# Lectures 16–17 Stochastic gradient descent

Stochastic approximation · fluctuations and averaging

**Reading guide.** Section 1 is Lecture 16, Section 2 is Lecture 17. The tools are those of Lectures 12–15, cited as "Theorem 4.2 of Lectures 12–15" and so on: orthogonal increments, the $L^2$ convergence theorem, the martingale CLT, Freedman's inequality. Examples marked WORKED IN CLASS are done on the board. Exercises are typed A, B, C as before; the homework for a lecture is the exercise block closing its section.

## 1 Stochastic approximation

Stochastic gradient descent on a loss $L$ with minimizer $\theta^*$ is the recursion $\theta_{k+1} = \theta_k - \eta_k\, g(\theta_k; z_{k+1})$, where $g(\theta_k; z_{k+1})$ is an unbiased but noisy estimate of $\nabla L(\theta_k)$ computed from a fresh sample. The two lectures prove two theorems about it, under strong convexity and smoothness of $L$ and a variance bound on the noise.

- *A limit theorem* (Theorem 1.5). With step sizes $\eta_k = \eta/(k + k_0)$ and $\eta\mu > 1$,

$$\mathbb{E}\|\theta_n - \theta^*\|^2 \le \frac{C}{n},$$

  with an explicit $C$; with a constant step the error instead settles at a noise floor proportional to the step. This says how many samples a given accuracy costs: the mean-square error after $n$ samples is of the same order as that of an estimator that uses all $n$ samples at once, although each sample is looked at only once and then discarded.

- *A central limit theorem* (Theorems 2.4 and 2.7). At the scale $n^{-1/2}$ the error is Gaussian, $\sqrt{n}(\theta_n - \theta^*) \Rightarrow \mathcal{N}(0, \Sigma_\eta)$, with a covariance that depends on the step size $\eta$ and is smallest at a value of $\eta$ one cannot know in advance; averaging the iterates, $\bar\theta_n = \frac{1}{n}\sum_{k<n}\theta_k$, gives $\sqrt{n}(\bar\theta_n - \theta^*) \Rightarrow \mathcal{N}(0, H^{-1}\Sigma H^{-1})$ for any choice of step: the smallest covariance among all step sizes, and the covariance of the estimator that minimizes the empirical loss on all $n$ samples at once (the Cramér–Rao bound when the loss is a log-likelihood). This is what a confidence interval for $\theta^*$ is built from, and it is the reason to average.

Both are consequences of the fact, from Lecture 12, that the noise in the recursion is a martingale difference: the limit theorem is the $L^2$ theory of Lectures 12–14 applied to that martingale, the central limit theorem is the martingale CLT of Lecture 14, and the trajectory bounds at the end of Lecture 17 are Freedman's inequality from Lecture 15.

### 1.1 Setting

Let $L : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable. We assume:

(A1) *Strong convexity.* $\langle \nabla L(\theta) - \nabla L(\theta'), \theta - \theta' \rangle \ge \mu\|\theta - \theta'\|^2$ for some $\mu > 0$.

(A2) *Smoothness.* $\|\nabla L(\theta) - \nabla L(\theta')\| \le \beta\|\theta - \theta'\|$ for some $\beta \ge \mu$.

Under (A1) $L$ has exactly one minimizer $\theta^*$, and it is the only point where $\nabla L$ vanishes. Indeed, with $\theta_t = \theta + t(\theta' - \theta)$, the fundamental theorem of calculus along the segment and then (A1) at the pair $(\theta_t, \theta)$, for which $\theta_t - \theta = t(\theta' - \theta)$, give

$$L(\theta') - L(\theta) - \langle \nabla L(\theta), \theta' - \theta \rangle = \int_0^1 \langle \nabla L(\theta_t) - \nabla L(\theta), \theta' - \theta \rangle \, dt \;\geq\; \int_0^1 \mu t \, \|\theta' - \theta\|^2 \, dt = \tfrac{\mu}{2} \|\theta' - \theta\|^2,$$

so $L(\theta') \to \infty$ as $\|\theta'\| \to \infty$ and a minimizer exists (a continuous function on a large closed ball attains its minimum); at a minimizer the gradient is zero, and the same inequality with $\nabla L(\theta) = 0$ shows $L(\theta') > L(\theta)$ for every $\theta' \neq \theta$, so there is only one such point. Write $e = \theta - \theta^*$ for the error. Taking $\theta' = \theta^*$ in (A1)–(A2),

$$\langle \nabla L(\theta), e \rangle \geq \mu \|e\|^2, \qquad \|\nabla L(\theta)\| \leq \beta \|e\|. \tag{1}$$

The algorithm sees $L$ only through stochastic gradients. Let $z_1, z_2, \ldots$ be i.i.d. samples and $g(\theta; z) \in \mathbb{R}^d$ a function of the parameter and the sample with

(A3) *Unbiased gradients with controlled variance.* $\mathbb{E} g(\theta; z) = \nabla L(\theta)$ and $\mathbb{E} \|g(\theta; z) - \nabla L(\theta)\|^2 \leq \sigma^2 + \kappa \|\theta - \theta^*\|^2$ for every $\theta$, for constants $\sigma^2 \geq 0$, $\kappa \geq 0$.

The typical case is $L(\theta) = \mathbb{E}\, \ell(\theta; z)$ and $g = \nabla_\theta \ell$. The term $\kappa \|e\|^2$ allows the noise to grow with the distance from the minimizer, which it does in least squares (Example 1.6).

*Stochastic gradient descent* with step sizes $\eta_0, \eta_1, \cdots > 0$ is

$$\theta_{k+1} = \theta_k - \eta_k \, g(\theta_k; z_{k+1}), \qquad \theta_0 \text{ fixed.}$$

With $\mathcal{F}_k = (z_1, \ldots, z_k)$, $\theta_k \in L^2(\mathcal{F}_k)$ (coordinatewise; we assume throughout that the needed second moments are finite), and by the freezing lemma (Lemma 1.5 of Lectures 12–15) with $U = \theta_k$ and $Z = z_{k+1}$,

$$\xi_{k+1} := g(\theta_k; z_{k+1}) - \nabla L(\theta_k) \quad \text{satisfies} \quad \mathbb{E}_k \xi_{k+1} = 0, \qquad s^2(\theta_k) := \mathbb{E}_k \|\xi_{k+1}\|^2 \leq \sigma^2 + \kappa \|e_k\|^2.$$

So, as in Example 2.4(d) of Lectures 12–15,

$$e_{k+1} = e_k - \eta_k \nabla L(\theta_k) - \eta_k \xi_{k+1}, \qquad M_n := \sum_{k<n} \eta_k \xi_{k+1} \text{ a martingale with respect to } (\mathcal{F}_n). \tag{2}$$

## 1.2 The one-step recursion

**Lemma 1.1** (one step)**.** *For every $k$,*

$$\mathbb{E}_k \|e_{k+1}\|^2 = \|e_k - \eta_k \nabla L(\theta_k)\|^2 + \eta_k^2 s^2(\theta_k) \;\leq\; \big(1 - 2\mu\eta_k + (\beta^2 + \kappa)\eta_k^2\big) \|e_k\|^2 + \eta_k^2 \sigma^2.$$

*In particular, if $\eta_k \leq \mu/(\beta^2 + \kappa)$, then $\mathbb{E}_k \|e_{k+1}\|^2 \leq (1 - \mu\eta_k) \|e_k\|^2 + \eta_k^2 \sigma^2$.*

*Proof.* Expand $\|e_{k+1}\|^2 = \|e_k - \eta_k \nabla L(\theta_k)\|^2 - 2\eta_k \langle e_k - \eta_k \nabla L(\theta_k), \xi_{k+1} \rangle + \eta_k^2 \|\xi_{k+1}\|^2$. The vector $e_k - \eta_k \nabla L(\theta_k)$ is in $L^2(\mathcal{F}_k)$, so by rule (iii) the cross term has $\mathbb{E}_k$ equal to zero, and $\mathbb{E}_k \|\xi_{k+1}\|^2 = s^2(\theta_k)$. For the inequality, by (1),

$$\|e_k - \eta_k \nabla L(\theta_k)\|^2 = \|e_k\|^2 - 2\eta_k \langle \nabla L(\theta_k), e_k \rangle + \eta_k^2 \|\nabla L(\theta_k)\|^2 \leq (1 - 2\mu\eta_k + \beta^2 \eta_k^2) \|e_k\|^2,$$

and (A3) bounds $\eta_k^2 s^2(\theta_k) \leq \eta_k^2 \sigma^2 + \eta_k^2 \kappa \|e_k\|^2$. The last claim uses $(\beta^2 + \kappa)\eta_k \leq \mu$. □

Taking expectations, $a_k := \mathbb{E}\|e_k\|^2$ satisfies, for $\eta_k \le \mu/(\beta^2+\kappa)$,

$$a_{k+1} \le (1-\mu\eta_k)\, a_k + \eta_k^2\sigma^2. \tag{3}$$

Unrolling (3) (induction on $n$) gives the explicit bound

$$a_n \;\le\; a_0 \prod_{k<n}(1-\mu\eta_k) \;+\; \sigma^2 \sum_{k<n} \eta_k^2 \prod_{j=k+1}^{n-1} (1-\mu\eta_j), \tag{4}$$

with the usual convention that an empty product is 1. The first term is the *bias*: the initial error, contracted by the factor $1-\mu\eta_k$ at each step, so by at most $\exp(-\mu\sum_{k<n}\eta_k)$ in total. The second is the *variance*: the noise $\eta_k^2\sigma^2$ injected at step $k$, contracted by every step after it. Everything in this lecture is read off (4).

