$$
$$\mathrm{Var}_d = \frac{n}{d}\left(1 - \frac{n}{d}\right) B_d^2 + \tau_d^2 \frac{n}{d - n - 1}.$$

Their sum is $\mathcal{E}_d - \sigma^2$, up to the factor $1/2$ in our loss. Thus double descent is not a new term outside bias–variance; it is a possible nonmonotone shape of their sum. In particular, variance can peak near interpolation and then decrease.

Related literatures. The mechanism is analytically visible in ridgeless or minimum-norm linear regression with random design [6]. Random-feature regression, equivalently a two-layer network whose random first-layer weights are frozen and whose output layer is fit by least squares, provides a nonlinear model with a precise double-descent analysis [8]. Model-wise and epoch-wise double descent have also been observed empirically for CNNs, ResNets, and Transformers trained close to interpolation [10]. These examples are not a sufficient condition: the peak depends on the capacity path, noise, regularization, early stopping, and the optimizer's implicit choice among interpolators. Ridge regularization or early stopping can substantially smooth or remove it.

Lecture 2

First-order optimization

The master decomposition in Theorem 1.3 shows that good population performance requires control of three distinct quantities:

$$L(\widetilde{f}_n) - L^\star \le \underbrace{L^\star_{\mathcal{F}} - L^\star}_{\text{approximation}} + \underbrace{2\Delta_n(\mathcal{F})}_{\text{statistical control}} + \underbrace{\varepsilon_{\text{opt}}}_{\text{optimization}} .$$

The model class determines the approximation term, and a later lecture will control the empirical-to-population deviation. This lecture addresses the remaining question: with finite computation, how close does an algorithm get to the best empirical objective value available inside the chosen class?

More concretely, suppose $\mathcal{F} = \{f_\theta : \theta \in \mathcal{K}\}$, define the empirical objective $\Phi_S(\theta) := \widehat{L}_n(f_\theta)$, and let an algorithm return θ_T after T updates. Its contribution to the master bound is

$$\varepsilon_{\text{opt,T}} := \Phi_S(\theta_T) - \inf_{\theta \in \mathcal{K}} \Phi_S(\theta) \ge 0.$$

Conditional on the training sample S, this is an ordinary optimization problem: full-batch methods are deterministic once initialized, whereas SGD adds mini-batch randomness. The goal below is to relate $\varepsilon_{\text{opt,T}}$ to curvature, step sizes, gradient noise, and the iteration budget. A small parameter error is not the definition of success; the quantity required by the learning decomposition is the objective gap displayed above.

2.1 Gradient descent: least squares and basic geometry

To avoid confusing the population risk L with a smoothness constant, write the optimization objective as $\Phi : \mathbb{R}^d \to \mathbb{R}$ and use β for smoothness.

Gradient descent for least squares

Begin with the concrete objective

$$\Phi(\theta) = \frac{1}{2} \|A\theta - b\|^2 . \tag{2.1.1}$$

Its gradient is

$$\nabla\Phi(\theta) = A^\top A\theta - A^\top b, \tag{2.1.2}$$

so gradient descent with a constant step size η becomes

$$\theta_{t+1} = (I - \eta A^\top A)\theta_t + \eta A^\top b. \tag{2.1.3}$$

If $\theta^\star$ satisfies the normal equation $A^\top A\theta^\star = A^\top b$, subtraction gives the exact error recursion

$$\theta_{t+1} - \theta^\star = (I - \eta A^\top A)(\theta_t - \theta^\star). \tag{2.1.4}$$

We now turn this recursion into an exact convergence rate. Assume first that A has full column rank, so

$$H := A^\top A \succ 0, \qquad \mu := \lambda_{\min}(H) > 0, \qquad \beta := \lambda_{\max}(H). \tag{2.1.5}$$

In this case least squares is μ-strongly convex and β-smooth.

Proposition 2.1 (Exact convergence of least-squares gradient descent)**.** *Let A have full column rank, let $\theta^\star$ be the unique least-squares minimizer, and use a constant step size $0 < \eta < 2/\beta$. Define*

$$q(\eta) := \max_{1\le j\le d} |1 - \eta\lambda_j(H)| < 1. \tag{2.1.6}$$

Then, for every integer $T \ge 0$,

$$\|\theta_T - \theta^\star\| \le q(\eta)^T \|\theta_0 - \theta^\star\|, \tag{2.1.7}$$

$$\Phi(\theta_T) - \Phi(\theta^\star) \le q(\eta)^{2T}\{\Phi(\theta_0) - \Phi(\theta^\star)\}. \tag{2.1.8}$$

The spectral proof is assigned in Homework 1, where the same calculation is carried out in the rank-deficient, overparameterized setting and also shows which interpolating solution gradient descent selects.

The upper bound $2/\beta$ is the exact stability boundary for this quadratic problem; the frequently used value $1/\beta$ is not the boundary. Indeed, since the spectrum of H lies in $[\mu, \beta]$, the worst spectral factor occurs at one of the two endpoints:

$$q(\eta) = \max\{|1 - \eta\mu|, |1 - \eta\beta|\}. \tag{2.1.9}$$

The different step-size ranges have visibly different dynamics.

(i) If $0 < \eta < 1/\beta$, then every multiplier $1 - \eta\lambda_j$ lies in $(0, 1)$. Every eigencoordinate contracts without changing sign, and
$$q(\eta) = 1 - \eta\mu. \tag{2.1.10}$$

(ii) If $\eta = 1/\beta$, the largest-curvature eigencoordinate is eliminated in one step, while all other eigencoordinates contract without changing sign. Here $q(\eta) = 1 - \mu/\beta$.

(iii) If $1/\beta < \eta < 2/\beta$, eigencoordinates with $\lambda_j > 1/\eta$ have negative multipliers. They alternate signs, so the iterates oscillate in those directions, but their magnitudes still shrink because $|1 - \eta\lambda_j| < 1$. In this range,
$$q(\eta) = \max\{1 - \eta\mu, \eta\beta - 1\} < 1, \tag{2.1.11}$$
and both the parameter error and objective gap still converge geometrically.

(iv) At $\eta = 2/\beta$, the multiplier in a β-eigendirection is -1, so that component generally does not decay. For $\eta > 2/\beta$, its magnitude exceeds one and that component diverges. Thus strict inequality in Proposition 2.1 is necessary for convergence from every initialization.

Thus $1/\beta$ separates nonoscillatory behavior from possible oscillation in the highest-curvature directions, whereas $2/\beta$ is the stability limit.

We can now choose the optimal constant step size. At $\eta = 1/\beta$,

$$\Phi(\theta_T) - \Phi(\theta^\star) \le \left(1 - \frac{1}{\kappa}\right)^{2T} \{\Phi(\theta_0) - \Phi(\theta^\star)\}, \qquad \kappa := \frac{\beta}{\mu}. \tag{2.1.12}$$

As η increases from zero, the slow low-curvature factor $1 - \eta\mu$ decreases, while beyond $1/\beta$ the oscillatory high-curvature factor $\eta\beta - 1$ increases. The maximum of the two is minimized when they are equal. Therefore

$$1 - \eta_{\text{opt}}\mu = \eta_{\text{opt}}\beta - 1, \qquad \eta_{\text{opt}} = \frac{2}{\mu + \beta}, \qquad q(\eta_{\text{opt}}) = \frac{\beta - \mu}{\beta + \mu} = \frac{\kappa - 1}{\kappa + 1}. \tag{2.1.13}$$

Consequently the objective gap contracts by $((\kappa - 1)/(\kappa + 1))^{2T}$. Both choices require $T = O(\kappa \log(1/\varepsilon))$ iterations to reduce the initial objective gap by a factor ε, but the optimized constant step has the sharper spectral factor. Unless $\mu = \beta$, this optimal step lies strictly between $1/\beta$ and $2/\beta$: it deliberately introduces controlled oscillation in the high-curvature directions to speed up the slow low-curvature directions.

This quadratic example motivates the geometric assumptions used for a general differentiable objective.

Basic geometric definitions

Definition 2.2 (Convexity)**.** A differentiable function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is convex if

$$\Phi(v) \ge \Phi(u) + \langle \nabla\Phi(u), v - u\rangle \qquad \text{for all } u, v \in \mathbb{R}^d. \tag{2.1.14}$$

Definition 2.3 (Strong convexity)**.** For $\mu > 0$, the function Φ is μ-strongly convex if

$$\Phi(v) \geq \Phi(u) + \langle \nabla\Phi(u), v - u\rangle + \frac{\mu}{2}\|v - u\|^2 \qquad \text{for all } u, v \in \mathbb{R}^d. \tag{2.1.15}$$

Definition 2.4 (Smoothness)**.** For $\beta > 0$, the function Φ is β-smooth if

$$\|\nabla\Phi(u) - \nabla\Phi(v)\| \leq \beta\|u - v\| \qquad \text{for all } u, v \in \mathbb{R}^d. \tag{2.1.16}$$

Convexity gives a global affine lower bound. Strong convexity adds positive quadratic curvature and yields a unique minimizer. Smoothness gives a quadratic upper bound and controls how aggressive a gradient step can be.

For a general differentiable objective, gradient descent has the form

$$\theta_{t+1} = \theta_t - \eta_t \nabla\Phi(\theta_t). \tag{2.1.17}$$

The least-squares Hessian is $A^\top A$. Its gradient has Lipschitz modulus $\|A\|_{\text{op}}^2$, so Φ is convex and β-smooth for any positive $\beta \geq \lambda_{\max}(A^\top A) = \|A\|_{\text{op}}^2$. It is μ-strongly convex exactly when $A^\top A$ is positive definite, in which case $\mu = \lambda_{\min}(A^\top A)$. These are the same spectral quantities that appeared directly in (2.1.4).

A common smoothness tool

Lemma 2.5 (Descent lemma)**.** *If Φ is β-smooth, then*

$$\Phi(v) \leq \Phi(u) + \langle \nabla\Phi(u), v - u\rangle + \frac{\beta}{2}\|v - u\|^2. \tag{2.1.18}$$

Proof. Let $d = v - u$. By the fundamental theorem of calculus,

$$\Phi(v) - \Phi(u) - \langle \nabla\Phi(u), d\rangle = \int_0^1 \langle \nabla\Phi(u + sd) - \nabla\Phi(u), d\rangle \, \mathrm{d}s.$$

Smoothness and Cauchy–Schwarz bound the integrand by $\beta s\|d\|^2$. Integrating over $s \in [0, 1]$ proves the claim. □

Corollary 2.6 (One gradient step decreases the objective)**.** *If Φ is β-smooth and $\theta^+ = \theta - \eta\nabla\Phi(\theta)$, then*

$$\Phi(\theta^+) \leq \Phi(\theta) - \eta\left(1 - \frac{\beta\eta}{2}\right)\|\nabla\Phi(\theta)\|^2. \tag{2.1.19}$$

In particular, if $0 < \eta \leq 1/\beta$, then

$$\Phi(\theta^+) \leq \Phi(\theta) - \frac{\eta}{2}\|\nabla\Phi(\theta)\|^2. \tag{2.1.20}$$

Proof. Insert $v = \theta - \eta\nabla\Phi(\theta)$ and $u = \theta$ into Lemma 2.5. □

The geometry behind the two assumptions is summarized in Figure 2.1. At a base point u, both models start with the tangent plane $\Phi(u) + \langle \nabla\Phi(u), v - u\rangle$. Smoothness adds an upward quadratic term to obtain an *upper* model, whereas strong convexity adds an upward quadratic term that remains a *lower* model.

Figure 2.2 isolates the distinction between convexity and strong convexity. Convexity requires a supporting affine lower bound, but it permits flat directions. Strong convexity requires a uniform positive quadratic term above that affine support.

