Solutions to the exercises, Lectures 16–17

Stochastic gradient descent

Numbering follows the notes *Lectures 16–17: Stochastic gradient descent*; (A1)–(A5), equations (3)–(7) and the theorems are those of the notes. Throughout $e_k = \theta_k - \theta^*$, $a_k = \mathbb{E}\|e_k\|^2$, $m = k + k_0$, $N = n + k_0$. Simulation numbers are averages over 200 independent runs.

Lecture 16

1.8 the quadratic

$\nabla L(\theta) = \frac{1}{2}(H + H^\top)\theta = H\theta$ since H is symmetric. (A1): $\langle H(\theta - \theta'), \theta - \theta'\rangle \geq \mu\|\theta - \theta'\|^2$ because $H \succeq \mu I$. (A2): $\|H(\theta - \theta')\| \leq \beta\|\theta - \theta'\|$ because $H \preceq \beta I$ and $H \succeq 0$ give $\|H\|_{\mathrm{op}} \leq \beta$. The minimizer is $\theta^* = 0$. For $g(\theta; z) = H\theta + z$: $\mathbb{E}g = H\theta = \nabla L(\theta)$ and $\mathbb{E}\|g - \nabla L\|^2 = \mathbb{E}\|z\|^2 = \operatorname{tr}\Sigma$, independent of θ, so (A3) holds with $\kappa = 0$, $\sigma^2 = \operatorname{tr}\Sigma$. This is *additive* noise: $\xi_{k+1} = z_{k+1}$ is independent of the past, so here (and only here) the martingale CLT of Lecture 17 is Lindeberg's theorem.

With $H = hI$ and constant step, $e_{k+1} = (1 - \eta h)e_k - \eta z_{k+1}$. The cross term has mean zero (z_{k+1} is centered and independent of e_k), so

$$a_{k+1} = (1 - \eta h)^2 a_k + \eta^2 \operatorname{tr}\Sigma, \qquad a_n = (1 - \eta h)^{2n} a_0 + \eta^2 \operatorname{tr}\Sigma \, \frac{1 - (1 - \eta h)^{2n}}{1 - (1 - \eta h)^2} \xrightarrow[n\to\infty]{} \frac{\eta \operatorname{tr}\Sigma}{h\,(2 - \eta h)}.$$

Theorem 1.2 ($\mu = \beta = h$) gives $(1 - \eta h)^n a_0 + \eta \operatorname{tr}\Sigma / h$: the exact contraction per step is $(1 - \eta h)^2 \approx 1 - 2\eta h$ rather than $1 - \eta h$, and the floor is smaller by the factor $2 - \eta h \in (1, 2]$ (Remark 1.3).

1.9 the slow rate when $\eta\mu < 1$

Why the naive induction fails. Suppose $a_k \leq Dm^{-c}$ and try to show $a_{k+1} \leq D(m+1)^{-c}$. The recursion gives $a_{k+1} \leq Dm^{-c} - cDm^{-c-1} + Cm^{-2}$, while $D(m+1)^{-c} = Dm^{-c} - cDm^{-c-1} + O(Dm^{-c-2})$. The two leading terms cancel exactly, and what is left compares Cm^{-2} with $O(Dm^{-c-2})$, which is smaller for large m whatever D is. The ansatz is right to leading order but has no room for the noise.

Unrolling. Induction on n (as for (4)) gives $a_n \leq a_0 \prod_{k<n}(1 - \frac{c}{m}) + C \sum_{k<n} m^{-2} \prod_{j=k+1}^{n-1}(1 - \frac{c}{j+k_0})$, and $\prod_{j=k}^{n-1}(1 - \frac{c}{j+k_0}) \leq \exp(-c \sum_{j=k}^{n-1} \frac{1}{j+k_0}) \leq (m/N)^c$ by comparison with $\int_m^N dx/x$. Hence

$$a_n \leq a_0 \Big(\frac{k_0}{N}\Big)^c + CN^{-c} \sum_{m=k_0}^{N-1} \frac{(m+1)^c}{m^2} \leq N^{-c}\Big[a_0 k_0^c + 2^c C \sum_{m \geq k_0} m^{c-2}\Big] \leq N^{-c}\Big[a_0 k_0^c + \frac{4C}{1-c}\Big],$$

using $(m+1)^c \leq 2^c m^c$ and $\sum_{m \geq k_0} m^{c-2} \leq k_0^{c-2} + \int_{k_0}^{\infty} x^{c-2} dx \leq \frac{2k_0^{c-1}}{1-c} \leq \frac{2}{1-c}$. So $D = a_0 k_0^c + 4C/(1-c)$ works. (The sum $\sum m^{c-2}$ converges because $c < 1$: the noise injected at early times is not forgotten, which is why the rate is N^{-c} and not $1/N$. For $c > 1$ the sum is $\asymp N^{c-1}$ and one recovers Lemma 1.4.)