### 1.3 Constant step size

**Theorem 1.2** (constant step)**.** *If $\eta_k \equiv \eta \le \mu/(\beta^2+\kappa)$, then for every $n$,*

$$\mathbb{E}\|\theta_n - \theta^*\|^2 \le (1-\mu\eta)^n \|\theta_0-\theta^*\|^2 + \frac{\eta\sigma^2}{\mu}.$$

*Proof.* In (4) the products are powers of $1-\mu\eta$: $a_n \le (1-\mu\eta)^n a_0 + \eta^2\sigma^2\sum_{j<n}(1-\mu\eta)^j \le (1-\mu\eta)^n a_0 + \eta^2\sigma^2/(\mu\eta)$. □

The initial error is forgotten exponentially fast, at rate $\mu\eta$ per step, and the iterates then fluctuate at the *noise floor* $\eta\sigma^2/\mu$. Halving the step halves the floor and doubles the time to reach it; decreasing step sizes are the way to have both.

*Remark* 1.3 (the floor is real). Suppose, in addition to (A1)–(A3), that the noise does not vanish anywhere: $s^2(\theta) \ge \sigma_-^2 > 0$ for all $\theta$. Then for a constant step $\eta \le \min(1/\beta, \mu/(\beta^2+\kappa))$,

$$\liminf_{n\to\infty} \mathbb{E}\|\theta_n-\theta^*\|^2 \;\ge\; \frac{\eta\,\sigma_-^2}{2\beta},$$

so Theorem 1.2 is sharp up to the factor $2\beta/\mu\cdot\sigma^2/\sigma_-^2$.

*Derivation.* Start from the identity in Lemma 1.1, before any inequality was used:

$$\mathbb{E}_k\|e_{k+1}\|^2 = \|e_k - \eta\nabla L(\theta_k)\|^2 + \eta^2 s^2(\theta_k).$$

*The drift term from below.* By the triangle inequality and (1), $\|e_k - \eta\nabla L(\theta_k)\| \ge \|e_k\| - \eta\|\nabla L(\theta_k)\| \ge (1-\beta\eta)\|e_k\|$, and the right side is $\ge 0$ because $\beta\eta \le 1$; squaring, $\|e_k-\eta\nabla L(\theta_k)\|^2 \ge (1-\beta\eta)^2\|e_k\|^2$. *The noise term from below.* $\eta^2 s^2(\theta_k) \ge \eta^2\sigma_-^2$ by hypothesis. *Take expectations.* With $a_k = \mathbb{E}\|e_k\|^2$ and $q := (1-\beta\eta)^2 \in [0,1)$,

$$a_{k+1} \;\ge\; q\, a_k + \eta^2\sigma_-^2 \qquad \text{for all } k \ge 0.$$

*Unroll.* Induction on $n$ (exactly as for (4), with the inequality reversed) gives

$$a_n \;\ge\; q^n a_0 + \eta^2\sigma_-^2 \sum_{j<n} q^j = q^n a_0 + \eta^2\sigma_-^2\,\frac{1-q^n}{1-q} \;\ge\; \eta^2\sigma_-^2\,\frac{1-q^n}{1-q}.$$

*Let $n \to \infty$.* Since $q < 1$, $q^n \to 0$ and $\liminf_n a_n \geq \eta^2\sigma_-^2/(1-q)$. Finally $1 - q = 1 - (1-\beta\eta)^2 = \beta\eta(2-\beta\eta) \leq 2\beta\eta$, so $\eta^2\sigma_-^2/(1-q) \geq \eta\sigma_-^2/(2\beta)$.

*Exactness for a quadratic.* For $L(\theta) = \frac{h}{2}(\theta - \theta^*)^2$ in $d = 1$ with noise of constant variance $\sigma^2$, both steps above are equalities: $\|e - \eta\nabla L\|^2 = (1-\eta h)^2 e^2$ and $s^2 \equiv \sigma^2$, so $a_{k+1} = (1-\eta h)^2 a_k + \eta^2\sigma^2$ and $a_n \to \eta^2\sigma^2/(1-(1-\eta h)^2) = \eta\sigma^2/(h(2-\eta h))$, which is Exercise 2.12(c) of Lectures 12–15 with $h = 1$. Here $\mu = \beta = h$ and the upper bound $\eta\sigma^2/\mu$ of Theorem 1.2 is off by the factor $2 - \eta h \in (1, 2]$.

*When the floor is absent.* The hypothesis $\sigma_- > 0$ cannot be dropped: if the noise vanishes at the minimizer, $s^2(\theta) \leq \kappa\|e\|^2$, then $a_{k+1} \leq (1-\mu\eta)a_k$ and $a_n \to 0$ at a constant step (Example 2.10).

## 1.4 Decreasing step sizes: the Robbins–Monro rate

**Lemma 1.4** (a recursion)**.** *Let $a_{k+1} \leq \big(1 - \frac{c}{k+k_0}\big)a_k + \frac{C}{(k+k_0)^2}$ for all $k \geq 0$, with $c > 1$, $C \geq 0$ and $k_0 \geq 1$. Then*

$$a_k \leq \frac{D}{k+k_0} \quad \text{for all } k \geq 0, \qquad D := \max\Big(\frac{C}{c-1},\ k_0 a_0\Big).$$

*Proof.* Induction. The case $k = 0$ holds since $D \geq k_0 a_0$. If $a_k \leq D/m$ with $m = k + k_0$, then $a_{k+1} \leq \frac{D(m-c)+C}{m^2}$, and $\frac{D(m-c)+C}{m^2} \leq \frac{D}{m+1}$ is equivalent to $m\big(C - D(c-1)\big) \leq Dc - C$, which holds because the left side is $\leq 0$ (as $D(c-1) \geq C$) and the right side is $\geq Dc - D(c-1) = D > 0$. □

**Theorem 1.5** (Robbins–Monro)**.** *Let $\eta_k = \dfrac{\eta}{k+k_0}$ with $\eta\mu > 1$ and $k_0 \geq \eta(\beta^2+\kappa)/\mu$. Then for every $n$,*

$$\mathbb{E}\|\theta_n - \theta^*\|^2 \leq \frac{1}{n+k_0}\max\Big(\frac{\eta^2\sigma^2}{\eta\mu - 1},\ k_0\|\theta_0 - \theta^*\|^2\Big).$$

*Proof.* The choice of $k_0$ gives $\eta_k \leq \eta/k_0 \leq \mu/(\beta^2+\kappa)$ for all $k$, so (3) holds and reads $a_{k+1} \leq (1 - \frac{\eta\mu}{k+k_0})a_k + \frac{\eta^2\sigma^2}{(k+k_0)^2}$. Apply Lemma 1.4 with $c = \eta\mu$, $C = \eta^2\sigma^2$. □

The same rate can be read off (4) directly: with $m = k + k_0$ and $N = n + k_0$, $\prod_{j=k+1}^{n-1}(1-\mu\eta_j) \leq \exp(-\eta\mu\sum_{j=k+1}^{n-1}\frac{1}{j+k_0}) \leq \big(\frac{m+1}{N}\big)^{\eta\mu}$, so the bias is $O(N^{-\eta\mu})$ and the variance term is at most $\sigma^2\eta^2 N^{-\eta\mu}\sum_{m<N} m^{-2}(m+1)^{\eta\mu} = O(N^{-1})$ when $\eta\mu > 1$, since $\sum_{m<N} m^{\eta\mu-2} \asymp N^{\eta\mu-1}$. Lemma 1.4 is the same computation with the constant made explicit.

The mean-square error decays like $1/n$, the rate of an average of $n$ independent observations, with constant $\eta^2\sigma^2/(\eta\mu - 1)$. Two features of the statement matter. The condition $\eta\mu > 1$ requires knowing a lower bound on the curvature $\mu$: if $\eta\mu < 1$ the recursion only gives the slower rate $n^{-\eta\mu}$ (Exercise 1.9). And the constant is minimized at $\eta = 2/\mu$, not at the largest admissible step. Lecture 17 refines both: the sharp threshold is $\eta\lambda_{\min}(\nabla^2 L(\theta^*)) > \frac{1}{2}$ when the recursion is linearized exactly, and averaging the iterates removes the need to know the curvature altogether.