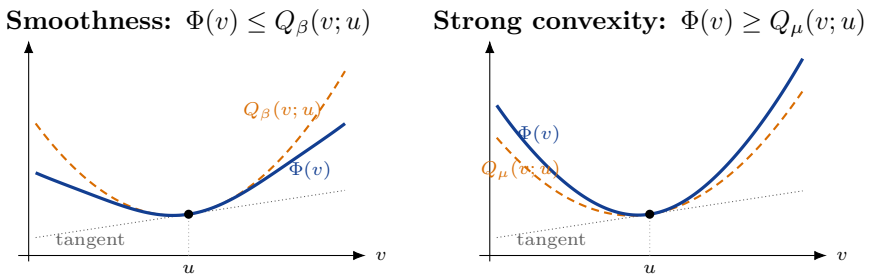

Figure 2.1: Quadratic models around a base point u. The blue curve is the objective, the gray dotted line is its tangent at u, and the orange dashed curve is the quadratic model. A β-smooth objective lies below its quadratic upper model; a μ-strongly convex objective lies above its quadratic lower model.

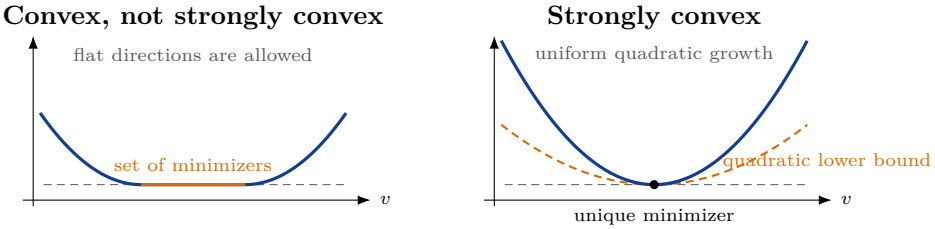

Figure 2.2: Convexity versus strong convexity. Convexity alone can allow a flat set of minimizers, as on the left. Strong convexity supplies uniform quadratic growth above the affine support and therefore forces a unique minimizer, as on the right. A unique minimizer by itself does not imply strong convexity.

2.2 Gradient descent under strong convexity

Lemma 2.7 (Strong convexity implies gradient domination)**.** *If Φ is differentiable and μ-strongly convex, then*

$$\|\nabla\Phi(\theta)\|^2 \ge 2\mu\{\Phi(\theta) - \Phi(\theta^\star)\}. \tag{2.2.1}$$

Proof. Strong convexity gives, for every v,

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

Taking the infimum of the right-hand side over v gives

$$\Phi(\theta^\star) \ge \Phi(\theta) - \frac{1}{2\mu}\|\nabla\Phi(\theta)\|^2.$$

Rearrange. □

Theorem 2.8 (Linear convergence of GD)**.** *Suppose Φ is μ-strongly convex and β-smooth. For a constant step size $0 < \eta \le 1/\beta$, gradient descent satisfies*

$$\Phi(\theta_T) - \Phi(\theta^\star) \le (1-\mu\eta)^T\{\Phi(\theta_0) - \Phi(\theta^\star)\}. \tag{2.2.2}$$

In particular, $\eta = 1/\beta$ gives

$$\Phi(\theta_T) - \Phi(\theta^\star) \le \left(1 - \frac{\mu}{\beta}\right)^T\{\Phi(\theta_0) - \Phi(\theta^\star)\}. \tag{2.2.3}$$

Proof. Corollary 2.6 and $\eta \le 1/\beta$ give

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

Apply Lemma 2.7:

$$\Phi(\theta_{t+1}) - \Phi(\theta^\star) \le (1 - \mu\eta)\{\Phi(\theta_t) - \Phi(\theta^\star)\}.$$

Iterating proves (2.2.2); substituting $\eta = 1/\beta$ gives (2.2.3). □

Theorem 2.9 (Parameter convergence of gradient descent)**.** *Under the assumptions of Theorem 2.8, for every integer $T \ge 0$,*

$$\begin{aligned}\|\theta_T - \theta^\star\|^2 &\le \frac{2}{\mu}\{\Phi(\theta_T) - \Phi(\theta^\star)\} \\ &\le \frac{2}{\mu}(1 - \mu\eta)^T\{\Phi(\theta_0) - \Phi(\theta^\star)\}. \end{aligned} \tag{2.2.4}$$

Consequently,

$$\|\theta_T - \theta^\star\| \le \sqrt{\frac{2\{\Phi(\theta_0) - \Phi(\theta^\star)\}}{\mu}}(1 - \mu\eta)^{T/2}. \tag{2.2.5}$$

Moreover,

$$\|\theta_T - \theta^\star\| \le \sqrt{\frac{\beta}{\mu}}(1 - \mu\eta)^{T/2}\,\|\theta_0 - \theta^\star\|\,. \tag{2.2.6}$$

Thus the iterates themselves converge geometrically to the unique minimizer.

Proof. Since $\nabla\Phi(\theta^\star) = 0$, strong convexity gives

$$\Phi(\theta) - \Phi(\theta^\star) \ge \frac{\mu}{2}\,\|\theta - \theta^\star\|^2\,. \tag{2.2.7}$$

Combining this inequality with Theorem 2.8 proves (2.2.4); taking square roots proves (2.2.5). Finally, smoothness and $\nabla\Phi(\theta^\star) = 0$ imply

$$\Phi(\theta_0) - \Phi(\theta^\star) \le \frac{\beta}{2}\,\|\theta_0 - \theta^\star\|^2\,.$$

Substitution proves (2.2.6). □

The square-root exponent in Theorem 2.9 appears because strong convexity controls the *squared* parameter error by the objective gap. For least squares, the direct spectral rate in Proposition 2.1 is sharper.

The condition number

$$\kappa := \frac{\beta}{\mu} \tag{2.2.8}$$

controls the complexity. Since $(1 - 1/\kappa)^T \le e^{-T/\kappa}$, reducing the initial objective gap by a factor ε takes $T = O(\kappa \log(1/\varepsilon))$ iterations. For full-column-rank least squares, $\kappa = \lambda_{\max}(A^\top A)/\lambda_{\min}(A^\top A)$. The exact quadratic analysis in Proposition 2.1 is sharper: it tracks each eigendirection and gives a squared contraction factor for the objective gap.

2.3 Gradient descent under convexity

Without strong convexity, the objective gap no longer controls $\|\theta - \theta^\star\|^2$: a convex objective may have an entire flat set of minimizers. Consequently, convergence of the objective alone does not generally imply convergence of θ_t to a particular minimizer, nor does it provide a uniform rate in parameter distance. Convexity and smoothness nevertheless still guarantee convergence of the objective value, now at the sublinear rate established below.

The next figure previews a second distinction that will matter when we turn to stochastic optimization. On the same objective, full-gradient descent follows a deterministic trajectory, whereas SGD follows a random, visibly noisier path because every update uses a gradient estimate.

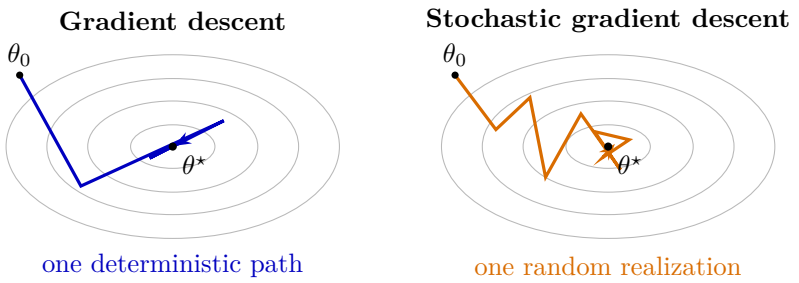

Figure 2.3: Two optimization paths on the same strongly convex objective. The contours are level sets. GD uses the exact gradient and hence has a deterministic path; SGD is perturbed by gradient noise and therefore follows a random path even when its average direction is correct.

Theorem 2.10 (GD on a convex smooth objective)**.** *Suppose* Φ *is convex and* β*-smooth and attains its minimum at* $\theta^\star$*. Let*

$$\theta_{t+1} = \theta_t - \eta \nabla \Phi(\theta_t), \qquad 0 < \eta \le 1/\beta. \tag{2.3.1}$$

Then for every integer $T \ge 1$*,*

$$\Phi(\theta_T) - \Phi(\theta^\star) \le \frac{\|\theta_0 - \theta^\star\|^2}{2\eta T}. \tag{2.3.2}$$

With $\eta = 1/\beta$*, the right-hand side is* $\beta \|\theta_0 - \theta^\star\|^2 / (2T)$*.*

Proof. Write $g_t = \nabla \Phi(\theta_t)$. Convexity gives

$$\Phi(\theta_t) - \Phi(\theta^\star) \le \langle g_t, \theta_t - \theta^\star \rangle.$$

To see the update identity explicitly, subtract $\theta^\star$ from both sides of (2.3.1) and expand the squared norm:

$$\begin{aligned} \|\theta_{t+1} - \theta^\star\|^2 &= \|\theta_t - \theta^\star - \eta g_t\|^2 \\ &= \|\theta_t - \theta^\star\|^2 - 2\eta \langle g_t, \theta_t - \theta^\star \rangle + \eta^2 \|g_t\|^2 . \end{aligned}$$

Move the inner-product term to the left and divide by 2η. This gives

$$\langle g_t, \theta_t - \theta^\star \rangle = \frac{\|\theta_t - \theta^\star\|^2 - \|\theta_{t+1} - \theta^\star\|^2}{2\eta} + \frac{\eta}{2} \|g_t\|^2 . \tag{2.3.3}$$

By Corollary 2.6,

$$\frac{\eta}{2} \|g_t\|^2 \le \Phi(\theta_t) - \Phi(\theta_{t+1}).$$

Combining the last three displays and canceling $\Phi(\theta_t)$ yields

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

Sum from $t = 0$ to $T - 1$. The squared distances telescope. Moreover, Corollary 2.6 shows that the objective values are nonincreasing, so each of the T objective gaps in the sum is at least the final gap. Therefore

$$T\{\Phi(\theta_T) - \Phi(\theta^\star)\} \le \frac{\|\theta_0 - \theta^\star\|^2}{2\eta}.$$

□

Why use a constant learning rate here? For deterministic full-gradient descent there is no persistent gradient noise to average away. Once the step size is small enough to guarantee descent, shrinking it only reduces the cumulative progress. To make this precise, allow

$$\theta_{t+1} = \theta_t - \eta_t \nabla \Phi(\theta_t), \qquad 0 < \eta_t \le 1/\beta. \tag{2.3.4}$$

The same calculation as above, now before dividing by η_t, gives

$$2\eta_t\{\Phi(\theta_{t+1}) - \Phi(\theta^\star)\} \le \|\theta_t - \theta^\star\|^2 - \|\theta_{t+1} - \theta^\star\|^2 .$$

Summing and using the fact that the objective gaps are nonincreasing yields the variable-step bound

$$\Phi(\theta_T) - \Phi(\theta^\star) \le \frac{\|\theta_0 - \theta^\star\|^2}{2\sum_{t=0}^{T-1} \eta_t}. \tag{2.3.5}$$

Thus the rate is governed by the cumulative learning rate. The largest constant choice allowed by this proof, $\eta_t = 1/\beta$, maximizes that quantity and recovers the $O(1/T)$ bound. For example, if $\eta_t = c(t+1)^{-\alpha}$, where $0 < c \le 1/\beta$, then

$$0 < \alpha < 1: \quad \Phi(\theta_T) - \Phi(\theta^\star) \le \frac{(1-\alpha)\|\theta_0 - \theta^\star\|^2}{2c\{(T+1)^{1-\alpha} - 1\}} = O\big(T^{-(1-\alpha)}\big), \tag{2.3.6}$$

$$\alpha = 1: \quad \Phi(\theta_T) - \Phi(\theta^\star) \le \frac{\|\theta_0 - \theta^\star\|^2}{2c\log(T+1)} = O\big(1/\log T\big). \tag{2.3.7}$$

For comparison, under μ-strong convexity the one-step contraction from Theorem 2.8 applies separately at every iteration:

$$\Phi(\theta_T) - \Phi(\theta^\star) \le \{\Phi(\theta_0) - \Phi(\theta^\star)\} \prod_{t=0}^{T-1} (1 - \mu\eta_t) \le \{\Phi(\theta_0) - \Phi(\theta^\star)\} \exp\left(-\mu \sum_{t=0}^{T-1} \eta_t\right). \tag{2.3.8}$$

Writing $\Delta_t := \Phi(\theta_t) - \Phi(\theta^\star)$ and $R := \|\theta_0 - \theta^\star\|$, the comparison is summarized below.