The corrected induction. With $b_k = Dm^{-c} - Em^{-1}$ one needs $(1 - \frac{c}{m})b_k + \frac{C}{m^2} \leq D(m+1)^{-c} - \frac{E}{m+1}$. Since $(1+x)^{-c} \geq 1 - cx$, $D(m+1)^{-c} \geq Dm^{-c} - cDm^{-c-1}$, so it suffices that $-\frac{E}{m} + \frac{cE + C}{m^2} \leq -\frac{E}{m+1}$,

i.e. $(cE+C)(m+1) \le Em$, i.e. $m\big(E(1-c)-C\big) \ge cE+C$. Take $E = 2C/(1-c)$: the condition becomes $m \ge (1+c)/(1-c)$, so $k_0 \ge (1+c)/(1-c)$ suffices, with $D = k_0^c(a_0 + E/k_0)$ for the base case. The extra term $-E/m$ supplies the missing order-m^{-2} decrease.

SGD. For $\eta_k = \eta/(k+k_0)$ with $\eta\mu < 1$, (3) is the recursion with $c = \eta\mu$, $C = \eta^2\sigma^2$, so $\mathbb{E}\|e_n\|^2 \le D(n+k_0)^{-\eta\mu}$.

The quadratic. For $L = \frac{h}{2}\theta^2$ in $d = 1$ with additive noise of variance σ^2 (Exercise 1.8), $a_{k+1} = (1-\eta_k h)^2 a_k + \eta_k^2\sigma^2$ exactly, and (3) replaces $(1-\eta_k h)^2$ by $1-\eta_k h$. Unrolling exactly, with $P_{n,k} = \prod_{j=k}^{n-1}(1-\eta_j h) = (m/N)^{\eta h}(1+O(1/m))$ (Lemma 2.6(i)),

$$a_n = P_{n,0}^2 a_0 + \sigma^2 \sum_{k<n} \eta_k^2 P_{n,k+1}^2 \;\asymp\; N^{-2\eta h}\Big[a_0 k_0^{2\eta h} + \eta^2\sigma^2 \sum_{m<N} m^{2\eta h - 2}\Big] \;\asymp\; N^{-2\eta h} \qquad (2\eta h < 1),$$

both terms being of this order (the sum converges). So the quadratic decays at rate $n^{-2\eta h}$, twice the exponent of the bound $n^{-\eta\mu}$, $\mu = h$; the bound is not attained, for the reason in Exercise 2.12: bounding $(1-\eta_k h)^2$ by $1-\eta_k h$ loses half the contraction. The recursion of the lemma is sharp for *sequences* satisfying it with equality, but SGD only produces (3) with equality when $\eta_k(\beta^2+\kappa) = \mu$, a constant step. The rate $n^{-2\eta h}$ is visible in Exercise 1.13(a) at $\eta = \frac{1}{2}$.

1.10 mini-batches

Let $\bar g(\theta) = \frac{1}{B}\sum_{b\le B} g(\theta; z^{(b)})$ with $z^{(1)}, \dots, z^{(B)}$ i.i.d. Then $\mathbb{E}\bar g = \nabla L(\theta)$, and since the B centered terms are independent, $\mathbb{E}\|\bar g - \nabla L\|^2 = \frac{1}{B^2}\sum_b \mathbb{E}\|g(\theta;z^{(b)}) - \nabla L\|^2 \le \frac{\sigma^2 + \kappa\|e\|^2}{B}$: (A3) with $(\sigma^2/B, \kappa/B)$. Lemma 1.1 now needs $\eta_k \le \mu/(\beta^2 + \kappa/B)$, Theorem 1.2 becomes $a_n \le (1-\mu\eta)^n a_0 + \eta\sigma^2/(B\mu)$ and Theorem 1.5 becomes $a_n \le \frac{1}{n+k_0}\max\big(\frac{\eta^2\sigma^2}{B(\eta\mu - 1)}, k_0 a_0\big)$ with $k_0 \ge \eta(\beta^2 + \kappa/B)/\mu$.