**Example 1.6** (least mean squares)**.** WORKED IN CLASS Linear regression with Gaussian design: $z = (x, y)$ with $x \sim \mathcal{N}(0, I_d)$, $y = \langle x, \theta^*\rangle + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ independent of $x$, and the loss $\ell(\theta; z) = \frac{1}{2}(\langle x, \theta\rangle - y)^2$. Then

$$L(\theta) = \tfrac{1}{2}\|\theta - \theta^*\|^2 + \tfrac{1}{2}\sigma_\varepsilon^2, \qquad \nabla L(\theta) = e, \qquad g(\theta; z) = x\,(\langle x, e\rangle - \varepsilon),$$

so $\mu = \beta = 1$ and SGD is the *least mean squares* algorithm $\theta_{k+1} = \theta_k - \eta_k x_{k+1}(\langle x_{k+1}, \theta_k\rangle - y_{k+1})$. The gradient noise is $\xi = x\langle x, e\rangle - e - x\varepsilon$, and by Wick's theorem ($\mathbb{E}[(x^\top e)^2\|x\|^2] = (d+2)\|e\|^2$, $\mathbb{E}\|x\|^2 = d$),

$$\mathbb{E}\|\xi\|^2 = \mathbb{E}\big[(x^\top e)^2\|x\|^2\big] - \|e\|^2 + \sigma_\varepsilon^2\mathbb{E}\|x\|^2 = (d+1)\|e\|^2 + d\sigma_\varepsilon^2.$$

So (A3) holds with $\kappa = d+1$ and $\sigma^2 = d\sigma_\varepsilon^2$, and nothing smaller: the noise is *multiplicative*, proportional to the current error, and its size is set by the dimension. Lemma 1.1 requires $\eta_k \le 1/(d+2)$; with $\eta_k = \eta/(k+k_0)$, $\eta > 1$, $k_0 \ge \eta(d+2)$, Theorem 1.5 gives

$$\mathbb{E}\|\theta_n - \theta^*\|^2 \le \frac{1}{n+k_0}\max\Big(\frac{\eta^2 d\,\sigma_\varepsilon^2}{\eta-1},\, k_0\|\theta_0-\theta^*\|^2\Big) \approx \frac{\eta^2}{\eta-1}\cdot\frac{d\sigma_\varepsilon^2}{n} \quad \text{for large } n,$$

which at $\eta = 2$ is $4d\sigma_\varepsilon^2/n$: four times the mean-square error $d\sigma_\varepsilon^2/n$ of the least squares estimator on the same $n$ samples. Lecture 17 shows how to remove the factor 4.

## 1.5 The Robbins–Monro conditions

**Proposition 1.7.** *Suppose $\eta_k \le \mu/(\beta^2+\kappa)$ for all $k$, $\sum_k \eta_k = \infty$ and $\eta_k \to 0$. Then $\mathbb{E}\|\theta_n - \theta^*\|^2 \to 0$.*

*Proof.* From (3), $a_{k+1} \le (1-\mu\eta_k)a_k + \eta_k^2\sigma^2$. Given $\epsilon > 0$ choose $n_0$ with $\eta_k\sigma^2 \le \mu\epsilon$ for $k \ge n_0$; then $a_{k+1} - \epsilon \le (1-\mu\eta_k)(a_k - \epsilon)$ for $k \ge n_0$, so $(a_n - \epsilon)_+ \le (a_{n_0} - \epsilon)_+ \exp(-\mu\sum_{k=n_0}^{n-1}\eta_k) \to 0$. Hence $\limsup a_n \le \epsilon$. □

The conditions $\sum_k \eta_k = \infty$ and $\sum_k \eta_k^2 < \infty$ are Robbins and Monro's (1951). The first lets the drift reach $\theta^*$ from anywhere, and is all the mean-square statement needs. The second makes the accumulated noise $M_n = \sum_{k<n}\eta_k\xi_{k+1}$ a martingale that is bounded in $L^2$, hence convergent (Theorem 4.2 of Lectures 12–15); it is what controls the whole trajectory rather than a fixed time $n$, and Section 2.3 does this with Freedman's inequality.

## 1.6 The ODE picture

Write $t_k = \sum_{j<k}\eta_j$ for the accumulated step size. Then (2) is

$$\theta_{k+1} - \theta_k = -(t_{k+1} - t_k)\,\nabla L(\theta_k) - \eta_k\xi_{k+1}\,:$$

Euler's scheme for the gradient flow $\dot\theta = -\nabla L(\theta)$ with time steps $\eta_k$, perturbed by a martingale whose quadratic variation up to time $t_k$ is $\sum_{j<k}\eta_j^2 s^2(\theta_j)$. Under the Robbins–Monro conditions the ODE time $t_k$ goes to infinity while the noise has finite total quadratic variation: on the ODE's time scale the noise is negligible. For strongly convex $L$ the flow satisfies $\|e(t)\| \le e^{-\mu t}\|e(0)\|$; with $\eta_k = \eta/k$, $t_k \approx \eta\log k$, so the bias decays like $k^{-\mu\eta}$, which is the rate $n^{-\eta\mu}$ seen for $\eta\mu < 1$ and the reason the $1/n$ rate needs $\eta\mu > 1$: for the bias to be forgotten faster than the variance accumulates, the ODE must contract faster than $1/\sqrt{k}$ in $k$.

## Exercises for Lecture 16

**Exercise 1.8.** TYPE A Let $L(\theta) = \frac{1}{2}\theta^\top H\theta$ with $H$ symmetric, $\mu I \preceq H \preceq \beta I$. Verify (A1)–(A2) and show $\nabla L(\theta) = H\theta$. For $g(\theta; z) = H\theta + z$ with $z$ centered of covariance $\Sigma$, verify (A3) with $\kappa = 0$, $\sigma^2 = \operatorname{tr}\Sigma$, and compute $\mathbb{E}\|e_n\|^2$ exactly for constant step in the case $H = hI$.

**Exercise 1.9.** type B Let $a_{k+1} \le (1 - \frac{c}{k+k_0})a_k + \frac{C}{(k+k_0)^2}$ with $0 < c < 1$. Show $a_k \le D(k+k_0)^{-c}$ for a constant $D$ depending on $a_0, C, c, k_0$ (unroll as in (4); or do the induction, but with the ansatz $D(k+k_0)^{-c} - E(k+k_0)^{-1}$: why does $D(k+k_0)^{-c}$ alone not close?). Conclude that for $\eta_k = \eta/(k+k_0)$ with $\eta\mu < 1$, Theorem 1.5 is replaced by the rate $n^{-\eta\mu}$. For the quadratic of Exercise 1.8 in $d = 1$, compute the exact rate of $\mathbb{E}e_n^2$ when $\eta h < \frac{1}{2}$ and compare: is $n^{-\eta\mu}$ attained?

**Exercise 1.10** (mini-batches)**.** type B Replace $g(\theta; z)$ by the average of $B$ independent stochastic gradients. Show (A3) holds with $(\sigma^2, \kappa)$ replaced by $(\sigma^2/B, \kappa/B)$, and derive the corresponding versions of Theorems 1.2 and 1.5. In Theorem 1.2, compare, for a fixed budget of $nB$ gradient evaluations, the floor reached by batch size $B$ with step $\eta$ against batch size 1 with step $\eta/B$.

**Exercise 1.11** (step size $\eta_k \propto k^{-\alpha}$)**.** type B Let $\eta_k = \eta(k+k_0)^{-\alpha}$ with $\alpha \in (\frac{1}{2}, 1)$ and $\eta_k \le \mu/(\beta^2+\kappa)$. Show from (3) that $a_k \le D\,\eta_k$ for a constant $D$ and all $k$, provided $k_0$ is large enough (induction with $D \ge \max(2\sigma^2/\mu, a_0/\eta_0)$, using $\eta_{k+1}/\eta_k \ge 1 - \alpha/(k+k_0)$ and $2\alpha(k+k_0)^{\alpha-1} \le \eta\mu$), hence $\mathbb{E}\|e_k\|^2 = O(k^{-\alpha})$. This is the rate Lecture 17 needs for averaging.

**Exercise 1.12** (non-convex smooth losses)**.** type C Drop (A1) and assume only (A2), $L \ge L_*$, and (A3) with $\kappa = 0$. (a) Use the descent inequality $L(\theta') \le L(\theta) + \langle \nabla L(\theta), \theta' - \theta\rangle + \frac{\beta}{2}\|\theta' - \theta\|^2$ to show $\mathbb{E}_k L(\theta_{k+1}) \le L(\theta_k) - \eta_k(1 - \frac{\beta\eta_k}{2})\|\nabla L(\theta_k)\|^2 + \frac{\beta}{2}\eta_k^2\sigma^2$. (b) Sum, for $\eta_k \le 1/\beta$, to obtain

$$\min_{k<n} \mathbb{E}\|\nabla L(\theta_k)\|^2 \;\le\; \frac{2\big(L(\theta_0) - L_*\big) + \beta\sigma^2 \sum_{k<n}\eta_k^2}{\sum_{k<n}\eta_k}.$$

(c) Optimize over a constant step for a horizon $n$, and over $\eta_k \propto k^{-1/2}$. What does the bound say, and what does it not say, about where the iterates end up?

**Exercise 1.13** (LMS, simulated)**.** type C Implement Example 1.6 with $d = 20$, $\sigma_\varepsilon = 1$, $\theta^*$ a fixed unit vector, $\theta_0 = 0$. (a) Run $\eta_k = \eta/(k+k_0)$ for $\eta \in \{0.5, 1, 2, 4\}$ with $k_0 = \eta(d+2)$, $n = 10^4$, 200 repetitions, and plot $n\,\mathbb{E}\|\theta_n - \theta^*\|^2$ against $\eta$; compare with $\eta^2 d/(\eta - 1)$ and with the least squares value $d$. (b) Run the constant step $\eta = 1/(d+2)$ and locate the noise floor. (c) Run $\eta_k = \eta k^{-0.7}$ with the averaged iterate $\bar\theta_n = \frac{1}{n}\sum_{k<n}\theta_k$ and compare $n\mathbb{E}\|\bar\theta_n - \theta^*\|^2$ with $d$.