Table 2.1: Objective-gap bounds for deterministic GD. Every displayed step size satisfies $0 < \eta_t \le 1/\beta$.

Learning-rate schedule	μ-strongly convex and β-smooth	Convex and β-smooth
Constant, $\eta_t \equiv \eta$	$\Delta_T \le \Delta_0(1-\mu\eta)^T = O(e^{-\mu\eta T})$	$\Delta_T \le R^2/(2\eta T) = O(T^{-1})$

The table compares objective gaps. Strong convexity additionally converts these bounds into parameter-distance bounds through $\|\theta_T - \theta^\star\|^2 \le 2\Delta_T/\mu$, with the corresponding rates square-rooted for the norm itself. In the merely convex case there is no analogous uniform parameter-distance rate. Decaying learning rates become essential later for SGD because stochastic-gradient noise persists; that noise–progress tradeoff is absent in deterministic GD.

2.4 Stochastic gradient descent

A full gradient of a finite-sum objective

$$\Phi(\theta) = \frac{1}{n}\sum_{i=1}^{n} \phi_i(\theta) \tag{2.4.1}$$

requires n per-example gradients. SGD replaces it by a random gradient estimate. The randomness must be described conditionally because θ_t depends on all previous random draws.

Throughout the SGD discussion, we consider the unconstrained update on $\mathbb{R}^d$:

$$\theta_{t+1} = \theta_t - \eta_t g_t, \qquad t = 1, \dots, T. \tag{2.4.2}$$

Let $\mathcal{H}_t$ contain the data, initialization, and all algorithmic randomness observed before g_t is drawn. We assume θ_t and the positive step size η_t are $\mathcal{H}_t$-measurable; the schedules used below are deterministic special cases.

A standard example is a mini-batch gradient. For the finite sum (2.4.1), conditionally sample $I_{t,1}, \dots, I_{t,b}$ independently and uniformly from $\{1, \dots, n\}$, and set

$$g_t := \frac{1}{b} \sum_{j=1}^{b} \nabla \phi_{I_{t,j}}(\theta_t). \tag{2.4.3}$$

Then

$$\mathbb{E}[g_t \mid \mathcal{H}_t] = \nabla \Phi(\theta_t), \tag{2.4.4}$$

$$\mathrm{Cov}(g_t \mid \mathcal{H}_t) = \frac{1}{b} \mathrm{Cov}(\nabla \phi_I(\theta_t) \mid \mathcal{H}_t), \tag{2.4.5}$$

where I is one conditional uniform draw. Thus independent mini-batching reduces the conditional covariance by a factor b, although it does not remove the randomness of the training sample itself. This construction motivates the following abstract oracle assumption.

Definition 2.11 (Stochastic gradient oracle)**.** The random vector g_t is a conditionally unbiased stochastic gradient with bounded conditional second moment if

$$\mathbb{E}[g_t \mid \mathcal{H}_t] = \nabla \Phi(\theta_t), \qquad \mathbb{E}[\|g_t\|^2 \mid \mathcal{H}_t] \le G^2 \quad \text{almost surely.} \tag{2.4.6}$$

Conditional unbiasedness concerns the gradient estimate. It does not say that the iterate, the objective value, or the test error is unbiased.

Strongly convex objectives

We begin with strong convexity. Its quadratic lower bound turns the expected alignment of the stochastic gradient with the parameter error into a contraction term. We first hold the learning rate constant. This separates a geometrically decaying optimization transient from the nonvanishing effect of gradient noise. We then decrease the learning rate to remove that noise floor and obtain an objective rate faster than the general convex rate considered afterward.

For the constant-step analysis, separate the stochastic gradient into signal and noise:

$$g_t = \nabla \Phi(\theta_t) + \xi_t, \qquad \mathbb{E}[\xi_t \mid \mathcal{H}_t] = 0, \qquad \mathbb{E}[\|\xi_t\|^2 \mid \mathcal{H}_t] \le \sigma^2. \tag{2.4.7}$$

The oracle assumption (2.4.6) implies the last bound with $\sigma^2 = G^2$, but the separate symbol σ^2 makes the role of gradient noise explicit.

Theorem 2.12 (Constant-step SGD under strong convexity)**.** *Suppose Φ is μ-strongly convex and β-smooth, attains its minimum at $\theta^\star$, and (2.4.7) holds. Use*

$$\theta_{t+1} = \theta_t - \eta g_t, \qquad 0 < \eta \le 1/\beta. \tag{2.4.8}$$

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

$$\begin{aligned} \mathbb{E}[\Phi(\theta_T)] - \Phi(\theta^\star) &\le (1 - \mu\eta)^{T-1} \{\mathbb{E}[\Phi(\theta_1)] - \Phi(\theta^\star)\} \\ &\quad + \frac{\beta \eta \sigma^2}{2\mu} \{1 - (1 - \mu\eta)^{T-1}\} \qquad (2.4.9) \\ &\le (1 - \mu\eta)^{T-1} \{\mathbb{E}[\Phi(\theta_1)] - \Phi(\theta^\star)\} + \frac{\beta \eta \sigma^2}{2\mu}. \qquad (2.4.10) \end{aligned}$$

Thus constant-step SGD approaches an $O(\eta)$ objective neighborhood of the optimum at a geometric rate.

Proof. First fix a realization of the stochastic gradient g_t. By β-smoothness and (2.4.8), the following pathwise inequality holds before taking any expectation:

$$\Phi(\theta_{t+1}) \le \Phi(\theta_t) + \langle \nabla \Phi(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{\beta}{2} \|\theta_{t+1} - \theta_t\|^2$$

$$= \Phi(\theta_t) - \eta\langle \nabla\Phi(\theta_t), g_t\rangle + \frac{\beta\eta^2}{2}\|g_t\|^2 .$$

Now condition on $\mathcal{H}_t$. Because θ_t and hence $\nabla\Phi(\theta_t)$ are $\mathcal{H}_t$-measurable, the preceding display gives

$$\mathbb{E}[\Phi(\theta_{t+1}) \mid \mathcal{H}_t] \le \Phi(\theta_t) - \eta\langle \nabla\Phi(\theta_t), \mathbb{E}[g_t \mid \mathcal{H}_t]\rangle + \frac{\beta\eta^2}{2}\mathbb{E}[\|g_t\|^2 \mid \mathcal{H}_t].$$

The conditional mean-zero property of ξ_t makes the cross term vanish, so

$$\mathbb{E}[\|g_t\|^2 \mid \mathcal{H}_t] = \|\nabla\Phi(\theta_t)\|^2 + \mathbb{E}[\|\xi_t\|^2 \mid \mathcal{H}_t].$$

Using (2.4.7) and $\eta \le 1/\beta$ therefore gives

$$\begin{aligned}\mathbb{E}[\Phi(\theta_{t+1}) \mid \mathcal{H}_t] &\le \Phi(\theta_t) - \eta\left(1 - \frac{\beta\eta}{2}\right)\|\nabla\Phi(\theta_t)\|^2 + \frac{\beta\eta^2\sigma^2}{2} \\ &\le \Phi(\theta_t) - \frac{\eta}{2}\|\nabla\Phi(\theta_t)\|^2 + \frac{\beta\eta^2\sigma^2}{2}.\end{aligned}$$

Strong convexity implies the Polyak–Łojasiewicz inequality

$$\|\nabla\Phi(\theta_t)\|^2 \ge 2\mu\{\Phi(\theta_t) - \Phi(\theta^\star)\}.$$

Taking total expectations and writing $\delta_t := \mathbb{E}[\Phi(\theta_t)] - \Phi(\theta^\star)$ yields

$$\delta_{t+1} \le (1 - \mu\eta)\delta_t + \frac{\beta\eta^2\sigma^2}{2}. \tag{2.4.11}$$

Iterating this affine recursion and summing the geometric series proves (2.4.9); dropping the two factors bounded by one gives (2.4.10). □

The two terms in (2.4.10) expose the basic tradeoff. A larger η makes the transient $(1-\mu\eta)^{T-1}$ disappear faster but raises the asymptotic noise floor; a smaller η lowers that floor but slows the transient. When $\sigma^2 = 0$, the noise floor disappears and the deterministic geometric rate is recovered. With persistent noise, exact convergence requires the learning rate to decrease.

Decreasing learning rates. We now let the step size shrink with t. With smoothness, the same objective-gap recursion used for constant-step SGD closes at every time and therefore gives a guarantee for the final iterate itself.

Proposition 2.13 (Final-iterate SGD with a decreasing step size)**.** *Suppose Φ is μ-strongly convex and β-smooth, attains its minimum at $\theta^\star$, and (2.4.7) holds. Let*

$$\eta_t = \frac{2}{\mu(t+\tau)}, \qquad \tau \ge \frac{2\beta}{\mu}. \tag{2.4.12}$$

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

$$\mathbb{E}[\Phi(\theta_T)] - \Phi(\theta^\star) \le \frac{\tau\{\mathbb{E}[\Phi(\theta_1)] - \Phi(\theta^\star)\} + 2\beta\sigma^2/\mu^2}{T + \tau - 1} = O\left(\frac{1}{T}\right). \tag{2.4.13}$$

Thus the decreasing step size removes the constant-step noise floor without averaging the iterates.

Proof. Write $\delta_t := \mathbb{E}[\Phi(\theta_t)] - \Phi(\theta^\star)$. Because $\eta_t \le 1/\beta$, the proof of Theorem 2.12, applied with the current value of η_t, gives

$$\delta_{t+1} \le (1 - \mu\eta_t)\delta_t + \frac{\beta\eta_t^2\sigma^2}{2} = \left(1 - \frac{2}{t+\tau}\right)\delta_t + \frac{C}{(t+\tau)^2}, \qquad C := \frac{2\beta\sigma^2}{\mu^2}. \tag{2.4.14}$$

Set $A := \tau\delta_1 + C$. We prove by induction that $\delta_t \le A/(t+\tau-1)$. The claim holds at $t = 1$. If it holds at time t, set $s = t + \tau$ in (2.4.14) to obtain

$$\delta_{t+1} \le \frac{A(s-2)}{s(s-1)} + \frac{C}{s^2} \le \frac{A}{s}.$$

The last inequality follows from $A \ge C$, since the difference between the first and last fractions involving A is $A/[s(s-1)] \ge C/s^2$. This proves (2.4.13). □

For compactness, write

$$\mathcal{A} := \tau\{\mathbb{E}[\Phi(\theta_1)] - \Phi(\theta^\star)\} + \frac{2\beta\sigma^2}{\mu^2}. \tag{2.4.15}$$

Corollary 2.14 (Parameter convergence of the final iterate)**.** *Under the assumptions and step-size schedule of Proposition 2.13,*

$$\mathbb{E}\,\|\theta_T - \theta^\star\|^2 \le \frac{2\mathcal{A}}{\mu(T+\tau-1)} = O\left(\frac{1}{T}\right), \tag{2.4.16}$$

$$\|\mathbb{E}[\theta_T] - \theta^\star\| \le \mathbb{E}\,\|\theta_T - \theta^\star\| \le \sqrt{\frac{2\mathcal{A}}{\mu(T+\tau-1)}} = O\left(\frac{1}{\sqrt{T}}\right). \tag{2.4.17}$$

Thus $\theta_T \to \theta^\star$ in mean square and in mean; the squared parameter error has the same $O(1/T)$ order as the objective gap, while the parameter norm itself has the square-root rate.