Fixed budget. Batch B, step η, n steps: bias $(1-\mu\eta)^n$, floor $\eta\sigma^2/(B\mu)$. Batch 1, step η/B, nB steps: floor $(\eta/B)\sigma^2/\mu$, the same; bias $(1-\mu\eta/B)^{nB}$, and $(1-x/B)^B \ge 1-x$ for $x \in [0,1]$, so the bias of the batch-1 scheme is slightly larger, but both are $\approx e^{-\mu\eta n}$. Per gradient evaluation the two are statistically equivalent: a batch of size B with step η is B small steps of size η/B taken at the same point instead of sequentially. What the batch buys is parallelism, and a larger admissible step: the constraint is $\eta \le \mu/(\beta^2 + \kappa/B)$ for the batch versus $\eta/B \le \mu/(\beta^2+\kappa)$, i.e. $\eta \le B\mu/(\beta^2+\kappa)$, for single samples. When $\kappa \gg \beta^2$ (multiplicative noise, as in least mean squares, $\kappa = d+1$) these are both $\approx B\mu/\kappa$ as long as $B \lesssim \kappa/\beta^2$; beyond that the batch constraint saturates at μ/β^2 while the single-sample scheme keeps scaling. So increasing B speeds up wall-clock convergence linearly up to a critical batch size of order κ/β^2, and is wasted beyond it.

1.11 step size $\eta_k \propto k^{-\alpha}$

The hypothesis $\eta_k \le \mu/(\beta^2+\kappa)$ makes (3) available. We show $a_k \le D\eta_k$ by induction, with $D \ge \max(2\sigma^2/\mu, a_0/\eta_0)$. The base case is $D \ge a_0/\eta_0$. If $a_k \le D\eta_k$, then by (3), $a_{k+1} \le D\eta_k - \mu D\eta_k^2 + \sigma^2\eta_k^2$, and we want this $\le D\eta_{k+1}$, i.e.

$$\eta_k^2\,(\mu D - \sigma^2) \;\ge\; D\,(\eta_k - \eta_{k+1}).$$

By the mean value theorem on $x \mapsto x^{-\alpha}$, $\eta_k - \eta_{k+1} = \eta\big(m^{-\alpha} - (m+1)^{-\alpha}\big) \le \eta\alpha m^{-\alpha-1} = \alpha\eta_k/m$ (equivalently $\eta_{k+1}/\eta_k \ge 1 - \alpha/m$). Since $\mu D - \sigma^2 \ge \mu D/2$, it suffices that $\eta_k^2\mu D/2 \ge D\alpha\eta_k/m$, i.e. $\eta_k\mu \ge 2\alpha/m$, i.e. $2\alpha m^{\alpha-1} \le \eta\mu$. As $\alpha < 1$ the left side decreases to 0, so this holds for all $m \ge k_0$ once $k_0 \ge (2\alpha/\eta\mu)^{1/(1-\alpha)}$. Hence $\mathbb{E}\|e_k\|^2 \le D\eta_k = O(k^{-\alpha})$, which is (7).

The mechanism: the contraction $\mu\eta_k$ per step removes a fraction $\mu\eta_k$ of the error, while the step size itself shrinks only by the fraction $\alpha/m \ll \mu\eta_k$; so the error tracks the moving target $\sigma^2\eta_k/\mu$

(the instantaneous floor of Theorem 1.2) with a lag. For $\alpha = 1$ the two fractions are of the same order $1/m$ and the comparison needs $\eta\mu > 1$: Lemma 1.4.

1.12 non-convex smooth losses

(a) Apply the descent inequality with $\theta = \theta_k$, $\theta' = \theta_{k+1} = \theta_k - \eta_k(\nabla L(\theta_k) + \xi_{k+1})$:

$$L(\theta_{k+1}) \le L(\theta_k) - \eta_k\langle \nabla L(\theta_k), \nabla L(\theta_k) + \xi_{k+1}\rangle + \tfrac{\beta}{2}\eta_k^2\|\nabla L(\theta_k) + \xi_{k+1}\|^2.$$

Take $\mathbb{E}_k$: $\nabla L(\theta_k) \in L^2(\mathcal{F}_k)$ and $\mathbb{E}_k\xi_{k+1} = 0$ kill the cross terms (rule (iii)), and $\mathbb{E}_k\|\xi_{k+1}\|^2 \le \sigma^2$, giving $\mathbb{E}_k L(\theta_{k+1}) \le L(\theta_k) - \eta_k(1 - \frac{\beta\eta_k}{2})\|\nabla L(\theta_k)\|^2 + \frac{\beta}{2}\eta_k^2\sigma^2$.