## 2 Fluctuations and averaging

Lecture 16 bounded the mean-square error. This lecture identifies the distribution of the error at scale $n^{-1/2}$, and shows that averaging the iterates attains the smallest limiting covariance available to SGD, without knowledge of the curvature. Everything is in $\mathbb{R}^d$. In addition to (A1)–(A3) we assume:

(A4) *Linearization.* $\nabla L(\theta) = H\,e + r(\theta)$ with $H = \nabla^2 L(\theta^*)$ symmetric, $\mu I \preceq H \preceq \beta I$, and $\|r(\theta)\| \le \frac{\gamma}{2}\|e\|^2$.

(A5) *Noise regularity.* $\mathbb{E}_k\|\xi_{k+1}\|^3 \le \rho^3$ for all $k$, and the conditional covariance $C(\theta) := \mathbb{E}\big[\xi\xi^\top\big]$, $\xi = g(\theta; z) - \nabla L(\theta)$, is continuous at $\theta^*$, with $\Sigma := C(\theta^*)$ positive definite.

With (A4) the recursion (2) becomes

$$e_{k+1} = (I - \eta_k H)\,e_k - \eta_k\xi_{k+1} - \eta_k r_k, \qquad r_k := r(\theta_k), \quad \|r_k\| \le \tfrac{\gamma}{2}\|e_k\|^2. \tag{5}$$

Note that $\operatorname{tr} C(\theta) = s^2(\theta) \le \sigma^2 + \kappa\|e\|^2$ by (A3), so $\|C(\theta)\|_{\mathrm{op}} \le \sigma^2 + \kappa\|e\|^2$ as well.

Two facts from outside the course are used to pass from one dimension to $d$.

QUOTED · convergence in distribution in $\mathbb{R}^d$

*Cramér–Wold.* Random vectors $X_n$ converge in distribution to $X$ in $\mathbb{R}^d$ if and only if $\langle v, X_n\rangle \Rightarrow \langle v, X\rangle$ for every fixed $v \in \mathbb{R}^d$. In particular $X_n \Rightarrow \mathcal{N}(0, S)$ if and only if $\langle v, X_n\rangle \Rightarrow \mathcal{N}(0, v^\top S v)$ for every $v$. *Slutsky.* If $X_n \Rightarrow X$ and $\|Y_n\| \to 0$ in probability then $X_n + Y_n \Rightarrow X$; and if $A$ is a fixed matrix then $AX_n \Rightarrow AX$.

The tool of the lecture is the martingale CLT of Lecture 14, in the following form.

**Theorem 2.1** (martingale CLT; Corollary 4.6 of Lectures 12–15)**.** *For each $n$ let $\zeta_{n,1}, \dots, \zeta_{n,k_n}$ be real martingale differences with respect to a filtration $(\mathcal{F}_i^{(n)})_{i\ge 0}$ (that is, $\zeta_{n,i} \in L^2(\mathcal{F}_i^{(n)})$ and $\mathbb{E}[\zeta_{n,i} \mid \mathcal{F}_{i-1}^{(n)}] = 0$), with conditional variances $\sigma_{n,i}^2 = \mathbb{E}[\zeta_{n,i}^2 \mid \mathcal{F}_{i-1}^{(n)}]$ and $V_n = \sum_{i\le k_n} \sigma_{n,i}^2$. If, for a constant $\tau^2 > 0$,*

$$(a)\ \ \mathbb{E}|V_n - \tau^2| \to 0, \qquad (b)\ \ \sum_{i\le k_n} \mathbb{E}|\zeta_{n,i}|^3 \to 0, \qquad (c)\ \ \mathbb{E}\max_{i\le k_n} \sigma_{n,i}^2 \to 0,$$

*then $\sum_{i\le k_n} \zeta_{n,i} \Rightarrow \mathcal{N}(0, \tau^2)$.*

*Proof.* Corollary 4.6 of Lectures 12–15 is the case $\tau = 1$; apply it to $\zeta_{n,i}/\tau$, whose conditional variances are $\sigma_{n,i}^2/\tau^2$ and whose third moments are $\mathbb{E}|\zeta_{n,i}|^3/\tau^3$. □

**Corollary 2.2** (vector form)**.** *Let $\zeta_{n,i} \in \mathbb{R}^d$ be martingale differences with conditional covariances $C_{n,i} = \mathbb{E}[\zeta_{n,i}\zeta_{n,i}^\top \mid \mathcal{F}_{i-1}^{(n)}]$, and let $S$ be a positive definite $d \times d$ matrix. If $\mathbb{E}\|\sum_i C_{n,i} - S\|_{\mathrm{op}} \to 0$, $\sum_i \mathbb{E}\|\zeta_{n,i}\|^3 \to 0$ and $\mathbb{E}\max_i \|C_{n,i}\|_{\mathrm{op}} \to 0$, then $\sum_i \zeta_{n,i} \Rightarrow \mathcal{N}(0, S)$.*

*Proof.* Fix $v \in \mathbb{R}^d$, $v \neq 0$. The scalars $\langle v, \zeta_{n,i}\rangle$ are martingale differences with conditional variances $v^\top C_{n,i} v$, so (a)–(c) hold with $\tau^2 = v^\top S v > 0$: $|\sum_i v^\top C_{n,i} v - v^\top S v| \le \|v\|^2 \|\sum_i C_{n,i} - S\|_{\mathrm{op}}$, $|\langle v, \zeta_{n,i}\rangle|^3 \le \|v\|^3 \|\zeta_{n,i}\|^3$, and $v^\top C_{n,i} v \le \|v\|^2 \|C_{n,i}\|_{\mathrm{op}}$. Theorem 2.1 gives $\langle v, \sum_i \zeta_{n,i}\rangle \Rightarrow \mathcal{N}(0, v^\top S v)$, and Cramér–Wold finishes. □

In both proofs below the array is indexed by $i = k + 1$, $k < n$, with $\mathcal{F}_i^{(n)} = \mathcal{F}_i = (z_1, \dots, z_i)$ the filtration of the samples, and the three hypotheses are checked in the order (a), (b), (c).

**Lemma 2.3.** *If $\mathbb{E}\|e_k\|^2 \to 0$ then $\mathbb{E}\|C(\theta_k) - \Sigma\|_{\mathrm{op}} \to 0$.*

*Proof.* Given $\epsilon > 0$ pick $\delta$ with $\|C(\theta) - \Sigma\|_{\mathrm{op}} \le \epsilon$ for $\|e\| \le \delta$. On $\{\|e_k\| > \delta\}$ use $\|C(\theta_k)\|_{\mathrm{op}} \le \sigma^2 + \kappa\|e_k\|^2$: $\mathbb{E}\|C(\theta_k) - \Sigma\|_{\mathrm{op}} \le \epsilon + (\sigma^2 + \|\Sigma\|_{\mathrm{op}})\mathbb{P}(\|e_k\| > \delta) + \kappa\mathbb{E}\|e_k\|^2 \le \epsilon + (\sigma^2 + \|\Sigma\|_{\mathrm{op}})\mathbb{E}\|e_k\|^2/\delta^2 + \kappa\mathbb{E}\|e_k\|^2$. □

## 2.1 Asymptotic normality of the iterates

Write $h_1, \dots, h_d \in [\mu, \beta]$ for the eigenvalues of $H$ and $u_1, \dots, u_d$ for an orthonormal eigenbasis; $\lambda_{\min} := \min_i h_i \ge \mu$.

**Theorem 2.4** (asymptotic normality)**.** *Let $\eta_k = \eta/(k + k_0)$ with $\eta\mu > 1$ and $k_0 \ge \max\big(\eta(\beta^2 + \kappa)/\mu,\ 2\eta\beta\big)$, and assume (A1)–(A5). Then*

$$\sqrt{n}\,(\theta_n - \theta^*) \ \Rightarrow\ \mathcal{N}(0, \Sigma_\eta), \qquad \Sigma_\eta := \sum_{i,j=1}^{d} \frac{\eta^2\, u_i^\top \Sigma u_j}{\eta(h_i + h_j) - 1}\, u_i u_j^\top,$$

*i.e. $\Sigma_\eta$ is the unique solution of the Lyapunov equation $(\eta H - \frac{1}{2}I)\Sigma_\eta + \Sigma_\eta(\eta H - \frac{1}{2}I) = \eta^2\Sigma$.*

In the eigenbasis of $H$, $(\Sigma_\eta)_{ij} = \eta^2\Sigma_{ij}/(\eta(h_i + h_j) - 1)$; for $d = 1$ this is $\eta^2\sigma^2/(2\eta h - 1)$, the variance found for the quadratic in Exercise 4.18 of Lectures 12–15, with the noise covariance *at the minimizer* in place of $\sigma^2$. When $H = hI$ the covariance $\frac{\eta^2}{2\eta h-1}\Sigma$ is minimized at $\eta = 1/h$, where it equals $\Sigma/h^2 = H^{-1}\Sigma H^{-1}$. In general $H^{-1}\Sigma H^{-1}$ is a lower bound that no step size attains unless $\eta H = I$:

*Remark* 2.5 (the covariance floor). For every $\eta$ with $\eta\lambda_{\min} > \frac{1}{2}$, $\Sigma_\eta \succeq H^{-1}\Sigma H^{-1}$ in the positive semidefinite order (so $v^\top\Sigma_\eta v \geq v^\top H^{-1}\Sigma H^{-1}v$ for every direction $v$, and likewise for the trace, the operator norm and the determinant), with equality if and only if $\eta H = I$. Indeed, in the eigenbasis of $H$,

$$(\Sigma_\eta)_{ij} - (H^{-1}\Sigma H^{-1})_{ij} = \frac{\Sigma_{ij}}{h_i h_j} \cdot (\eta h_i - 1)(\eta h_j - 1) \cdot \frac{1}{\eta(h_i + h_j) - 1},$$

and $\frac{1}{\eta(h_i+h_j)-1} = \int_0^\infty e^{-(\eta h_i - \frac{1}{2})t} e^{-(\eta h_j - \frac{1}{2})t}\, dt$ is a Gram matrix, hence positive semidefinite; multiplying its entries by $a_i a_j$, $a_i = \eta h_i - 1$, keeps it so, and the entrywise (Schur) product of two positive semidefinite matrices is positive semidefinite. For $d = 1$ the difference is $\sigma^2(\eta h - 1)^2/(h^2(2\eta h - 1))$.