Proof. Strong convexity at the minimizer gives

$$\Phi(\theta) - \Phi(\theta^\star) \ge \frac{\mu}{2}\,\|\theta - \theta^\star\|^2.$$

Combine this inequality with (2.4.13). Jensen's inequality gives the first inequality in (2.4.17), and Cauchy–Schwarz gives the second. □

Proposition 2.15 (Polyak averaging in expectation)**.** *Under the assumptions and step-size schedule of Proposition 2.13, define the arithmetic Polyak average*

$$\overline{\theta}_T^{\mathrm{PR}} := \frac{1}{T}\sum_{t=1}^{T}\theta_t. \tag{2.4.18}$$

Then

$$\mathbb{E}\left\|\overline{\theta}_T^{\mathrm{PR}} - \theta^\star\right\|^2 \le \frac{8\mathcal{A}}{\mu T}, \tag{2.4.19}$$

$$\mathbb{E}[\Phi(\overline{\theta}_T^{\mathrm{PR}})] - \Phi(\theta^\star) \le \frac{4\beta\mathcal{A}}{\mu T} = O\left(\frac{1}{T}\right). \tag{2.4.20}$$

Consequently, the arithmetic average converges to $\theta^\star$ in mean square, and its objective converges to the optimum in expectation. The mean-square parameter error and the expected objective gap are both $O(1/T)$.

Proof. For $e_t := \theta_t - \theta^\star$, strong convexity and (2.4.13) imply

$$\mathbb{E}\,\|e_t\|^2 \le \frac{2\mathcal{A}}{\mu(t+\tau-1)} \le \frac{2\mathcal{A}}{\mu t}.$$

Minkowski's inequality in L^2, followed by $\sum_{t=1}^{T} t^{-1/2} \le 2\sqrt{T}$, gives

$$\begin{aligned}\left(\mathbb{E}\left\|\overline{\theta}_T^{\mathrm{PR}} - \theta^\star\right\|^2\right)^{1/2} &\le \frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}\,\|e_t\|^2\right)^{1/2} \\ &\le \frac{1}{T}\sqrt{\frac{2\mathcal{A}}{\mu}}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \le \sqrt{\frac{8\mathcal{A}}{\mu T}}.\end{aligned}$$

This proves (2.4.19). Smoothness at the minimizer, where $\nabla\Phi(\theta^\star) = 0$, gives the quadratic upper bound

$$\Phi(\theta) - \Phi(\theta^\star) \le \frac{\beta}{2}\,\|\theta - \theta^\star\|^2.$$

Apply it at $\overline{\theta}_T^{\mathrm{PR}}$, take expectations, and use (2.4.19). □

Remark 2.16 (What changes without a smoothness assumption)**.** The arithmetic-average objective rates with and without smoothness are not the same. Smoothness gives the $O(1/T)$ rate in Proposition 2.15: after averaging the parameter errors, its quadratic upper model converts their mean-square rate into an objective-gap rate. If we assume only differentiability, μ-strong convexity, and the bounded conditional second moment in (2.4.6), expanding the squared parameter distance instead gives

$$2\eta_t\delta_t \le (1-\mu\eta_t)d_t - d_{t+1} + \eta_t^2 G^2, \qquad d_t := \mathbb{E}\left\|\theta_t - \theta^\star\right\|^2. \tag{2.4.21}$$

Without a quadratic upper model, this inequality naturally controls a sum of objective gaps rather than converting an averaged parameter error into an objective gap. The stochastic objective values need not decrease monotonically, so the same calculation also cannot isolate δ_T. In particular, $\eta_t = 1/(\mu t)$ and the uniform average give

$$\mathbb{E}\left[\Phi\left(\frac{1}{T}\sum_{t=1}^{T}\theta_t\right)\right] - \Phi(\theta^\star) \le \frac{G^2 H_T}{2\mu T} = O\left(\frac{\log T}{T}\right),$$

whereas $\eta_t = 2/[\mu(t+1)]$ together with weights proportional to t gives $2G^2/[\mu(T+1)] = O(1/T)$. Thus the standard arithmetic Polyak average has an $O(1/T)$ expected objective rate with smoothness but only the general $O(\log T/T)$ guarantee without it. A different, linearly weighted average recovers $O(1/T)$ without smoothness.

Convex objectives

Without strong convexity, the contraction term disappears. The same squared-distance calculation still controls an average objective gap, but at the slower $O(T^{-1/2})$ rate.

Theorem 2.17 (Convex SGD with step-size-weighted averaging)**.** *Suppose Φ is convex and differentiable, attains its minimum at $\theta^\star$, and (2.4.6) holds. Let $\eta_1, \dots, \eta_T > 0$ be deterministic, and define*

$$S_T := \sum_{t=1}^{T}\eta_t, \qquad Q_T := \sum_{t=1}^{T}\eta_t^2, \qquad \overline{\theta}_T^{(\eta)} := \frac{1}{S_T}\sum_{t=1}^{T}\eta_t\theta_t. \tag{2.4.22}$$

Suppose the possibly random initialization satisfies $\mathbb{E}\left\|\theta_1 - \theta^\star\right\|^2 \le R^2$. Then

$$\mathbb{E}[\Phi(\overline{\theta}_T^{(\eta)})] - \Phi(\theta^\star) \le \frac{R^2 + G^2 Q_T}{2S_T}. \tag{2.4.23}$$

For the constant schedule $\eta_t \equiv \eta$, the weighted average is the uniform average

$$\overline{\theta}_T := \frac{1}{T}\sum_{t=1}^{T}\theta_t, \tag{2.4.24}$$

and (2.4.23) becomes

$$\mathbb{E}[\Phi(\overline{\theta}_T)] - \Phi(\theta^\star) \le \frac{R^2}{2\eta T} + \frac{\eta G^2}{2}. \tag{2.4.25}$$

For $R > 0$ and $G > 0$, choosing $\eta = R/(G\sqrt{T})$ gives

$$\mathbb{E}[\Phi(\overline{\theta}_T)] - \Phi(\theta^\star) \le \frac{RG}{\sqrt{T}}. \tag{2.4.26}$$

For a genuinely decreasing schedule, let $\eta_t = ct^{-\alpha}$, where $c > 0$ and $0 < \alpha \le 1$, and write $H_{T,p} := \sum_{t=1}^{T} t^{-p}$. Then

$$\mathbb{E}[\Phi(\overline{\theta}_T^{(\eta)})] - \Phi(\theta^\star) \le \frac{R^2 + c^2 G^2 H_{T,2\alpha}}{2cH_{T,\alpha}} = \begin{cases} O(T^{-\alpha}), & 0 < \alpha < 1/2, \\ O(\log T/\sqrt{T}), & \alpha = 1/2, \\ O(T^{-(1-\alpha)}), & 1/2 < \alpha < 1, \\ O(1/\log T), & \alpha = 1. \end{cases} \tag{2.4.27}$$

The decreasing schedule does not require the horizon T in advance, but the critical choice $\eta_t = c/\sqrt{t}$ pays a logarithmic factor in this elementary weighted-average bound. The horizon-tuned constant step attains the clean $O(T^{-1/2})$ rate. If $\alpha > 1$, then S_T stays bounded, so (2.4.23) need not converge to zero.

Remark 2.18 (What the theorem does and does not say)**.** The guarantee is in expectation over the algorithmic randomness. A high-probability guarantee needs stronger tail control than a bounded conditional second moment. The theorem also controls the averaged iterate, not automatically the final iterate. These distinctions should not be hidden inside big-O notation.

For a common notation in the comparison below, write $\Delta(\theta) := \Phi(\theta) - \Phi(\theta^\star)$, $\Delta_T := \Delta(\theta_T)$, and $e_T := \theta_T - \theta^\star$. Expectations are included for stochastic methods.

Table 2.2: Convergence rates obtained through Section 2.4. The reported iterate and assumptions are part of each guarantee.

Method	Geometry and reported iterate	Objective-gap rate	Parameter conclusion
GD	μ-strongly convex, β-smooth; final	$\Delta_T = O(e^{-T/\kappa})$	$\|e_T\| = O(e^{-T/(2\kappa)})$ from the general bound
GD	Convex, β-smooth; final	$\Delta_T = O(T^{-1})$	No uniform rate to a particular minimizer
SGD, constant η	μ-strongly convex, β-smooth; final	$\mathbb{E}\Delta_T = O(e^{-\mu\eta T}) + O(\eta\sigma^2)$	$\mathbb{E}\,\|e_T\|^2$ has the same transient–floor form
SGD, $\eta_t \asymp 1/t$	μ-strongly convex, β-smooth; final or arithmetic average	$\mathbb{E}\Delta_T = O(T^{-1})$	$\mathbb{E}\,\|e_T\|^2 = O(T^{-1})$, also for the average
SGD, horizon-tuned η	Convex with bounded second moment; uniform average	$\mathbb{E}\Delta(\overline{\theta}_T) = O(T^{-1/2})$	No uniform parameter rate
SGD, $\eta_t = c/\sqrt{t}$	Convex with bounded second moment; step-size-weighted average	$\mathbb{E}\Delta(\overline{\theta}_T^{(\eta)}) = O(\log T/\sqrt{T})$	Horizon-free, but no uniform parameter rate

The two deterministic rows have no gradient-noise floor. Strong convexity is what turns an objective rate into a parameter rate; averaging is what permits the general convex SGD guarantee to concern a reported iterate.

2.5 Uncertainty and asymptotic normality of SGD

The preceding results measure optimization error through expected objective gaps. Under randomness, however, how should we measure the uncertainty of the estimated parameter itself? A central limit theorem answers this by giving an asymptotic distribution, and hence an approximate covariance and confidence region. We first study the final SGD iterate and then show how Polyak–Ruppert averaging changes its limiting covariance.

A local stochastic-approximation model

Consider the unconstrained recursion

$$\theta_{t+1} = \theta_t - \eta_t \widehat{g}_{t+1}(\theta_t), \qquad \xi_{t+1} := \widehat{g}_{t+1}(\theta_t) - \nabla\Phi(\theta_t). \tag{2.5.1}$$

Here $\mathcal{H}_t$ contains the history before the new stochastic gradient is drawn. Write $e_t := \theta_t - \theta^\star$. To make the two central limit theorems transparent, assume:

1. $\theta_t \to \theta^\star$ in probability, where $\nabla\Phi(\theta^\star) = 0$. For each step-size schedule below, the local mean-square rate satisfies $\mathbb{E}\,\|e_t\|^2 = O(\eta_t)$.
2. $H := \nabla^2\Phi(\theta^\star)$ is positive definite and the Hessian is locally Lipschitz. Thus, near $\theta^\star$,

$$\nabla\Phi(\theta_t) = He_t + r_t, \qquad \|r_t\| \le C\,\|e_t\|^2. \tag{2.5.2}$$

3. $\mathbb{E}[\xi_{t+1} \mid \mathcal{H}_t] = 0$, and for some positive-semidefinite matrix S,
$$\mathbb{E}[\xi_{t+1}\xi_{t+1}^\top \mid \mathcal{H}_t] \longrightarrow S \quad \text{in probability;} \tag{2.5.3}$$
4. the noise has a locally uniform conditional $(2+\delta)$-moment for some $\delta > 0$. In particular, the conditional Lindeberg condition needed by the martingale central limit theorem holds.

The first condition is a stability and rate assumption, not a consequence of the local calculation. Global regularity or a separate convergence theorem is needed to show that the recursion enters this neighborhood and has the stated mean-square rate.