(b) For $\eta_k \le 1/\beta$, $1 - \beta\eta_k/2 \ge \frac{1}{2}$. Take expectations and sum over $k < n$:

$$\tfrac{1}{2}\sum_{k<n}\eta_k\,\mathbb{E}\|\nabla L(\theta_k)\|^2 \le L(\theta_0) - \mathbb{E}L(\theta_n) + \tfrac{\beta}{2}\sigma^2\sum_{k<n}\eta_k^2 \le L(\theta_0) - L_* + \tfrac{\beta}{2}\sigma^2\sum_{k<n}\eta_k^2.$$

Since $\sum_k \eta_k\mathbb{E}\|\nabla L(\theta_k)\|^2 \ge \big(\min_{k<n}\mathbb{E}\|\nabla L(\theta_k)\|^2\big)\sum_{k<n}\eta_k$, the displayed bound follows.

(c) Constant $\eta \le 1/\beta$: the bound is $\frac{2\Delta}{n\eta} + \beta\sigma^2\eta$, $\Delta = L(\theta_0) - L_*$, minimized at $\eta = \sqrt{2\Delta/(\beta\sigma^2 n)}$ (admissible for n large) with value $2\sigma\sqrt{2\beta\Delta/n}$: the rate $n^{-1/2}$ for the squared gradient norm, with a step that depends on the horizon. For $\eta_k = \eta k^{-1/2}$ ($k \ge 1$), $\sum_{k\le n}\eta_k \approx 2\eta\sqrt{n}$ and $\sum\eta_k^2 \approx \eta^2\log n$, so the bound is $O\big((\Delta/\eta + \beta\sigma^2\eta\log n)/\sqrt{n}\big)$: the same rate up to $\log n$, with no horizon to choose.

What it says and does not say. It says that *some* iterate among the first n has small expected squared gradient: the algorithm visits an approximate stationary point. (The minimizing index is not observable; the standard device is to output θ_K with K drawn at random with $\mathbb{P}(K = k) \propto \eta_k$, for which $\mathbb{E}\|\nabla L(\theta_K)\|^2$ equals the weighted average bounded above.) It does not say that the iterates converge, that the last iterate is good, that $L(\theta_k)$ is close to L_*, or anything about the distance to a minimizer: a saddle point or a local maximum is a stationary point, and without convexity a small gradient says nothing about the value of L. The constant-step floor $\beta\sigma^2\eta$ also reappears: with a constant step the iterates never settle, they wander with gradient norm of order $\sqrt{\beta\sigma^2\eta}$.

1.13 least mean squares, simulated

$d = 20$, $\sigma_\varepsilon = 1$, $\theta^* = \mathbf{1}/\sqrt{d}$, $\theta_0 = 0$, $n = 10^4$. Because the design is Gaussian, Example 1.6 makes a_k exactly computable, which checks the simulation: $\mathbb{E}_k\|e_{k+1}\|^2 = \mathbb{E}\|(I - \eta_k xx^\top)e_k\|^2 + \eta_k^2\mathbb{E}\|x\varepsilon\|^2 = (1 - 2\eta_k + (d+2)\eta_k^2)\|e_k\|^2 + \eta_k^2 d$ by Wick, so $a_{k+1} = (1 - 2\eta_k + (d+2)\eta_k^2)a_k + d\eta_k^2$.

(a) $\eta_k = \eta/(k + k_0)$, $k_0 = \eta(d+2)$. Values of $n\,\mathbb{E}\|\theta_n - \theta^*\|^2$:

η	0.5	1	2	4
simulation	52.8	20.7	26.6	45.1
exact recursion	54.5	20.3	26.7	45.6
Theorem 1.5: $\eta^2 d/(\eta - 1)$	–	–	80	107
Theorem 2.4: $\eta^2 d/(2\eta - 1)$	$\asymp \log n$	20	26.7	45.7

The simulation agrees with the exact recursion and with the CLT variance $\operatorname{tr}\Sigma_\eta = \eta^2 d/(2\eta - 1)$ of Lecture 17 ($H = I$, $\Sigma = \mathbb{E}[xx^\top\varepsilon^2] = I$), not with the bound $\eta^2 d/(\eta - 1)$ of Theorem 1.5, which is off by the factor $(2\eta - 1)/(\eta - 1)$ (3 at $\eta = 2$) for the reason in Exercise 2.12. At $\eta = 1$ ($\eta h = 1$) the last iterate reaches the least squares value d; at $\eta = \frac{1}{2}$ one is at the threshold $\eta h = \frac{1}{2}$ and $n\,a_n \approx \eta^2 d\log N = 0.25 \cdot 20 \cdot \log 10011 \approx 46$ plus lower-order terms, growing logarithmically. Theorem 1.5 does not apply at $\eta \le 1$ at all, yet $\eta = 1$ is the best choice.