Choosing $\eta$ well therefore requires knowing $H$, and when the eigenvalues of $H$ are spread no scalar $\eta$ reaches the floor in every direction (Exercise 2.11). The floor itself has a meaning beyond SGD: $H^{-1}\Sigma H^{-1}$ is the asymptotic covariance of the estimator $\hat\theta_n$ that minimizes $\frac{1}{n}\sum_{i\leq n}\ell(\theta; z_i)$ with all $n$ samples in hand (expand $\nabla$ of the empirical loss around $\theta^*$: $\sqrt{n}(\hat\theta_n - \theta^*) \approx -H^{-1}\frac{1}{\sqrt{n}}\sum_i \xi_i$), and when $\ell$ is a negative log-likelihood of a correctly specified model, $H = \Sigma$ is the Fisher information and $H^{-1}\Sigma H^{-1} = H^{-1}$ is the Cramér–Rao bound.

**Plan of the proof**

The recursion (5) is linear in $e_k$ up to the small term $r_k$, so it can be solved: the error at time $n$ is a weighted sum of the initial error, the noises $\xi_1, \ldots, \xi_n$, and the remainders $r_0, \ldots, r_{n-1}$, the weights being matrices. Multiplied by $\sqrt{n}$, the initial-error term and the remainder term will go to zero, and the noise term is a martingale with small, dependent increments, to which the martingale CLT applies through its one-dimensional projections. The weights are polynomials in the single symmetric matrix $H$, so they commute and are diagonal in the eigenbasis of $H$; every estimate reduces to the scalar case, eigenvalue by eigenvalue.

*Step 0: solving the recursion.* Write (5) as $e_{k+1} = (I - \eta_k H)e_k - \eta_k(\xi_{k+1} + r_k)$ and iterate: each later step multiplies what is already there by its factor $I - \eta_j H$ and adds its own contribution. With

$$P_{n,k} := \prod_{j=k}^{n-1}(I - \eta_j H) \quad (P_{n,n} := I), \qquad w_{n,k} := \eta_k\, P_{n,k+1},$$

one checks by induction on $n$ that

$$e_n = P_{n,0}\, e_0 - \sum_{k<n} w_{n,k}\, \xi_{k+1} - \sum_{k<n} w_{n,k}\, r_k. \tag{6}$$

Here $P_{n,k}$ is the total contraction applied between times $k$ and $n$, and $w_{n,k}$ is the weight with which the noise injected at step $k$ is still felt at time $n$. All these matrices are functions of $H$: in the

eigenbasis,

$$P_{n,k}u_i = P_{n,k}^{(i)}u_i, \quad P_{n,k}^{(i)} := \prod_{j=k}^{n-1}(1-\eta_j h_i), \qquad w_{n,k}u_i = w_{n,k}^{(i)}u_i, \quad w_{n,k}^{(i)} := \eta_k P_{n,k+1}^{(i)}.$$

*Step 1: the weights.* All the hypotheses enter through the following scalar estimates. We write

$$m := k + k_0, \qquad N := n + k_0,$$

so that $\eta_k = \eta/m$ and sums over $k < n$ become sums over $m$ from $k_0$ to $N-1$.

**Lemma 2.6** (the weights)**.** *Under the hypotheses of Theorem 2.4, for every eigenvalue $h = h_i$ (so $\eta h \geq \eta\mu > 1$):*

*(i)* $0 \leq P_{n,k}^{(i)} \leq (m/N)^{\eta h}$ *and* $w_{n,k}^{(i)} \leq \dfrac{\eta}{m}\Big(\dfrac{m}{N}\Big)^{\eta h}$*; moreover* $w_{n,k}^{(i)} = \dfrac{\eta}{m}\Big(\dfrac{m}{N}\Big)^{\eta h}\big(1+O(1/m)\big)$.

*(ii) For every pair $i, j$:* $n\displaystyle\sum_{k<n} w_{n,k}^{(i)} w_{n,k}^{(j)} \;\to\; \frac{\eta^2}{\eta(h_i+h_j)-1}$.

*(iii) For each fixed $K$,* $n\displaystyle\sum_{k<K}\big(w_{n,k}^{(i)}\big)^2 \to 0$*; and* $n\max_{k<n}\big(w_{n,k}^{(i)}\big)^2 = O(1/n)$.

*(iv)* $\sqrt{n}\displaystyle\sum_{k<n}\frac{w_{n,k}^{(i)}}{m} = O(n^{-1/2})$ *and* $n^{3/2}\displaystyle\sum_{k<n}\big(w_{n,k}^{(i)}\big)^3 = O(n^{-1/2})$.

*Consequently the operator norms satisfy $\|P_{n,k}\|_{\mathrm{op}} \leq (m/N)^{\eta\lambda_{\min}}$ and $\|w_{n,k}\|_{\mathrm{op}} \leq \frac{\eta}{m}(m/N)^{\eta\lambda_{\min}}$, and (iii)–(iv) hold with $w_{n,k}^{(i)}$ replaced by $\|w_{n,k}\|_{\mathrm{op}}$.*

*Proof.* (i) Since $k_0 \geq 2\eta\beta$, every $\eta_j h \leq \frac{1}{2}$, so $0 \leq 1-\eta_j h \leq e^{-\eta_j h}$ and

$$P_{n,k}^{(i)} \leq \exp\Big(-\eta h\sum_{j=k}^{n-1}\frac{1}{j+k_0}\Big) \leq \exp\Big(-\eta h\log\frac{N}{m}\Big) = \Big(\frac{m}{N}\Big)^{\eta h},$$

by comparing the sum with $\int_m^N dx/x$. In the other direction $\log(1-x) \geq -x-x^2$ on $[0,\frac{1}{2}]$, and $\sum_{j=k}^{n-1}\frac{1}{j+k_0} \leq \log\frac{N}{m} + \frac{1}{m}$, $\sum_{j\geq k}\eta_j^2 h^2 \leq \frac{\eta^2 h^2}{m-1}$, so $P_{n,k}^{(i)} \geq (m/N)^{\eta h}e^{-O(1/m)}$; multiply by $\eta_k = \eta/m$.
(ii) By (i), with $p := \eta(h_i+h_j) > 2$,

$$n\sum_{k<n} w_{n,k}^{(i)} w_{n,k}^{(j)} = (1+o(1))\,\eta^2\,N^{1-p}\sum_{m=k_0}^{N-1} m^{p-2} \qquad (N\to\infty),$$

the finitely many small $m$ whose $O(1/m)$ factor is not negligible contributing $O(N^{1-p}) \to 0$. Since $p-2 > -1$, $\sum_{m<N} m^{p-2} \sim N^{p-1}/(p-1)$, and the claim follows.
(iii) The terms $k < K$ are $O(N^{1-2\eta h}) \to 0$ since $2\eta h > 1$. The bound $\frac{\eta}{m}(m/N)^{\eta h} = \eta m^{\eta h-1}N^{-\eta h}$ is increasing in $m$ because $\eta h > 1$, so $\max_k w_{n,k}^{(i)} \leq \eta/N$ and $n\max_k(w_{n,k}^{(i)})^2 \leq \eta^2 n/N^2$.
(iv) $\sqrt{n}\sum_k \frac{w_{n,k}^{(i)}}{m} \leq \sqrt{n}\,\eta N^{-\eta h}\sum_{m<N} m^{\eta h-2} = O(\sqrt{n}\,N^{-\eta h}N^{\eta h-1}) = O(n^{-1/2})$, using $\eta h - 2 > -1$. Likewise $n^{3/2}\sum_k (w_{n,k}^{(i)})^3 \leq n^{3/2}\eta^3 N^{-3\eta h}\sum_{m<N} m^{3\eta h-3} = O(n^{3/2}N^{-3\eta h}N^{3\eta h-2}) = O(n^{-1/2})$.
The operator-norm statements follow because $P_{n,k}$ and $w_{n,k}$ are diagonal in the eigenbasis with the nonnegative entries $P_{n,k}^{(i)}$, $w_{n,k}^{(i)}$, and $(m/N)^{\eta h}$ is largest for the smallest eigenvalue. □

*Proof of Theorem 2.4.* Multiply (6) by $\sqrt{n}$:

$$\sqrt{n}\,e_n = \underbrace{\sqrt{n}\,P_{n,0}e_0}_{\text{(I)}} \;-\; \underbrace{\sqrt{n}\sum_{k<n} w_{n,k}\xi_{k+1}}_{\text{(II)}} \;-\; \underbrace{\sqrt{n}\sum_{k<n} w_{n,k}r_k}_{\text{(III)}}.$$

We show that $\|\text{(I)}\|$ and $\|\text{(III)}\|$ tend to 0 in probability and that $\text{(II)} \Rightarrow \mathcal{N}(0,\Sigma_\eta)$. Slutsky's lemma then gives the theorem.