Asymptotic normality of the final SGD iterate

Theorem 2.19 (Last-iterate SGD central limit theorem)**.** *Under the local assumptions above, let $\eta_t = a/t$ with $a > 0$. If $2a\lambda_{\min}(H) > 1$, then*
$$\sqrt{T}(\theta_T - \theta^\star) \xrightarrow{d} \mathcal{N}(0, V_a), \tag{2.5.4}$$
where V_a is the unique symmetric solution of the Lyapunov equation
$$\left(aH - \frac{1}{2}I\right)V_a + V_a\left(aH - \frac{1}{2}I\right)^\top = a^2 S. \tag{2.5.5}$$

Proof. By (2.5.2), the error recursion is
$$e_{t+1} = \left(I - \frac{aH}{t}\right)e_t - \frac{a}{t}\xi_{t+1} - \frac{a}{t}r_t. \tag{2.5.6}$$
For $1 \le t < T$, define the transition matrix
$$\Gamma_{T,t} := \prod_{j=t+1}^{T-1}\left(I - \frac{aH}{j}\right), \qquad \Gamma_{T,0} := \prod_{j=1}^{T-1}\left(I - \frac{aH}{j}\right),$$
where an empty product is the identity. Iterating (2.5.6) gives
$$\sqrt{T}e_T = \sqrt{T}\Gamma_{T,0}e_1 - \sum_{t=1}^{T-1} W_{T,t}\xi_{t+1} - \sum_{t=1}^{T-1} W_{T,t}r_t, \qquad W_{T,t} := \sqrt{T}\frac{a}{t}\Gamma_{T,t}. \tag{2.5.7}$$
Because H is positive definite, $\|\Gamma_{T,t}\| \le C(t/T)^{a\lambda_{\min}(H)}$. Hence $2a\lambda_{\min}(H) > 1$ makes the initialization term in (2.5.7) vanish. The quadratic remainder bound in (2.5.2), together with $\mathbb{E}\,\|e_t\|^2 = O(t^{-1})$, similarly gives
$$\sum_{t=1}^{T-1} W_{T,t}r_t = o_{\mathbb{P}}(1).$$
It remains to analyze the weighted martingale differences. For $x \in (0,1]$, write $x^{aH} := \exp(aH\log x)$. The product approximation $\Gamma_{T,t} = (t/T)^{aH} + o(1)$, the covariance assumption, and a Riemann-sum argument give
$$\sum_{t=1}^{T-1} W_{T,t}\mathbb{E}[\xi_{t+1}\xi_{t+1}^\top \mid \mathcal{H}_t]W_{T,t}^\top \xrightarrow{\mathbb{P}} V_a := a^2\int_0^1 x^{aH-I}Sx^{aH^\top - I}\,dx. \tag{2.5.8}$$
The integral is finite under the displayed stability condition. The conditional moment assumption gives the corresponding Lindeberg condition, so the martingale triangular-array central limit theorem yields $\sum_t W_{T,t}\xi_{t+1} \Rightarrow \mathcal{N}(0, V_a)$.

Finally, with $A = aH - I/2$ and the change of variables $x = e^{-s}$,
$$V_a = a^2\int_0^\infty e^{-As}Se^{-A^\top s}\,ds.$$
Differentiating the integrand and integrating from zero to infinity gives $AV_a + V_aA^\top = a^2S$. Slutsky's theorem applied to (2.5.7) proves (2.5.4). □

In one dimension, if $H = h > 0$ and $S = \sigma_g^2$, then

$$V_a = \frac{a^2\sigma_g^2}{2ah - 1}, \qquad ah > \frac{1}{2}. \tag{2.5.9}$$

This variance is minimized at $a = 1/h$, where it equals σ_g^2/h^2. Thus the last iterate is asymptotically normal, but its covariance can be sensitive to the learning-rate constant and the curvature.

Polyak–Ruppert averaging

Proposition 2.15 gave a finite-time $O(1/T)$ expectation bound for the arithmetic average under global smooth strong convexity and a $1/t$-scale step size. We now ask a finer question: after normalization, what is the limiting distribution of that average? For this covariance-efficient limit, we use a slightly slower decay $t^{-\alpha}$ with $1/2 < \alpha < 1$ and the local assumptions above.

Theorem 2.20 (Polyak–Ruppert central limit theorem)**.** *Under the same local assumptions, use*

$$\eta_t = at^{-\alpha}, \qquad a > 0, \qquad \frac{1}{2} < \alpha < 1, \tag{2.5.10}$$

and report the arithmetic average

$$\bar{\theta}_T^{\mathrm{PR}} := \frac{1}{T}\sum_{t=1}^{T}\theta_t. \tag{2.5.11}$$

Then

$$\sqrt{T}(\bar{\theta}_T^{\mathrm{PR}} - \theta^\star) \xrightarrow{d} \mathcal{N}\big(0, H^{-1}S(H^{-1})^\top\big). \tag{2.5.12}$$

In particular, the leading covariance does not depend on a or α within this range.

Proof. Substitute (2.5.2) into the SGD recursion:

$$e_{t+1} - e_t = -\eta_t(He_t + \xi_{t+1} + r_t).$$

Rearrange and average from $t = 1$ to T:

$$H(\bar{\theta}_T^{\mathrm{PR}} - \theta^\star) = -\frac{1}{T}\sum_{t=1}^{T}\xi_{t+1} - \frac{1}{T}\sum_{t=1}^{T}\frac{\theta_{t+1} - \theta_t}{\eta_t} - \frac{1}{T}\sum_{t=1}^{T}r_t. \tag{2.5.13}$$

For the increment term, summation by parts gives

$$\sum_{t=1}^{T}\frac{e_{t+1} - e_t}{\eta_t} = \frac{e_{T+1}}{\eta_T} - \frac{e_1}{\eta_1} + \sum_{t=2}^{T}e_t\left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t}\right). \tag{2.5.14}$$

Here $\eta_t^{-1} = a^{-1}t^\alpha$ and the local mean-square bound implies $e_t = O_{\mathbb{P}}(t^{-\alpha/2})$. Consequently, the right-hand side of (2.5.14) is $O_{\mathbb{P}}(T^{\alpha/2})$, so after division by $\sqrt{T}$ it is $o_{\mathbb{P}}(1)$ because $\alpha < 1$.

The quadratic Taylor remainder and $\mathbb{E}\,\|e_t\|^2 = O(t^{-\alpha})$ give

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}r_t = O_{\mathbb{P}}(T^{1/2-\alpha}) = o_{\mathbb{P}}(1),$$

where the last step uses $\alpha > 1/2$. Multiplying (2.5.13) by $\sqrt{T}$ therefore yields the asymptotic linear representation

$$\sqrt{T}(\bar{\theta}_T^{\mathrm{PR}} - \theta^\star) = -H^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_{t+1} + o_{\mathbb{P}}(1). \tag{2.5.15}$$

By (2.5.3), the Cesàro average of the conditional covariances converges in probability to S. The conditional Lindeberg condition then gives

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_{t+1} \xrightarrow{d} \mathcal{N}(0, S)$$

by the martingale central limit theorem. Apply Slutsky's theorem to (2.5.15) to obtain (2.5.12). □

The limiting covariance coincides with the covariance obtained by using the ideal matrix gain H^{-1} in the linearized recursion; this is the sense in which Polyak–Ruppert averaging is asymptotically covariance-efficient. The learning-rate parameters still affect transient behavior and finite-time error. The restriction $1/2 < \alpha < 1$ balances $\sum_t \eta_t = \infty$ with $\sum_t \eta_t^2 < \infty$; the endpoint $\alpha = 1$ requires additional conditions and can retain tuning effects.

Example 2.21 (Online least squares)**:** Let

$$\Phi(\theta) = \frac{1}{2}\mathbb{E}[(Y - X^\top\theta)^2], \qquad \widehat{g}(\theta; X, Y) = (X^\top\theta - Y)X.$$

Write $\varepsilon = Y - X^\top\theta^\star$, and assume $H = \mathbb{E}[XX^\top]$ is positive definite. At the optimum,

$$S = \mathbb{E}[\varepsilon^2 XX^\top], \qquad \sqrt{T}(\bar{\theta}_T^{\mathrm{PR}} - \theta^\star) \xrightarrow{d} \mathcal{N}\big(0, H^{-1}SH^{-1}\big)\,. \tag{2.5.16}$$

Under homoskedastic noise, $\mathbb{E}[\varepsilon^2 \mid X] = \sigma^2$, so $S = \sigma^2 H$ and the limiting covariance simplifies to $\sigma^2 H^{-1}$. This is the familiar least-squares covariance.

Output and step size	$\sqrt{T}$-scale limit	Main point
Final iterate, $\eta_t = a/t$	$\mathcal{N}(0, V_a)$, with V_a from (2.5.5))	Covariance depends on the scalar gain and curvature.
PR arithmetic average, $\eta_t = at^{-\alpha}$	$\mathcal{N}(0, H^{-1}S(H^{-1})^\top)$	Efficient leading covariance without tuning a to H.

These limits concern the number T of stochastic-approximation iterations for a fixed population objective and noise mechanism. Conditional finite-sum SGD has an analogous algorithmic limit around its empirical minimizer, but this is distinct from an $n \to \infty$ generalization or statistical sampling limit.

2.6 Momentum under strong convexity and smoothness

Throughout this section, the objective Φ is μ-strongly convex and β-smooth, so it has a unique minimizer $\theta^\star$. Momentum is a change to the optimization dynamics, not an average of the reported iterates. We first isolate the deterministic mechanism and then add stochastic gradients.

Deterministic heavy-ball momentum

With a constant step size $\eta > 0$ and momentum parameter $0 \le \rho < 1$, heavy-ball momentum can be written as

$$m_t = \rho m_{t-1} + \nabla\Phi(\theta_t), \qquad \theta_{t+1} = \theta_t - \eta m_t, \qquad m_{-1} = 0, \tag{2.6.1}$$

or, equivalently for $t \ge 1$,

$$\theta_{t+1} = \theta_t - \eta\nabla\Phi(\theta_t) + \rho(\theta_t - \theta_{t-1}). \tag{2.6.2}$$

The extra term preserves motion in directions whose gradients remain consistent and damps directions whose gradients repeatedly reverse.

The acceleration mechanism is exact for strongly convex quadratics. Let

$$\Phi(\theta) = \frac{1}{2}(\theta - \theta^\star)^\top H(\theta - \theta^\star), \qquad \mu I \preceq H \preceq \beta I, \tag{2.6.3}$$

where $\mu = \lambda_{\min}(H)$ and $\beta = \lambda_{\max}(H)$. Along an eigenvector of H with eigenvalue λ, the scalar error satisfies

$$z_{t+1} = (1 + \rho - \eta\lambda)z_t - \rho z_{t-1}. \tag{2.6.4}$$

Proposition 2.22 (Exact quadratic heavy-ball rate)**.** *For* (2.6.3)*, define*

$$r_{\pm}(\lambda) := \frac{1+\rho-\eta\lambda \pm \sqrt{(1+\rho-\eta\lambda)^2-4\rho}}{2}. \tag{2.6.5}$$

The iteration is stable in every eigendirection if and only if

$$0 \le \rho < 1, \qquad 0 < \eta < \frac{2(1+\rho)}{\beta}. \tag{2.6.6}$$

Its asymptotic parameter contraction factor is

$$q(\eta,\rho) = \max_{\lambda\in\mathrm{spec}(H)} \max\{|r_+(\lambda)|, |r_-(\lambda)|\}. \tag{2.6.7}$$

The quadratic-optimal parameters are

$$\eta_{\mathrm{HB}} = \frac{4}{(\sqrt{\beta}+\sqrt{\mu})^2}, \qquad \rho_{\mathrm{HB}} = \left(\frac{\sqrt{\beta}-\sqrt{\mu}}{\sqrt{\beta}+\sqrt{\mu}}\right)^2, \tag{2.6.8}$$

and they give

$$\limsup_{t\to\infty} \|\theta_t - \theta^\star\|^{1/t} \le q_{\mathrm{HB}} := \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}, \qquad \kappa := \frac{\beta}{\mu}. \tag{2.6.9}$$

By comparison, optimally tuned GD has parameter factor $(\kappa-1)/(\kappa+1)$*.*

Proof. Diagonalizing H reduces the vector recursion to (2.6.4), whose characteristic polynomial is $r^2 - (1+\rho-\eta\lambda)r + \rho$. The second-order Jury conditions place both roots inside the unit disk exactly when $\rho < 1$, $\eta\lambda > 0$, and $\eta\lambda < 2(1+\rho)$. Enforcing these inequalities over $[\mu,\beta]$ gives (2.6.6). Substituting (2.6.8) into the roots shows that their maximum modulus over $[\mu,\beta]$ is q_{HB}. At an endpoint the root can be repeated, producing a polynomial prefactor such as tq_{HB}^t, but not changing the asymptotic root factor in (2.6.9). □

This accelerated factor is a spectral statement for quadratics. For a general nonlinear μ-strongly convex, β-smooth objective, the same quadratic-optimal parameters need not provide a global guarantee; a valid global theorem requires a separate Lyapunov argument and generally a more conservative parameter region. Thus the quadratic calculation explains the mechanism without silently extending it beyond its assumptions [12].