(b) Constant $\eta = 1/(d+2)$: the time average of $\mathbb{E}\|e_k\|^2$ over the second half of the run is 0.913. The exact floor is $d\eta^2/(2\eta - (d+2)\eta^2) = d/(d+2) = 0.909$, and Theorem 1.2 gives $\eta\sigma^2/\mu = d/(d+2)$: the same number, because at $\eta = \mu/(\beta^2+\kappa)$ the inequality in Lemma 1.1 is an equality.

(c) $\eta_k = \eta k^{-0.7}$ with $\eta = 1/(d+2)$, the largest step admissible in Theorem 2.7 (and $\eta = 0.2/(d+2)$, $0.5/(d+2)$): $n\,\mathbb{E}\|\bar\theta_n - \theta^*\|^2 = 549$ (5096, 1989), nowhere near $d = 20$. Nothing is wrong with Theorem 2.7; the asymptotic regime has not been reached. The ODE time is $\sum_{k\le n}\eta_k \approx \eta n^{0.3}/0.3 = 2.3$ at $n = 10^4$ (4.7 at $n = 10^5$), so the initial error has been contracted only by $e^{-2.3} \approx 0.1$ at the end of the run and by much less for most of it; the average is dominated by the transient, i.e. by the term (I) of the proof, whose $O(n^{(\alpha-1)/2})$ decay has a large constant. At $n = 10^5$ the full average gives 313, and the tail average of Exercise 2.13 gives 51 against its prediction $2d = 40$. With $1/k$ steps the same ODE time is $\eta\log n = 18$ at $\eta = 2$, which is why (a) is already asymptotic. The lesson is the one in Exercise 2.13: averaging is only as good as the iterates it averages.

Lecture 17

2.11 the Lyapunov equation

In the eigenbasis the Lyapunov equation reads $(\eta h_i - \frac12 + \eta h_j - \frac12)(\Sigma_\eta)_{ij} = \eta^2\Sigma_{ij}$. With Σ diagonal the off-diagonal entries vanish, and $(2\eta h_i - 1)(\Sigma_\eta)_{ii} = \eta^2 s_i^2$ gives the claim; the solution is unique because $2\eta h_i - 1 > 0$. If $h_i \equiv h$, $\operatorname{tr}\Sigma_\eta = \frac{\eta^2}{2\eta h-1}\sum_i s_i^2$, and $\frac{d}{d\eta}\frac{\eta^2}{2\eta h-1} = \frac{2\eta(\eta h-1)}{(2\eta h-1)^2}$ vanishes at $\eta = 1/h$, where the value is $\sum_i s_i^2/h^2 = \operatorname{tr}(H^{-1}\Sigma H^{-1})$: the floor is attained because $\eta H = I$. For $d = 2$, $h = (1, 10)$, $s \equiv 1$: $f(\eta) = \frac{\eta^2}{2\eta-1} + \frac{\eta^2}{20\eta-1}$ on $\eta > \frac12$ has its minimum 1.052 at $\eta \approx 0.977$, while $\operatorname{tr}(H^{-1}\Sigma H^{-1}) = 1 + 0.01 = 1.01$. The gap is only 4% here because the stiff direction contributes little to the trace in any case; the point is rather that η must be tuned to $1/h_{\min}$, and that $\eta = 1/h_{\max} = 0.1$ is not even admissible ($\eta h_{\min} = 0.1 < \frac12$: the slow direction then has $n\,\mathbb{E}e_n^2 \to \infty$, Exercise 2.12). In the direction u_2 the attained variance is $0.977^2/18.5 = 0.052$ against the floor 0.01, five times larger.

2.12 the threshold $\eta h > \frac12$ is sharp

Unrolling (5) with $r \equiv 0$, $e_n = P_{n,0}e_0 - \sum_{k<n} w_{n,k}\xi_{k+1}$, with independent centered ξ's of variance σ^2, so $\mathbb{E}e_n^2 = P_{n,0}^2 e_0^2 + \sigma^2\sum_{k<n} w_{n,k}^2$. By the two-sided estimate in the proof of Lemma 2.6(i) (which only needs $k_0 \ge 2\eta h$), $w_{n,k} = \frac{\eta}{m}(m/N)^{\eta h}(1 + O(1/m))$ with the $O(1/m)$ bounded in sign and size, so

$$n\sum_{k<n} w_{n,k}^2 \;\asymp\; \eta^2 N^{1-2\eta h}\sum_{m=k_0}^{N-1} m^{2\eta h-2} \;\asymp\; \begin{cases} N^{1-2\eta h}\to\infty, & \eta h < \frac12 \text{ (the sum converges)},\\ \eta^2\log N, & \eta h = \frac12 \text{ (the sum is } \sum 1/m),\\ \eta^2/(2\eta h - 1), & \eta h > \frac12 \text{ (Lemma 2.6(ii))}. \end{cases}$$