*(I), the initial condition.* By Lemma 2.6, $\|\text{(I)}\| \le \sqrt{n}\,(k_0/N)^{\eta\lambda_{\min}}\|e_0\|$, which tends to 0 because $\eta\lambda_{\min} > \frac{1}{2}$. The initial condition is forgotten faster than $n^{-1/2}$.

*(III), the curvature remainder.* By (A4), $\|r_k\| \le \frac{\gamma}{2}\|e_k\|^2$, and by Theorem 1.5, $\mathbb{E}\|e_k\|^2 \le D/m$ for a constant $D$. Hence

$$\mathbb{E}\|\text{(III)}\| \le \frac{\gamma}{2}\sqrt{n}\sum_{k<n}\|w_{n,k}\|_{\mathrm{op}}\,\mathbb{E}\|e_k\|^2 \le \frac{\gamma D}{2}\,\sqrt{n}\sum_{k<n}\frac{\|w_{n,k}\|_{\mathrm{op}}}{m} = O(n^{-1/2})$$

by Lemma 2.6(iv). So (III) tends to 0 in $L^1$, hence in probability. This is where the $1/n$ rate of Lecture 16 is used: the non-linearity contributes at order $\|e_k\|^2 \asymp 1/k$, which after weighting is $O(1/n)$, below the $n^{-1/2}$ scale.

*(II), the noise.* By Cramér–Wold it suffices to fix $v \in \mathbb{R}^d$ and show $\langle v, \text{(II)}\rangle \Rightarrow \mathcal{N}(0, v^\top\Sigma_\eta v)$. Fix $n$ and put

$$\zeta_{n,k+1} := \langle v, \sqrt{n}\,w_{n,k}\xi_{k+1}\rangle = \sqrt{n}\,\langle w_{n,k}v, \xi_{k+1}\rangle, \qquad k < n,$$

using that $w_{n,k}$ is symmetric. These are scalar martingale differences with respect to $(\mathcal{F}_k)$: the vector $\sqrt{n}w_{n,k}v$ is deterministic, and $\mathbb{E}_k\xi_{k+1} = 0$. Their conditional variances and third moments are

$$\mathbb{E}_k\zeta_{n,k+1}^2 = n\,(w_{n,k}v)^\top C(\theta_k)\,(w_{n,k}v), \qquad \mathbb{E}_k|\zeta_{n,k+1}|^3 \le n^{3/2}\|w_{n,k}\|_{\mathrm{op}}^3\|v\|^3\rho^3.$$

They are not independent: the law of $\xi_{k+1}$ and its conditional covariance $C(\theta_k)$ depend on $\theta_k$, hence on $z_1,\dots,z_k$ (in Example 1.6, $\xi = xx^\top e - e - x\varepsilon$). So this is the martingale CLT and not Lindeberg's theorem; the two coincide only for additive noise $g(\theta;z) = \nabla L(\theta) + z$ with $z$ independent of $\theta$, as in Exercise 4.18 of Lectures 12–15.

We apply Theorem 2.1 with $\tau^2 = \tau_v^2 := v^\top\Sigma_\eta v > 0$, which requires three things.

*(a) The quadratic variation concentrates:* $\mathbb{E}|V_n - \tau_v^2| \to 0$, where

$$V_n = n\sum_{k<n}(w_{n,k}v)^\top C(\theta_k)\,(w_{n,k}v).$$

The idea is that the weights put almost all their mass on large $k$, where $\theta_k$ is close to $\theta^*$ and so $C(\theta_k)$ is close to $\Sigma$. Precisely, fix $K$ and split the sum at $K$. The part $k < K$ has expectation at most $\|v\|^2\big(\sigma^2 + \kappa\max_{k<K}\mathbb{E}\|e_k\|^2\big)\cdot n\sum_{k<K}\|w_{n,k}\|_{\mathrm{op}}^2$, which tends to 0 by Lemma 2.6(iii). For the part $k \ge K$,

$$\mathbb{E}\Big|n\sum_{k\ge K}(w_{n,k}v)^\top\big(C(\theta_k)-\Sigma\big)(w_{n,k}v)\Big| \;\le\; \Big(\sup_{k\ge K}\mathbb{E}\|C(\theta_k)-\Sigma\|_{\mathrm{op}}\Big)\cdot\|v\|^2\,n\sum_{k<n}\|w_{n,k}\|_{\mathrm{op}}^2,$$

and the supremum is small for $K$ large by Lemma 2.3, since $\mathbb{E}\|e_k\|^2 \to 0$. What remains is the deterministic sum $n\sum_{k<n}(w_{n,k}v)^\top\Sigma(w_{n,k}v)$; expanding $v = \sum_i v_i u_i$ in the eigenbasis, $w_{n,k}v = \sum_i v_i w_{n,k}^{(i)}u_i$, so this sum is

$$\sum_{i,j} v_iv_j\,(u_i^\top\Sigma u_j)\cdot n\sum_{k<n} w_{n,k}^{(i)}w_{n,k}^{(j)} \;\longrightarrow\; \sum_{i,j} v_iv_j\,\frac{\eta^2\,u_i^\top\Sigma u_j}{\eta(h_i+h_j)-1} = v^\top\Sigma_\eta v = \tau_v^2$$

by Lemma 2.6(ii). Letting $n \to \infty$ and then $K \to \infty$ gives $\mathbb{E}|V_n - \tau_v^2| \to 0$.

*(b) The third moments vanish:* $\sum_{k<n} \mathbb{E}|\zeta_{n,k+1}|^3 \le \rho^3 \|v\|^3 n^{3/2} \sum_{k<n} \|w_{n,k}\|_{\mathrm{op}}^3 = O(n^{-1/2})$ by Lemma 2.6(iv).

*(c) No single step dominates:* since $\|C(\theta_k)\|_{\mathrm{op}} \le \sigma^2 + \kappa\|e_k\|^2$,

$$\mathbb{E} \max_{k<n} \mathbb{E}_k \zeta_{n,k+1}^2 \;\le\; n \max_{k<n} \|w_{n,k}\|_{\mathrm{op}}^2 \, \|v\|^2 \Big(\sigma^2 + \kappa\, \mathbb{E} \max_{k<n} \|e_k\|^2\Big),$$

and $n \max_k \|w_{n,k}\|_{\mathrm{op}}^2 = O(1/n)$ by Lemma 2.6(iii), while

$$\mathbb{E} \max_{k<n} \|e_k\|^2 \le \sum_{k<n} \mathbb{E}\|e_k\|^2 \le \sum_{k<n} \frac{D}{m} = O(\log n);$$

the product is $O(\log n/n) \to 0$.

Hence $\langle v, (\mathrm{II})\rangle \Rightarrow \mathcal{N}(0, \tau_v^2)$ for every $v$, i.e. (II) $\Rightarrow \mathcal{N}(0, \Sigma_\eta)$, which with (I) and (III) completes the proof. (That $\Sigma_\eta$ solves the Lyapunov equation is checked entry by entry in the eigenbasis: $(\eta h_i - \frac{1}{2} + \eta h_j - \frac{1}{2})(\Sigma_\eta)_{ij} = \eta^2 \Sigma_{ij}$.) □

## 2.2 Polyak–Ruppert averaging

By Remark 2.5, $H^{-1}\Sigma H^{-1}$ is the smallest covariance Theorem 2.4 can produce, and reaching it needs $\eta H = I$, which a scalar step cannot achieve unless $H$ is a multiple of the identity. Averaging the iterates obtains it with no knowledge of $H$, for any step-size constant, and with a larger step size, which also forgets the initial condition faster.

**Theorem 2.7** (Polyak–Ruppert)**.** *Let $\eta_k = \eta(k + k_0)^{-\alpha}$ with $\alpha \in (\frac{1}{2}, 1)$ and $\eta_k \le \mu/(\beta^2 + \kappa)$, let (A1)–(A5) hold, and let $\bar\theta_n = \frac{1}{n}\sum_{k<n} \theta_k$. Then*

$$\sqrt{n}\,(\bar\theta_n - \theta^*) \;\Rightarrow\; \mathcal{N}\big(0,\; H^{-1}\Sigma H^{-1}\big).$$

### Plan of the proof

Instead of solving the recursion, read it backwards: (5) expresses $He_k$, the quantity whose average we want, as a difference of consecutive errors plus noise plus a remainder,

$$H\, e_k = \frac{e_k - e_{k+1}}{\eta_k} - \xi_{k+1} - r_k.$$

Averaging over $k < n$ turns the first term into a telescoping sum, which is small because the errors change slowly when the step is large; the second into $\frac{1}{n}\sum \xi_{k+1}$, a plain martingale average whose fluctuations are of order $n^{-1/2}$; and the third into an average of $\|e_k\|^2$, which is small because the errors are small. The only input from Lecture 16 is a rate: for $\eta_k = \eta(k+k_0)^{-\alpha}$, Exercise 1.11 gives

$$\mathbb{E}\|e_k\|^2 \le D\,\eta_k = O(k^{-\alpha}), \qquad \text{hence} \qquad \mathbb{E}\|e_k\| \le \sqrt{D\eta_k} = O(k^{-\alpha/2}), \tag{7}$$

for a constant $D$, once $k_0$ is large enough. The proof is three estimates and the martingale CLT; $H$ plays no role until the last line.