Stochastic momentum

For the stochastic result, specialize the finite sum $\Phi = n^{-1}\sum_{i=1}^n \phi_i$: assume that every ϕ_i is convex and $L_{\max}$-smooth, and retain the assumption that Φ is μ-strongly convex. Sample I_t conditionally uniformly and define

$$\sigma_\star^2 := \mathbb{E}_I \left\|\nabla\phi_I(\theta^\star)\right\|^2. \tag{2.6.10}$$

Convex component smoothness gives the variance-transfer inequality

$$\mathbb{E}_I \left\|\nabla\phi_I(\theta)\right\|^2 \le 4L_{\max}\{\Phi(\theta) - \Phi(\theta^\star)\} + 2\sigma_\star^2. \tag{2.6.11}$$

This is stronger structure than conditional unbiasedness alone.

Consider the time-varying stochastic momentum recursion

$$m_t = \rho_t m_{t-1} + \nabla\phi_{I_t}(\theta_t), \qquad \theta_{t+1} = \theta_t - \gamma_t m_t, \qquad m_{-1} = 0. \tag{2.6.12}$$

Theorem 2.23 (A final-iterate bound for stochastic momentum)**.** *Under the assumptions above, choose*

$$\gamma_t = \frac{2\eta}{t+3}, \qquad \rho_t = \frac{t}{t+2}, \qquad 0 < \eta \le \frac{1}{4L_{\max}}. \tag{2.6.13}$$

If $R := \|\theta_0 - \theta^\star\|$, then, for every $T \geq 0$,

$$\mathbb{E}[\Phi(\theta_T)] - \Phi(\theta^\star) \leq \frac{R^2}{\eta(T+1)} + 2\eta\sigma_\star^2, \tag{2.6.14}$$

$$\mathbb{E}\,\|\theta_T - \theta^\star\|^2 \leq \frac{2R^2}{\mu\eta(T+1)} + \frac{4\eta\sigma_\star^2}{\mu}. \tag{2.6.15}$$

Thus a fixed base value η gives a transient plus a noise floor. If $R > 0$, $\sigma_\star > 0$, and the step-size restriction is satisfied, the horizon-dependent choice $\eta = R/[\sigma_\star\sqrt{2(T+1)}]$ gives

$$\mathbb{E}[\Phi(\theta_T)] - \Phi(\theta^\star) \leq \frac{2\sqrt{2}R\sigma_\star}{\sqrt{T+1}}. \tag{2.6.16}$$

Proof. First, convexity and smoothness of each component imply

$$\begin{aligned}\mathbb{E}_I\,\|\nabla\phi_I(\theta)\|^2 &\leq 2\mathbb{E}_I\,\|\nabla\phi_I(\theta) - \nabla\phi_I(\theta^\star)\|^2 + 2\sigma_\star^2 \\ &\leq 4L_{\max}\{\Phi(\theta) - \Phi(\theta^\star)\} + 2\sigma_\star^2,\end{aligned}$$

where the linear terms at $\theta^\star$ cancel after averaging. This proves (2.6.11).

To expose the telescoping structure, set $\lambda_t = t/2$, $z_{-1} = \theta_0$, and rewrite (2.6.12) as

$$z_t = z_{t-1} - \eta\nabla\phi_{I_t}(\theta_t), \qquad \theta_{t+1} = \frac{\lambda_{t+1}}{\lambda_{t+1}+1}\theta_t + \frac{1}{\lambda_{t+1}+1}z_t. \tag{2.6.17}$$

A direct substitution verifies the equivalence. Write $\Delta_t := \Phi(\theta_t) - \Phi(\theta^\star)$. Expanding the squared distance of z_t, conditioning on $\mathcal{H}_t$, and using convexity together with (2.6.11) gives

$$\mathbb{E}[\|z_t - \theta^\star\|^2 \mid \mathcal{H}_t] \leq \|z_{t-1} - \theta^\star\|^2 - 2\eta\lambda_{t+1}\Delta_t + 2\eta\lambda_t\Delta_{t-1} + 2\eta^2\sigma_\star^2.$$

The term multiplied by $\lambda_0 = 0$ is omitted when $t = 0$. The step-size restriction is used through $1+\lambda_t-2\eta L_{\max} \geq \lambda_{t+1}$. Taking total expectations and summing from $t = 0$ to T cancels all intermediate objective gaps and yields

$$\mathbb{E}\,\|z_T - \theta^\star\|^2 \leq R^2 - \eta(T+1)\mathbb{E}\Delta_T + 2(T+1)\eta^2\sigma_\star^2.$$

Drop the nonnegative left-hand side to obtain (2.6.14). Strong convexity gives $\Delta_T \geq (\mu/2)\,\|\theta_T - \theta^\star\|^2$, proving (2.6.15). The final choice of η balances the two terms. This argument follows the iterate-moving-average analysis in [4]. □

Although the objective is strongly convex here, the stochastic theorem uses strong convexity only to convert objective error into parameter error. Its $O(T^{-1/2})$ noisy rate is therefore not a strong-convex acceleration result. When $\sigma_\star = 0$, the noise-floor term vanishes and the stated schedule gives the final-iterate rate $O(1/T)$.

The GD row is a global smooth strongly convex theorem. The sharper heavy-ball factor is an exact quadratic statement and should be compared with the exact quadratic GD factor $q_{\mathrm{GD}} = (\kappa-1)/(\kappa+1)$. The merely convex rates remain in Table 2.2.

Table 2.3: Strongly convex, smooth convergence guarantees obtained through Section 2.6.

Method and output	Additional structure or tuning	Guarantee	Interpretation
GD, final	$\eta = 1/\beta$	$\Delta_T = O((1-1/\kappa)^T)$	Global geometric objective convergence
SGD, final, constant step	Unbiased noise; fixed η	$\mathbb{E}\Delta_T = O(e^{-\mu\eta T}) + O(\eta\sigma^2)$	Geometric transient plus noise floor
SGD, final or PR average	Unbiased noise; $\eta_t \asymp 1/t$	$\mathbb{E}\Delta_T = O(T^{-1})$, $\mathbb{E}\left\|e_T\right\|^2 = O(T^{-1})$	Exact convergence in expectation
Heavy-ball, final	Deterministic strongly convex quadratic; optimal η, ρ	$\limsup_T \left\|e_T\right\|^{1/T} \le q_{\mathrm{HB}} = (\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$	Accelerated quadratic root factor
Stochastic momentum, final	Convex smooth components; schedule in (2.6.13)	$\mathbb{E}\Delta_T = O((\eta T)^{-1}) + O(\eta\sigma_\star^2)$; horizon tuning gives $O(T^{-1/2})$	$O(T^{-1})$ under interpolation $\sigma_\star = 0$

2.7 Adam under strong convexity and smoothness

Continue to assume that Φ is μ-strongly convex and β-smooth, and let g_t be a stochastic gradient. Adam combines a first-moment exponential average with a coordinatewise second-moment average. Initialize $m_0 = v_0 = 0$, choose $0 \le \beta_1, \beta_2 < 1$, $\varepsilon_{\mathrm{adam}} > 0$, and base step sizes $\alpha_t > 0$. For $t \ge 1$, set

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \qquad v_t = \beta_2 v_{t-1} + (1-\beta_2)(g_t \odot g_t), \tag{2.7.1}$$

$$\widehat{m}_t = \frac{m_t}{1-\beta_1^t}, \qquad \widehat{v}_t = \frac{v_t}{1-\beta_2^t}, \tag{2.7.2}$$

$$D_t := \operatorname{diag}\left(\frac{1}{\sqrt{\widehat{v}_t} + \varepsilon_{\mathrm{adam}}}\right), \qquad \theta_{t+1} = \theta_t - \alpha_t D_t \widehat{m}_t. \tag{2.7.3}$$

Products, square roots, and division are coordinatewise. Bias correction removes the zero-initialization factor from the two exponential averages; D_t then assigns a different effective learning rate to each coordinate [7].

Smoothness gives the pathwise inequality

$$\Phi(\theta_{t+1}) \le \Phi(\theta_t) - \alpha_t \langle \nabla\Phi(\theta_t), D_t\widehat{m}_t\rangle + \frac{\beta\alpha_t^2}{2}\left\|D_t\widehat{m}_t\right\|^2. \tag{2.7.4}$$

For SGD, conditional unbiasedness turns the alignment term into a squared gradient norm. For Adam, D_t and $\widehat{m}_t$ depend on the same current and past gradients, so the analogous simplification is not available.

The following deterministic benchmark isolates what a stable diagonal preconditioner would provide under our geometric assumptions.

Proposition 2.24 (A frozen-preconditioner benchmark)**.** *Suppose D is a fixed positive-definite diagonal matrix satisfying*

$$d_- I \preceq D \preceq d_+ I, \qquad 0 < d_- \le d_+, \tag{2.7.5}$$

and consider $\theta_{t+1} = \theta_t - \alpha D\nabla\Phi(\theta_t)$. If $0 < \alpha \le 1/(\beta d_+)$, then

$$\Phi(\theta_t) - \Phi(\theta^\star) \le (1 - \alpha\mu d_-)^t\{\Phi(\theta_0) - \Phi(\theta^\star)\}. \tag{2.7.6}$$

Proof. Because $D^2 \preceq d_+ D$, smoothness gives

$$\begin{aligned}\Phi(\theta_{t+1}) &\le \Phi(\theta_t) - \alpha\nabla\Phi(\theta_t)^\top D\nabla\Phi(\theta_t) + \frac{\beta\alpha^2}{2}\nabla\Phi(\theta_t)^\top D^2\nabla\Phi(\theta_t)\\ &\le \Phi(\theta_t) - \frac{\alpha}{2}d_-\left\|\nabla\Phi(\theta_t)\right\|^2.\end{aligned}$$

Strong convexity implies $\left\|\nabla\Phi(\theta_t)\right\|^2 \ge 2\mu\{\Phi(\theta_t) - \Phi(\theta^\star)\}$. Substitute this inequality and iterate the resulting contraction. □

Actual Adam does not have a frozen preconditioner: D_t is random, time-varying, and coupled to the momentum term. Strong convexity and smoothness provide a unique optimum and gradient domination, but they do not by themselves control the two adaptive quantities in (2.7.4). Broad convergence claims for vanilla Adam require care; nonconvergent examples exist even in simple convex online settings [14]. A valid stochastic theorem needs additional conditions controlling the effective coordinatewise steps and the noise, such as bounded preconditioners, a decreasing base step, and often a long-memory variant such as AMSGrad. Accordingly, this section derives the strongly convex, smooth benchmark but does not claim that vanilla Adam inherits its linear rate.

Table 2.4: Strongly convex, smooth convergence comparison through Section 2.7.