The bias term behaves the same way: $nP_{n,0}^2 e_0^2 \asymp N^{1-2\eta h}k_0^{2\eta h}e_0^2$, which also diverges for $\eta h < \frac12$. So $n\,\mathbb{E}e_n^2 \to \infty$ for $\eta h < \frac12$ and $\asymp \log n$ for $\eta h = \frac12$, and the rate is $1/n$ exactly when $\eta h > \frac12$. The recursion (3) has contraction $1 - \eta_k h$ where the truth is $(1-\eta_k h)^2 \approx 1 - 2\eta_k h$; its threshold $\eta\mu > 1$ is the sharp one for a recursion with the halved contraction. For least mean squares this is the row $\eta = \frac12$ of Exercise 1.13(a).

2.13 tail averaging

Sum the identity $He_k = \frac{e_k - e_{k+1}}{\eta_k} - \xi_{k+1} - r_k$ over $n/2 \le k < n$ and multiply by $2/\sqrt{n}$:

$$\sqrt{n}\, H\bar{e}_n^{(1/2)} = \frac{2}{\sqrt{n}} \sum_{n/2\le k<n} \frac{e_k - e_{k+1}}{\eta_k} \;-\; \frac{2}{\sqrt{n}} \sum_{n/2\le k<n} \xi_{k+1} \;-\; \frac{2}{\sqrt{n}} \sum_{n/2\le k<n} r_k.$$

(I): summation by parts from $n/2$ gives $\frac{e_{n/2}}{\eta_{n/2}} - \frac{e_n}{\eta_{n-1}} + \sum_{n/2<k<n} e_k(\frac{1}{\eta_k} - \frac{1}{\eta_{k-1}})$; with (7) every term is $O(n^{-\alpha/2}n^{\alpha})$ and there are at most $n/2$ of them with coefficients $O(n^{\alpha-1})$, so $\mathbb{E}\|(\mathrm{I})\| = O(n^{-1/2}n^{\alpha/2}) = O(n^{(\alpha-1)/2}) \to 0$, as in the proof of Theorem 2.7 (the boundary term e_0/η_0 is replaced by $e_{n/2}/\eta_{n/2}$, which is smaller). (III): $\mathbb{E}\|(\mathrm{III})\| \le \frac{\gamma}{\sqrt{n}} \sum_{k\ge n/2} D\eta_k = O(n^{1/2-\alpha}) \to 0$. (II): for fixed v, the $\zeta_{n,k+1} = 2\langle v, \xi_{k+1}\rangle/\sqrt{n}$, $n/2 \le k < n$, are martingale differences with $V_n = \frac{4}{n}\sum_{n/2\le k<n} v^\top C(\theta_k)v$; (a) $\mathbb{E}|V_n - 2v^\top\Sigma v| \le \frac{4}{n}\sum_{n/2\le k<n} \|v\|^2\mathbb{E}\|C(\theta_k) - \Sigma\|_{\mathrm{op}} \to 0$ since there are $n/2$ terms, each tending to 0 by Lemma 2.3; (b) $\sum \mathbb{E}|\zeta|^3 \le \frac{n}{2}\cdot 8\|v\|^3\rho^3 n^{-3/2} \to 0$; (c) $\mathbb{E}\max \mathbb{E}_k\zeta^2 \le \frac{4}{n}\|v\|^2(\sigma^2 + \kappa\mathbb{E}\max_{k<n}\|e_k\|^2) = O(n^{-\alpha}) \to 0$. Hence (II) $\Rightarrow \mathcal{N}(0, 2\Sigma)$, and by Slutsky $\sqrt{n}\,\bar{e}_n^{(1/2)} \Rightarrow \mathcal{N}(0, 2H^{-1}\Sigma H^{-1})$: the variance of an average of $n/2$ terms.