*Proof of Theorem 2.7.* Average the identity above over $k < n$ and multiply by $\sqrt{n}$:

$$\sqrt{n}\, H\, \bar{e}_n = \underbrace{\frac{1}{\sqrt{n}} \sum_{k<n} \frac{e_k - e_{k+1}}{\eta_k}}_{\text{(I)}} \;-\; \underbrace{\frac{1}{\sqrt{n}} \sum_{k<n} \xi_{k+1}}_{\text{(II)}} \;-\; \underbrace{\frac{1}{\sqrt{n}} \sum_{k<n} r_k}_{\text{(III)}}, \qquad \bar{e}_n := \bar{\theta}_n - \theta^*.$$

We show that $\|\text{(I)}\|$ and $\|\text{(III)}\|$ tend to 0 in probability and that (II) $\Rightarrow \mathcal{N}(0, \Sigma)$; Slutsky's lemma then gives $\sqrt{n}\, H\bar{e}_n \Rightarrow \mathcal{N}(0, \Sigma)$, and applying the fixed matrix $H^{-1}$ gives the claim, since $H^{-1}\mathcal{N}(0,\Sigma) = \mathcal{N}(0, H^{-1}\Sigma H^{-1})$.

*(I), the telescoping term.* Summation by parts moves the differences from the errors to the step sizes:

$$\sum_{k<n} \frac{e_k - e_{k+1}}{\eta_k} = \frac{e_0}{\eta_0} - \frac{e_n}{\eta_{n-1}} + \sum_{k=1}^{n-1} e_k \Big( \frac{1}{\eta_k} - \frac{1}{\eta_{k-1}} \Big).$$

(Check: the coefficient of $e_k$ on the left is $\frac{1}{\eta_k} - \frac{1}{\eta_{k-1}}$ for $1 \le k \le n-1$, $\frac{1}{\eta_0}$ for $k = 0$, and $-\frac{1}{\eta_{n-1}}$ for $k = n$.) Now $\frac{1}{\eta_k} = \frac{(k+k_0)^\alpha}{\eta}$ is increasing, and by the mean value theorem applied to $x \mapsto x^\alpha$ on $[k-1+k_0, k+k_0]$, whose derivative $\alpha x^{\alpha-1}$ is decreasing,

$$0 \;\le\; \frac{1}{\eta_k} - \frac{1}{\eta_{k-1}} \;\le\; \frac{\alpha}{\eta}\,(k-1+k_0)^{\alpha-1} \;\le\; \frac{2\alpha}{\eta}\,(k+k_0)^{\alpha-1}\,:$$

consecutive step sizes differ little, by a relative amount of order $1/k$. Take norms and expectations, using (7):

$$\mathbb{E}\|\text{(I)}\| \;\le\; \frac{\|e_0\|}{\sqrt{n}\,\eta_0} + \frac{\mathbb{E}\|e_n\|}{\sqrt{n}\,\eta_{n-1}} + \frac{1}{\sqrt{n}} \sum_{k=1}^{n-1} \frac{2\alpha}{\eta} (k+k_0)^{\alpha-1}\, \mathbb{E}\|e_k\|.$$

The first term is $O(n^{-1/2})$. The second is $O\big(n^{-1/2} \cdot n^{-\alpha/2} \cdot n^{\alpha}\big) = O(n^{(\alpha-1)/2})$, which tends to 0 because $\alpha < 1$. The third is $O\big(n^{-1/2} \sum_{k<n} k^{\alpha-1} k^{-\alpha/2}\big) = O\big(n^{-1/2}\, n^{\alpha/2}\big) = O(n^{(\alpha-1)/2})$, since $\alpha/2 - 1 > -1$; again $\to 0$ because $\alpha < 1$. So (I) tends to 0 in $L^1$, hence in probability. This is the step that needs $\alpha < 1$: a step size decaying like $1/k$ would make $1/\eta_k$ grow linearly and the second term would not vanish.

*(III), the curvature remainder.* By (A4) and (7),

$$\mathbb{E}\|\text{(III)}\| \le \frac{\gamma}{2\sqrt{n}} \sum_{k<n} \mathbb{E}\|e_k\|^2 \le \frac{\gamma D}{2\sqrt{n}} \sum_{k<n} \eta_k = O\big(n^{-1/2}\, n^{1-\alpha}\big) = O(n^{1/2-\alpha}),$$

which tends to 0 because $\alpha > \frac{1}{2}$. This is the step that needs $\alpha > \frac{1}{2}$: the errors must decay fast enough that their squared norms, summed, are $o(\sqrt{n})$.

*(II), the noise.* By Cramér–Wold, fix $v \in \mathbb{R}^d$. The summands $\langle v, \xi_{k+1}\rangle/\sqrt{n}$, $k < n$, are scalar martingale differences with respect to $(\mathcal{F}_k)$, with

$$\mathbb{E}_k\Big[\frac{\langle v, \xi_{k+1}\rangle^2}{n}\Big] = \frac{v^\top C(\theta_k) v}{n}, \qquad \mathbb{E}_k\Big|\frac{\langle v, \xi_{k+1}\rangle}{\sqrt{n}}\Big|^3 \le \frac{\|v\|^3 \rho^3}{n^{3/2}}.$$

As in Theorem 2.4 they are dependent, with a random conditional variance, so the martingale CLT is the tool. The weights are now all equal, which makes the three hypotheses of Theorem 2.1 (with $\tau^2 = \tau_v^2 := v^\top \Sigma v > 0$) easier than before.

*(a) The quadratic variation concentrates:* $V_n = \frac{1}{n}\sum_{k<n} v^\top C(\theta_k)v$, and

$$\mathbb{E}|V_n - \tau_v^2| \le \|v\|^2 \cdot \frac{1}{n}\sum_{k<n}\mathbb{E}\|C(\theta_k)-\Sigma\|_{\mathrm{op}} \;\longrightarrow\; 0,$$

because $\mathbb{E}\|C(\theta_k)-\Sigma\|_{\mathrm{op}} \to 0$ by Lemma 2.3 (as $\mathbb{E}\|e_k\|^2 \to 0$) and Cesàro averages of a null sequence are null.

*(b) The third moments vanish:* $\sum_{k<n}\mathbb{E}|\langle v, \xi_{k+1}\rangle/\sqrt{n}|^3 \le n \cdot \|v\|^3\rho^3 n^{-3/2} \to 0$.

*(c) No single step dominates:* $\mathbb{E}\max_{k<n}\frac{v^\top C(\theta_k)v}{n} \le \|v\|^2\,\frac{\sigma^2+\kappa\,\mathbb{E}\max_{k<n}\|e_k\|^2}{n}$, and $\mathbb{E}\max_{k<n}\|e_k\|^2 \le \sum_{k<n}\mathbb{E}\|e_k\|^2 \le D\sum_{k<n}\eta_k = O(n^{1-\alpha})$, so the bound is $O(n^{-\alpha}) \to 0$.

Hence $\langle v, (\mathrm{II})\rangle \Rightarrow \mathcal{N}(0, v^\top\Sigma v)$ for every $v$, i.e. $(\mathrm{II}) \Rightarrow \mathcal{N}(0,\Sigma)$, which with (I) and (III) completes the proof. □

Both constraints on $\alpha$ are visible in the proof: $\alpha < 1$ makes the iterates move slowly enough that the telescoping term is negligible, $\alpha > \frac{1}{2}$ makes $\mathbb{E}\|e_k\|^2$ decay fast enough that the curvature error is negligible. The averaged iterate has, to first order, the fluctuations of $H^{-1}\frac{1}{n}\sum\xi_{k+1}$: it behaves like the full-sample empirical minimizer of Remark 2.5, whose covariance is $H^{-1}\Sigma H^{-1}$. For least mean squares (Example 1.6) $H = I$ and $\Sigma = \mathbb{E}[xx^\top\varepsilon^2] = \sigma_\varepsilon^2 I$, so $\mathbb{E}\|\bar\theta_n - \theta^*\|^2 \approx d\sigma_\varepsilon^2/n$, the least squares value.

## 2.3 High-probability bounds along the trajectory

Theorems 1.5 and 2.4 describe a fixed time $n$. Freedman's inequality controls the whole trajectory at once, through the noise martingale $M_n = \sum_{k<n}\eta_k\xi_{k+1}$ and its realized quadratic variation. We state it for a quadratic loss, one eigen-coordinate of $H$ at a time.