Method and output	Extra conditions or tuning	Guarantee	Status
GD, final	$\eta = 1/\beta$	$\Delta_T = O((1-1/\kappa)^T)$	Global theorem
SGD, final, constant step	Unbiased noise; fixed η	$\mathbb{E}\Delta_T = O(e^{-\mu\eta T}) + O(\eta\sigma^2)$	Noise floor
SGD, final or PR average	Unbiased noise; $\eta_t \asymp 1/t$	$\mathbb{E}\Delta_T = O(T^{-1})$, $\mathbb{E}\,\|e_T\|^2 = O(T^{-1})$	Exact convergence in expectation
Heavy-ball, final	Deterministic quadratic; optimal η, ρ	Parameter root factor $q_{\mathrm{HB}} = (\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$	Exact quadratic theorem
Stochastic momentum, final	Convex smooth components; time-varying momentum	$O((\eta T)^{-1}) + O(\eta\sigma_\star^2)$; horizon tuning gives $O(T^{-1/2})$	$O(T^{-1})$ if $\sigma_\star = 0$
Fixed diagonal preconditioner, final	$d_- I \preceq D \preceq d_+ I$, $\alpha \le 1/(\beta d_+)$	$\Delta_T \le (1-\alpha\mu d_-)^T \Delta_0$	Deterministic Adam benchmark
Vanilla Adam, final	Only strong convexity and smoothness	No general rate from these assumptions alone	Adaptive quantities need further control

The final row is deliberately a non-result: the geometric rate immediately above belongs to the frozen-preconditioner benchmark, not automatically to Adam's random, time-varying preconditioner. Convex-only rates are recorded in Table 2.2.

2.9 Summary

1. Least-squares gradient descent is an exactly solvable linear recursion. When A has full column rank, the problem is strongly convex and the eigenvalues of $A^\top A$ determine the entire stable step-size interval and the exact geometric rate. Rank-deficient least squares is convex, but not strongly convex in the parameters.
2. Smooth strongly convex GD converges geometrically at a rate governed by the condition number β/μ. Without strong convexity, the corresponding smooth convex guarantee is $O(1/T)$ in objective gap.
3. Under smooth strong convexity, constant-step SGD has a geometrically decaying transient but an $O(\eta)$ noise floor. A decreasing $1/t$-scale step size removes that floor and gives an $O(1/T)$ final-iterate objective guarantee and $O(1/T)$ mean-square parameter error. Arithmetic Polyak averaging also has an $O(1/T)$ expected objective rate with smoothness. Without smoothness, its general rate is $O(\log T/T)$, although linearly weighted averaging recovers $O(1/T)$. Without strong convexity, convex SGD gives the slower $O(1/\sqrt{T})$ rate with a horizon-tuned constant step. General step-size-weighted averaging also handles decreasing steps; in particular, $\eta_t \asymp t^{-1/2}$ gives $O(\log T/\sqrt{T})$ under the elementary bound used here.
4. Near a smooth minimizer, final-iterate SGD with step size a/t is asymptotically normal with a gain-dependent Lyapunov covariance. Polyak–Ruppert averaging gives the covariance-efficient sandwich limit $H^{-1}S(H^{-1})^\top$.
5. Deterministic heavy-ball momentum is exactly analyzable on strongly convex quadratics: its optimal asymptotic factor depends on $\sqrt{\kappa}$, rather than κ. For stochastic momentum, the displayed final-iterate guarantee separates an optimization transient from a noise floor. Adam adds a data-dependent diagonal preconditioner to a momentum-like gradient average; strong convexity and smoothness alone do not make that adaptive preconditioner behave like a fixed matrix.

Appendix A

Mathematical background

A.1 Probability review

This section is a compact reference for the probability notation used in the course. It follows the sequence most useful for asymptotic statistics: convergence modes, the LLN and CLT, continuous mappings and Slutsky's theorem, tightness, stochastic orders, and the delta method. Unless stated otherwise, random vectors take values in $\mathbb{R}^d$, $\|\cdot\|$ is the Euclidean norm, and $X, X_1, X_2, \ldots$ are defined on a common probability space whenever the definition compares their sample paths.

Modes of convergence

Definition A.1 (Four modes of convergence)**.** Let $p > 0$.

(a) *Convergence in probability*, written $X_n \xrightarrow{\mathbb{P}} X$, means that, for every $\varepsilon > 0$,

$$\mathbb{P}(\|X_n - X\| > \varepsilon) \longrightarrow 0.$$

(b) *Convergence in distribution*, also called *convergence in law* or *weak convergence*, is written $X_n \xrightarrow{d} X$. For scalar random variables it means

$$F_{X_n}(x) \longrightarrow F_X(x)$$

at every continuity point x of F_X. For random vectors, the same definition uses the joint cdf and componentwise inequalities.

(c) *Convergence in p-th moment*, or *in L^p*, written $X_n \xrightarrow{L^p} X$, means

$$\mathbb{E}\,\|X_n - X\|^p \longrightarrow 0.$$

(d) *Almost-sure convergence*, written $X_n \xrightarrow{\text{a.s.}} X$, means

$$\mathbb{P}\Big(\lim_{n\to\infty} \|X_n - X\| = 0\Big) = 1.$$

Almost-sure convergence concerns entire sample paths. Convergence in probability and in moment also depend on the joint construction of X_n and X. Convergence in distribution is weaker: it compares only their laws, so the variables may even be defined on different probability spaces.

Proposition A.2 (Implications between modes)**.** *For $p > 0$,*

$$X_n \xrightarrow{\text{a.s.}} X \quad\Longrightarrow\quad X_n \xrightarrow{\mathbb{P}} X \quad\Longrightarrow\quad X_n \xrightarrow{d} X,$$

and

$$X_n \xrightarrow{L^p} X \quad\Longrightarrow\quad X_n \xrightarrow{\mathbb{P}} X.$$

If $p \geq q > 0$, then L^p-convergence also implies L^q-convergence.

Proof. Almost-sure convergence implies convergence in probability by applying bounded convergence to $\mathbf{1}\{\|X_n - X\| > \varepsilon\}$. Markov's inequality, $\mathbb{P}(\|X_n - X\| > \varepsilon) \leq \varepsilon^{-p}\mathbb{E}\,\|X_n - X\|^p$, proves the moment implication. The implication from probability to distribution follows from the subsequence characterization and bounded convergence. Finally, monotonicity of L^r norms on a probability space gives $\{\mathbb{E}\,\|X_n - X\|^q\}^{1/q} \leq \{\mathbb{E}\,\|X_n - X\|^p\}^{1/p}$. □

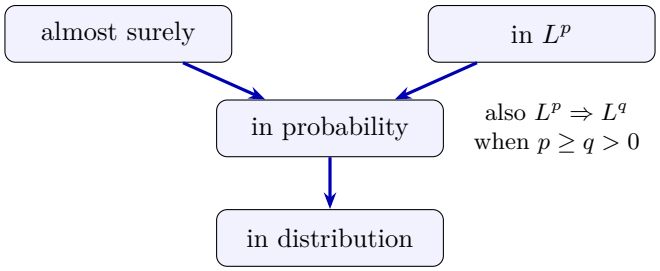

Figure A.1: The standard implications. In general, none of the displayed arrows can be reversed. Almost-sure and moment convergence are not ordered without additional assumptions.

The missing reverse arrows are genuine. The following examples are useful tests against overclaiming.

(i) Let X be Rademacher and set $X_n = -X$. Then $X_n \xrightarrow{d} X$, because X_n and X have the same law, but $\mathbb{P}(|X_n - X| > 1) = 1$. Thus distributional convergence need not imply convergence in probability.

(ii) Let A_n be independent with $\mathbb{P}(A_n) = 1/n$, and set $X_n = \mathbf{1}_{A_n}$. Then $X_n \to 0$ in every L^p, and hence in probability, but the second Borel–Cantelli lemma gives $X_n = 1$ infinitely often almost surely. Thus neither convergence in probability nor convergence in moment implies almost-sure convergence.

(iii) For any fixed $p > 0$, if $U \sim \mathrm{Unif}(0, 1)$, set $X_n = n^{1/p}\mathbf{1}\{U \leq 1/n\}$. Then $X_n \to 0$ almost surely, but $\mathbb{E}|X_n|^p = 1$. Thus almost-sure convergence need not imply L^p-convergence.

Laws of large numbers and the central limit theorem

Let $X_1, X_2, \ldots$ be iid scalar random variables and $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$.

Theorem A.3 (Weak law of large numbers)**.** *If $\mathbb{E}|X_1| < \infty$ and $\mu := \mathbb{E}X_1$, then*

$$\overline{X}_n \xrightarrow{\mathbb{P}} \mu.$$

Theorem A.4 (Strong law of large numbers)**.** *Under the same iid and integrability assumptions,*

$$\overline{X}_n \xrightarrow{\text{a.s.}} \mu.$$

The SLLN therefore implies the WLLN, although the names refer to the modes of convergence rather than to different limits.

Theorem A.5 (Central limit theorem)**.** *Let $X_i \in \mathbb{R}^d$ be iid with mean μ and finite covariance matrix Σ. Then*

$$\sqrt{n}\,(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma). \tag{A.1.1}$$

For $d = 1$, this reads $\sqrt{n}(\overline{X}_n - \mu)/\sigma \xrightarrow{d} \mathcal{N}(0, 1)$ when $0 < \sigma^2 := \mathrm{Var}(X_1) < \infty$.

The laws of large numbers identify the probability limit of the unscaled average. The CLT studies the remaining fluctuations on the finer $n^{-1/2}$ scale. It does not say that $\overline{X}_n - \mu$ converges to a nondegenerate normal random variable without the $\sqrt{n}$ rescaling.

Continuous mappings and Slutsky's theorem

Theorem A.6 (Continuous mapping theorem)**.** *Let $g : \mathbb{R}^d \to \mathbb{R}^k$ be measurable and continuous on a set C such that $\mathbb{P}(X \in C) = 1$. Then*

$$\begin{aligned} X_n \xrightarrow{\mathbb{P}} X \quad &\Longrightarrow \quad g(X_n) \xrightarrow{\mathbb{P}} g(X), \\ X_n \xrightarrow{\text{a.s.}} X \quad &\Longrightarrow \quad g(X_n) \xrightarrow{\text{a.s.}} g(X), \\ X_n \xrightarrow{d} X \quad &\Longrightarrow \quad g(X_n) \xrightarrow{d} g(X). \end{aligned}$$

The condition is continuity at the values that the limit can actually take; global continuity is sufficient but not necessary. For instance, squaring, matrix multiplication, and inversion away from singular matrices can all be handled by this theorem.

Theorem A.7 (Slutsky's theorem)**.** *The following three forms will be used repeatedly.*

(i) If $X_n \xrightarrow{d} c$*, where* c *is constant, then* $X_n \xrightarrow{\mathbb{P}} c$*.*

(ii) If $X_n \xrightarrow{d} X$ *and* $\|X_n - Y_n\| \xrightarrow{\mathbb{P}} 0$*, then* $Y_n \xrightarrow{d} X$*.*

(iii) If $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{\mathbb{P}} c$*, where* c *is constant, then*

$$(X_n, Y_n) \xrightarrow{d} (X, c).$$

Combining part (iii) with the continuous mapping theorem gives, whenever the operations are well defined,

$$X_n + Y_n \xrightarrow{d} X + c, \qquad X_n Y_n \xrightarrow{d} cX, \qquad \frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c} \quad (c \neq 0).$$

Vector–matrix versions follow by applying the relevant continuous map. The constant-limit condition matters: knowing only the two marginal limits does not identify the joint limit. For example, if $X_n \sim \mathcal{N}(0,1)$ and $Y_n = -X_n$, both marginals are standard normal, but (X_n, Y_n) is supported on the line $y = -x$, not distributed as two independent standard normals.

For example, if the scalar CLT holds and $\widehat{\sigma}_n \xrightarrow{\mathbb{P}} \sigma > 0$, then

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\widehat{\sigma}_n} \xrightarrow{d} \mathcal{N}(0,1).$$

Replacing an unknown constant by a consistent estimator is one of the most common uses of Slutsky's theorem.