Why accept the factor 2. The limit theorem is about $n \to \infty$; at a fixed n the full average $\bar\theta_n$ contains the whole transient. The iterates forget θ_0 at the rate $\exp(-\mu\sum_{k<j}\eta_k) \approx \exp(-\mu\eta j^{1-\alpha}/(1-\alpha))$, which for α near 1 and the small steps that Lemma 1.1 allows is slow: in Exercise 1.13(c) the ODE time $\sum_{k<n}\eta_k$ is 2.3 at $n = 10^4$, so a large fraction of the averaged iterates are still at distance $\asymp \|e_0\|$ from θ^*, and the term (I), although $O(n^{(\alpha-1)/2})$, has a constant proportional to $\|e_0\|$ with a huge prefactor. The tail average depends on θ_0 only through $e_{n/2}$, which is already contracted by $\exp(-\mu\sum_{k<n/2}\eta_k)$; its transient is exponentially smaller, at the price of a constant factor 2 in the variance that the limit theorem does see. The numbers in 1.13(c) (549 versus 230 at $n = 10^4$; 313 versus 51 at $n = 10^5$, floor 20 versus 40) show which effect dominates at realistic horizons. (Averaging the last $(1-c)n$ iterates with $c \to 0$ slowly recovers the constant 1 in the limit while still discarding the transient.)

2.14 constant step plus averaging

With $d = 1$, $\nabla L = he$, constant η and i.i.d. noise, $e_{k+1} = \varphi e_k - \eta\xi_{k+1}$ with $\varphi = 1 - \eta h \in (0,1)$: an AR(1) process. Solving, $e_k = \varphi^k e_0 - \eta\sum_{j<k}\varphi^{k-1-j}\xi_{j+1}$, and exchanging the order of summation,

$$\bar e_n = \frac{1}{n}\sum_{k<n} e_k = \frac{1}{n}\Big[e_0\frac{1-\varphi^n}{1-\varphi} - \eta\sum_{j<n}\xi_{j+1}\frac{1-\varphi^{n-1-j}}{1-\varphi}\Big].$$

The ξ's are independent, so

$$n\operatorname{Var}\bar e_n = \frac{\eta^2\sigma^2}{(1-\varphi)^2}\cdot\frac{1}{n}\sum_{j<n}\big(1-\varphi^{n-1-j}\big)^2 \;\longrightarrow\; \frac{\eta^2\sigma^2}{(1-\varphi)^2} = \frac{\eta^2\sigma^2}{\eta^2h^2} = \frac{\sigma^2}{h^2},$$

since all but $O(1/(1-\varphi))$ of the n terms are close to 1. This is $H^{-1}\Sigma H^{-1}$ in $d = 1$, with no condition on η beyond stability.

Where the step size went. The noise $\eta\xi_{j+1}$ injected at step j is remembered by the iterates for about $1/(\eta h)$ steps, with geometric weights φ^i; the average collects it with total weight $\eta\sum_{i\ge0}\varphi^i = \eta/(\eta h) = 1/h$. A larger step injects more noise per step but forgets it faster, and the two effects cancel exactly: $\bar e_n \approx -\frac{1}{h}\cdot\frac{1}{n}\sum_j\xi_{j+1}$, the full-sample estimator $-H^{-1}\bar\xi$ of Remark 2.5. The step size survives only in the bias $e_0(1-\varphi^n)/(n(1-\varphi)) \approx e_0/(n\eta h)$, i.e. in the length of the transient, which is $o(n^{-1/2})$ but larger for smaller η. The last iterate, by contrast, has variance $\eta\sigma^2/(h(2-\eta h))$

(Remark 1.3), which does not go to zero: averaging converts the $O(\eta)$ fluctuation of a stationary iterate into a $1/n$ variance.

In the linear case, then, a constant step plus averaging is efficient. It is the non-linearity that forces decreasing steps: when $r \not\equiv 0$ the stationary distribution of the constant-step iterates is not centered at θ^* (the remainder $\mathbb{E}r(\theta_k) \asymp \gamma\,\mathbb{E}\|e_k\|^2 \asymp \gamma\eta\sigma^2/\mu$ does not average out), so $\bar\theta_n$ converges to a point at distance $O(\eta)$ from θ^*, and η must go to zero for consistency.