**Proposition 2.8.** *Let $L(\theta) = \frac{1}{2}e^\top He$ (so $r \equiv 0$), let the steps be nonincreasing with $\eta_k\beta \le 1$, and let $\|\xi_k\| \le c$. For an eigenvector $u$ of $H$ with eigenvalue $h$ write $e_n^{(u)} = \langle u, e_n\rangle$, $M_n^{(u)} = \langle u, M_n\rangle$, and $P_{n,0}^{(u)} = \prod_{j<n}(1-\eta_j h)$. Then for every $n$,*

$$\big|e_n^{(u)} - P_{n,0}^{(u)}e_0^{(u)}\big| \;\le\; 2\max_{k\le n}|M_k^{(u)}|,$$

*and consequently, for all $N$ and $u_0, v > 0$,*

$$\mathbb{P}\Big(\exists\, n \le N:\ |e_n^{(u)} - P_{n,0}^{(u)}e_0^{(u)}| \ge 2u_0 \quad \text{and} \quad \sum_{k<n}\eta_k^2\, u^\top C(\theta_k)u \le v\Big) \;\le\; 2\exp\Big(-\frac{u_0^2}{2(v+\eta_0 c u_0/3)}\Big).$$

*Proof.* Take the inner product of (6) (with $r \equiv 0$) with $u$: $e_n^{(u)} - P_{n,0}^{(u)}e_0^{(u)} = -\sum_{k<n}P_{n,k+1}^{(u)}(M_{k+1}^{(u)} - M_k^{(u)})$, where $P_{n,k}^{(u)} = \prod_{j=k}^{n-1}(1-\eta_j h)$. Summation by parts, with $P_{n,n}^{(u)} = 1$, $M_0^{(u)} = 0$ and $P_{n,k+1}^{(u)} - P_{n,k}^{(u)} = \eta_k h P_{n,k+1}^{(u)} \ge 0$:

$$\sum_{k<n}P_{n,k+1}^{(u)}(M_{k+1}^{(u)} - M_k^{(u)}) = M_n^{(u)} - \sum_{k=1}^{n-1}\big(P_{n,k+1}^{(u)} - P_{n,k}^{(u)}\big)M_k^{(u)}, \qquad \sum_{k=1}^{n-1}\big(P_{n,k+1}^{(u)} - P_{n,k}^{(u)}\big) = 1 - P_{n,1}^{(u)} \le 1.$$

So the sum is bounded by $|M_n^{(u)}| + \max_{k<n}|M_k^{(u)}|$. For the probability bound, $M^{(u)}$ is a scalar martingale with steps at most $\eta_0 c$ and $\langle M^{(u)}\rangle_n = \sum_{k<n}\eta_k^2\, u^\top C(\theta_k)u$, nondecreasing in $n$, so on

the event some $k \le n$ has $|M_k^{(u)}| \ge u_0$ and $\langle M^{(u)} \rangle_k \le v$; apply Theorem 5.10 of Lectures 12–15 to $M^{(u)}$ and to $-M^{(u)}$. □

The bound is uniform over the horizon and sees the variance actually accumulated; a union bound over the $d$ eigen-coordinates controls $\|e_n - P_{n,0} e_0\|$. Two consequences.

**Example 2.9** (tail of the trajectory)**.** With $\eta_k = \eta/(k + k_0)$ and $u^\top C(\theta) u \le \sigma^2$, apply the proposition to the shifted martingale $M_n^{(u)} - M_m^{(u)}$: for $n \ge m$ the noise contribution after time $m$ is at most $2 \max_{m \le k \le n} |M_k^{(u)} - M_m^{(u)}|$, and $\sum_{k \ge m} \eta_k^2 \sigma^2 \le \eta^2 \sigma^2/(m + k_0 - 1)$. So with probability at least $1 - \delta$, simultaneously for all $n \ge m$, the noise contributes at most $2\eta\sigma \sqrt{2 \log(2/\delta)/(m + k_0 - 1)}\,(1 + o(1))$ in the direction $u$: the $1/\sqrt{m}$ rate of Theorem 1.5, now for the whole future at once, with a $\sqrt{\log(1/\delta)}$ price.

**Example 2.10** (interpolation)**.** WORKED IN CLASS If the noise vanishes at the minimizer, $s^2(\theta) \le \kappa \|e\|^2$ (so $\sigma = 0$: every sample gradient is zero at $\theta^*$, as for a model that fits all the data), then Lemma 1.1 has no additive term: $a_{k+1} \le (1 - \mu\eta_k) a_k$ for $\eta_k \le \mu/(\beta^2 + \kappa)$. A constant step gives $\mathbb{E}\|e_n\|^2 \le (1 - \mu\eta)^n \|e_0\|^2$ with no noise floor, and $\eta_k = \eta/(k + k_0)$ gives $O(n^{-\eta\mu})$, faster than $1/n$ when $\eta\mu > 1$. In Proposition 2.8 the realized variance $\sum_{k<n} \eta_k^2 u^\top C(\theta_k) u \le \kappa \sum_{k<n} \eta_k^2 \|e_k\|^2$ is then itself small along the trajectory, and Freedman's bound improves with it, while Azuma's, which only knows $\|\xi_k\| \le c$, does not (Exercise 5.17 of Lectures 12–15).

## Exercises for Lecture 17

**Exercise 2.11.** TYPE A For $H = \operatorname{diag}(h_1, \dots, h_d)$ and $\Sigma = \operatorname{diag}(s_1^2, \dots, s_d^2)$ verify directly that $\Sigma_\eta = \operatorname{diag}\big(\eta^2 s_i^2/(2\eta h_i - 1)\big)$ solves the Lyapunov equation, so $\operatorname{tr} \Sigma_\eta = \sum_i \eta^2 s_i^2/(2\eta h_i - 1)$. Which $\eta$ minimizes it when all $h_i$ are equal? When they are not, compare $\min_\eta \operatorname{tr} \Sigma_\eta$ with $\operatorname{tr}(H^{-1} \Sigma H^{-1})$ for $d = 2$, $h = (1, 10)$, $s \equiv 1$.

**Exercise 2.12.** TYPE B In the linear case $L(\theta) = \frac{h}{2}\theta^2$, $d = 1$, with i.i.d. noise of variance $\sigma^2$ and $\eta_k = \eta/(k + k_0)$, show that $n\,\mathbb{E} e_n^2 \to \infty$ when $\eta h < \frac{1}{2}$ and $n \mathbb{E} e_n^2 \asymp \log n$ when $\eta h = \frac{1}{2}$ (compute $n \sum_k w_{n,k}^2$ as in Lemma 2.6). So the threshold $\eta \lambda_{\min} > \frac{1}{2}$ is sharp, and Theorem 1.5's $\eta\mu > 1$ is an artefact of bounding $(1 - \eta_k h)^2$ by $1 - \eta_k h$.

**Exercise 2.13** (tail averaging)**.** TYPE B Let $\bar\theta_n^{(1/2)} = \frac{2}{n} \sum_{n/2 \le k < n} \theta_k$ (average of the second half). Adapt the proof of Theorem 2.7 to show $\sqrt{n}(\bar\theta_n^{(1/2)} - \theta^*) \Rightarrow \mathcal{N}(0, 2H^{-1} \Sigma H^{-1})$, and explain in one paragraph why one might accept the factor 2 (hint: how do $\bar\theta_n$ and $\bar\theta_n^{(1/2)}$ depend on $\theta_0$ for moderate $n$?).

**Exercise 2.14** (constant step plus averaging)**.** TYPE B For the linear case in $d = 1$ with constant step $\eta$ ($\eta h < 1$) and i.i.d. noise, compute $\lim_n n \operatorname{Var} \bar\theta_n$ exactly, and show it equals $\sigma^2/h^2$: with a constant step the averaged iterate is still efficient, although the last iterate is not (Theorem 1.2). Where did the step size go?

**Exercise 2.15** (averaging, simulated)**.** TYPE C For least mean squares (Example 1.6) with $d = 20$, compare over 200 runs and $n = 10^4$: the last iterate with $\eta_k = 2/(k + k_0)$; the last iterate with $\eta_k = 1/(k + k_0)$ (the optimal $\eta = 1/h$); the averaged iterate with $\eta_k = \eta k^{-0.7}$ for $\eta \in \{0.2, 0.5, 1\}/(d + 2)$. Report $n\mathbb{E}\|\cdot - \theta^*\|^2$ for each and compare with $4d$, $d$ and $d$ respectively. Then repeat with a badly conditioned design $x \sim \mathcal{N}(0, \operatorname{diag}(1, \dots, 1, 0.05))$ and comment on which methods needed to know the conditioning.

**Exercise 2.16** (non-gradient fields)**.** type C Stochastic approximation solves $F(\theta^*) = 0$ for a general vector field $F$, by $\theta_{k+1} = \theta_k - \eta_k(F(\theta_k) + \xi_{k+1})$; the Jacobian $J = DF(\theta^*)$ need not be symmetric. (a) Identify every place in the proof of Theorem 2.4 where symmetry of $H$ was used. (b) Suppose $J$ is diagonalizable with real eigenvalues $\geq \mu > 0$, $J = Q\Lambda Q^{-1}$. Show the proof goes through after the change of variables $e \mapsto Q^{-1}e$, and write the limiting covariance. (c) What changes if $J$ has complex eigenvalues with positive real parts? (Only the weights lemma; state the condition on $\eta \operatorname{Re}\lambda_{\min}$ that replaces $\eta\lambda_{\min} > \frac{1}{2}$.)