Tightness and escaping mass

Definition A.8 (Uniform tightness)**.** A collection $\{X_\alpha : \alpha \in A\}$ in $\mathbb{R}^d$ is *uniformly tight* if, for every $\varepsilon > 0$, there is a compact set $K_\varepsilon \subset \mathbb{R}^d$ such that

$$\sup_{\alpha \in A} \mathbb{P}(X_\alpha \notin K_\varepsilon) < \varepsilon.$$

In $\mathbb{R}^d$, this is equivalent to requiring some finite M such that $\sup_{\alpha \in A} \mathbb{P}(\|X_\alpha\| > M) < \varepsilon$. A sequence is called *tight* when its associated collection is uniformly tight.

Proposition A.9 (Weak convergence implies tightness)**.** *If* $X_n \xrightarrow{d} X$*, then* (X_n) *is tight.*

Proof. Given $\varepsilon > 0$, choose a continuity point M_0 of the cdf of $\|X\|$ such that $\mathbb{P}(\|X\| > M_0) < \varepsilon/2$. The continuous mapping theorem and the scalar definition of weak convergence give $\mathbb{P}(\|X_n\| > M_0) \to \mathbb{P}(\|X\| > M_0)$, so the desired bound holds for all sufficiently large n. Enlarging M_0 to control the finitely many remaining random variables proves the claim. □

Theorem A.10 (Prokhorov's theorem in $\mathbb{R}^d$)**.** *If* (X_n) *is tight, then every subsequence contains a further subsequence that converges in distribution to some* $\mathbb{R}^d$*-valued random vector.*

Thus tightness prevents mass from escaping and supplies candidate subsequential limits; a separate argument must still identify the limit and prove its uniqueness.

Example A.11 (Why tightness matters: mass escaping to infinity)**:** Let $X_n = n$ deterministically. For any compact $K \subset \mathbb{R}$, all sufficiently large n satisfy $\mathbb{P}(X_n \notin K) = 1$, so the sequence is not tight and has no weakly convergent subsequence on $\mathbb{R}$. Its cdf is

$$F_n(x) = \mathbf{1}\{x \geq n\}.$$

For every fixed x, $F_n(x) \to 0$. The pointwise limit is not a cdf, because it does not approach one as $x \to \infty$. Thus pointwise-looking limits can lose all probability mass when tightness is absent.

Proposition A.12 (A convenient moment criterion)**.** *If, for some $p > 0$, $\sup_{\alpha \in A} \mathbb{E}\,\|X_\alpha\|^p \le C < \infty$, then $\{X_\alpha : \alpha \in A\}$ is uniformly tight.*

Proof. Markov's inequality gives

$$\sup_{\alpha \in A} \mathbb{P}(\|X_\alpha\| > M) \le \frac{C}{M^p} \longrightarrow 0 \qquad (M \to \infty).$$

□

A standard weak-convergence proof consequently has three steps:

1. prove tightness;
2. take an arbitrary convergent subsequence;
3. identify its limit and show that every subsequence has the same one.

Stochastic orders $o_{\mathbb{P}}$ and $O_{\mathbb{P}}$

Definition A.13 (Stochastic order)**.** Let $a_n > 0$ be deterministic.

(a) $X_n = O_{\mathbb{P}}(a_n)$ means that X_n/a_n is tight. Equivalently, for every $\varepsilon > 0$, there are finite M and N such that

$$\mathbb{P}(\|X_n\| > M a_n) < \varepsilon, \qquad n \ge N.$$

(b) $X_n = o_{\mathbb{P}}(a_n)$ means $X_n/a_n \xrightarrow{\mathbb{P}} 0$.

Thus $O_{\mathbb{P}}(1)$ means bounded in probability and $o_{\mathbb{P}}(1)$ means convergence to zero in probability.

Proposition A.14 (Calculation rules)**.** *Let $a_n, b_n > 0$, and let products below be scalar or dimensionally compatible matrix–vector products.*

$$\begin{aligned}
X_n = O_{\mathbb{P}}(a_n),\ Y_n = O_{\mathbb{P}}(b_n) \quad &\Longrightarrow \quad X_n + Y_n = O_{\mathbb{P}}(a_n + b_n),\\
X_n = o_{\mathbb{P}}(a_n),\ Y_n = o_{\mathbb{P}}(b_n) \quad &\Longrightarrow \quad X_n + Y_n = o_{\mathbb{P}}(a_n + b_n),\\
X_n = O_{\mathbb{P}}(a_n),\ Y_n = O_{\mathbb{P}}(b_n) \quad &\Longrightarrow \quad X_n Y_n = O_{\mathbb{P}}(a_n b_n),\\
X_n = o_{\mathbb{P}}(a_n),\ Y_n = O_{\mathbb{P}}(b_n) \quad &\Longrightarrow \quad X_n Y_n = o_{\mathbb{P}}(a_n b_n).
\end{aligned}$$

Also, $X_n = o_{\mathbb{P}}(a_n)$ implies $X_n = O_{\mathbb{P}}(a_n)$. If $a_n = O(b_n)$, then $O_{\mathbb{P}}(a_n)$ is also $O_{\mathbb{P}}(b_n)$; if $a_n = o(b_n)$, then $O_{\mathbb{P}}(a_n)$ is $o_{\mathbb{P}}(b_n)$. Finally, if $X_n/a_n \xrightarrow{\mathbb{P}} c \neq 0$, then

$$X_n^{-1} = a_n^{-1}\{c^{-1} + o_{\mathbb{P}}(1)\}.$$

The last rule requires a nonzero probability limit; stochastic order alone does not justify taking reciprocals.

The sum rules follow from the triangle inequality and union bounds. The product rules follow by first restricting the $O_{\mathbb{P}}(1)$ factor to a high-probability bounded set. The inverse rule is the continuous mapping theorem applied away from zero.

Lemma A.15 (Continuous transformations of small remainders)**.** *Let $R : \mathbb{R}^d \to \mathbb{R}^k$ be measurable, let $q > 0$, and suppose $X_n \xrightarrow{\mathbb{P}} 0$. Define the ratio below to be zero when $X_n = 0$.*

(i) If $\|R(h)\| = o(\|h\|^q)$ as $h \to 0$, then

$$\frac{\|R(X_n)\|}{\|X_n\|^q} \xrightarrow{\mathbb{P}} 0.$$

(ii) If $\|R(h)\| = O(\|h\|^q)$ as $h \to 0$, then

$$\frac{\|R(X_n)\|}{\|X_n\|^q} = O_{\mathbb{P}}(1).$$

Proof. For $h \neq 0$, set $G(h) = \|R(h)\| \,/\, \|h\|^q$, with $G(0) = 0$. In case (i), G is continuous at zero and the continuous mapping theorem gives $G(X_n) \xrightarrow{\mathbb{P}} 0$. In case (ii), G is bounded in a neighborhood of zero; since $X_n \xrightarrow{\mathbb{P}} 0$, this gives $G(X_n) = O_{\mathbb{P}}(1)$. □

Example A.16 (Reading a CLT as a stochastic rate)**:** If $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} Z$, then the scaled sequence is tight, so

$$\widehat{\theta}_n - \theta = O_{\mathbb{P}}(n^{-1/2}).$$

If an expansion additionally contains $r_n = o_{\mathbb{P}}(n^{-1/2})$, then

$$\sqrt{n}\{\widehat{\theta}_n - \theta + r_n\} = \sqrt{n}(\widehat{\theta}_n - \theta) + o_{\mathbb{P}}(1),$$

and Slutsky's theorem shows that the limiting distribution is unchanged.

The delta method

Suppose T_n estimates θ and $r_n(T_n - \theta) \xrightarrow{d} Z$. The delta method answers the natural question: what happens to the same asymptotic approximation after applying a smooth transformation to T_n?

Theorem A.17 (Delta method)**.** *Suppose $r_n \to \infty$ and*

$$r_n(T_n - \theta) \xrightarrow{d} Z, \qquad T_n, \theta \in \mathbb{R}^d.$$

If $g : \mathbb{R}^d \to \mathbb{R}^k$ is differentiable at θ, with Jacobian $G = Dg(\theta)$, then

$$r_n\{g(T_n) - g(\theta)\} \xrightarrow{d} GZ. \tag{A.1.2}$$

Moreover, the delta method gives the asymptotic linearization

$$r_n\{g(T_n) - g(\theta)\} - G\, r_n(T_n - \theta) \xrightarrow{\mathbb{P}} 0. \tag{A.1.3}$$

In particular, if $\sqrt{n}(T_n - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma)$, then

$$\sqrt{n}\{g(T_n) - g(\theta)\} \xrightarrow{d} \mathcal{N}(0, G\Sigma G^\top). \tag{A.1.4}$$

Proof. Convergence in distribution implies $r_n(T_n - \theta) = O_{\mathbb{P}}(1)$, hence $T_n - \theta = o_{\mathbb{P}}(1)$. Differentiability gives

$$g(T_n) - g(\theta) = G(T_n - \theta) + R_n, \qquad \frac{\|R_n\|}{\|T_n - \theta\|} \xrightarrow{\mathbb{P}} 0.$$

The stochastic-order product rule yields $r_n R_n = o_{\mathbb{P}}(1)$, which is (A.1.3). Equation (A.1.2) now follows from Slutsky's theorem; the normal case is the linear transformation of a Gaussian vector. □

Example A.18 (Logarithm of a sample mean)**:** Suppose $\mu > 0$ and $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$. Since $g(x) = \log x$ has $g'(\mu) = 1/\mu$,

$$\sqrt{n}\{\log \overline{X}_n - \log \mu\} \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2}{\mu^2}\right).$$

Because $\overline{X}_n \xrightarrow{\mathbb{P}} \mu > 0$, the logarithm is defined with probability tending to one.

If $Dg(\theta) = 0$, the first-order delta method gives a degenerate limit and a second-order scaling may be needed. For example, if $\sqrt{n}T_n \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, then the continuous mapping theorem gives

$$nT_n^2 \xrightarrow{d} \sigma^2 \chi_1^2.$$

This is the simplest second-order delta-method calculation.

Bibliographical notes

The population-risk and empirical-risk framework is standard; see Mohri, Rostamizadeh, and Talwalkar [9] and Shalev-Shwartz and Ben-David [16]. The model-class examples, capacity discussion, validation protocol, and estimator bias–variance material draw on Goodfellow, Bengio, and Courville [5], with notation adjusted to emphasize the repeated-sampling interpretation. The double-descent terminology and unified curve are due to Belkin et al. [1]; the ridgeless linear calculation, random-feature theory, and deep-network experiments are represented by [6, 8, 10]. The convex-optimization proofs follow the standard first-order analysis in Nesterov [11] and Bubeck [3].

Stochastic approximation begins with Robbins and Monro [15]. Classical trajectory averaging and its covariance-efficient asymptotic normality, often called Polyak–Ruppert averaging, are treated by Polyak and Juditsky in [13]. Bottou, Curtis, and Nocedal [2] connect these ideas to large-scale machine learning. Deterministic heavy-ball momentum originates with Polyak [12]. The final-iterate stochastic-momentum argument used here follows the iterate-moving-average analysis in the handbook of Garrigos and Gower [4].

Neural-network optimization and the practical role of momentum are developed in Goodfellow, Bengio, and Courville [5]; the Adam update is due to Kingma and Ba [7]. Reddi, Kale, and Kumar [14] give nonconvergence examples for Adam and introduce AMSGrad.

Bibliography

[1] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[2] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[3] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3–4):231–357, 2015.

[4] G. Garrigos and R. M. Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

[5] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[6] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

[7] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[8] S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

[9] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. MIT Press, second edition, 2018.

[10] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020.

[11] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

[12] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[13] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[14] S. J. Reddi, S. Kale, and S. Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

[15] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[16] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.