2.15 averaging, simulated

Same setting as Exercise 1.13 ($d = 20$, $n = 10^4$, 200 runs); $n\,\mathbb{E}\|\cdot - \theta^*\|^2$ reported. Well-conditioned design, $H = \Sigma = I$:

	simulation	prediction
last, $\eta_k = 2/(k+k_0)$	26.5	Thm 1.5: $4d = 80$; Thm 2.4: $\eta^2 d/(2\eta - 1) = 26.7$
last, $\eta_k = 1/(k+k_0)$	20.7	Thm 2.4: $d = 20$ (floor attained, $\eta H = I$)
average, $\eta_k = \frac{c}{d+2}k^{-0.7}$	5096, 1989, 549 ($c = 0.2, 0.5, 1$)	Thm 2.7: $d = 20$

The last iterates are exactly where Lecture 17 puts them (and far below the Lecture 16 bound). The averaged iterates are in the transient, as explained in 1.13(c): the ODE time $\sum_k \eta_k = \frac{c}{d+2}\cdot\frac{n^{0.3}}{0.3} \le 2.3$ is too short to forget θ_0, and the smaller c the worse. At $n = 10^5$ the $c = 1$ average gives 313 and the tail average of Exercise 2.13 gives 51 (prediction $2d = 40$).

Ill-conditioned design, $x \sim \mathcal{N}(0, \mathrm{diag}(1, \dots, 1, 0.05))$, so $H = \mathrm{diag}(1, \dots, 1, 0.05)$, $\Sigma = \mathbb{E}[xx^\top \varepsilon^2] = H$, $\mu = \lambda_{\min} = 0.05$, and $H^{-1}\Sigma H^{-1} = H^{-1}$ has trace $19 + 20 = 39$:

	$n = 10^4$	$n = 10^5$
last iterate, $\eta_k = 2/(k+k_0)$	213	1273
last iterate, $\eta_k = 1/(k+k_0)$	314	2310
tail average, $\eta_k = \frac{1}{d+2}k^{-0.7}$	632	3381

Everything is now dominated by the slow coordinate. For the last iterates $\eta\lambda_{\min} = 0.1$ and 0.05 are below the threshold $\frac{1}{2}$, so by Exercise 2.12 $n\,\mathbb{E}e_n^2$ grows like $n^{1-2\eta\lambda_{\min}}$: $n^{0.8}$, i.e. by the factor 6.3 per decade, as observed ($213 \to 1273$). The 19 fast coordinates contribute their asymptotic $19\cdot 4/3 = 25$ and 19 respectively, invisible under the slow one. For averaging, the slow coordinate has been contracted by $\exp(-0.05\sum_k \eta_k) = e^{-0.23}$ at $n = 10^5$, i.e. not at all: the transient $n\,(e_0^{(d)})^2 = n/20$ is what is measured (5000 at $n = 10^5$, times $e^{-0.46}$).

Which methods needed to know the conditioning. The last iterate with $1/k$ steps needs $\eta > 1/(2\lambda_{\min})$ to have any $1/n$ rate at all, and $\eta = 1/\lambda_{\min}$ to be efficient in the slow direction (which then costs the factor $\eta^2 h/(2\eta h - 1) \approx \eta h/2 = 10$ in each fast direction): there is no good scalar step without knowing the spectrum. Averaging is asymptotically efficient for every c without knowing H, which is the content of Theorem 2.7; but the *time* it needs to enter the asymptotic regime is $\sum_k \eta_k \gg 1/\lambda_{\min}$, which no averaging can shorten, and at $\alpha = 0.7$ with the admissible η this is $n \gg (0.3\cdot(d+2)/\lambda_{\min})^{1/0.3} \gtrsim 10^7$ here. Knowing the conditioning is not needed to choose the method, but it is needed to know the horizon at which the guarantees begin, and in this example that horizon is out of reach: in practice one would precondition, or use a larger constant step for a burn-in phase (Exercise 2.14) before decreasing it.

2.16 non-gradient fields

(a) Symmetry of H enters at four places. (1) H has an orthonormal eigenbasis with real eigenvalues $h_i \in [\mu, \beta]$; the weights $P_{n,k}$, $w_{n,k}$, being polynomials in H, are diagonal in it, which reduces

Lemma 2.6 to scalars and gives $\|w_{n,k}\|_{\mathrm{op}} = \max_i w_{n,k}^{(i)}$. (2) In (II), $\langle v, w_{n,k}\xi_{k+1}\rangle = \langle w_{n,k}v, \xi_{k+1}\rangle$ uses $w_{n,k}^\top = w_{n,k}$. (3) The covariance Σ_η is written in the orthonormal eigenbasis, and the Lyapunov equation has $\Sigma_\eta(\eta H - \frac{1}{2}I)$ rather than $\Sigma_\eta(\eta H - \frac{1}{2}I)^\top$. (4) Behind Theorem 2.4 stands the rate $\mathbb{E}\|e_k\|^2 = O(1/k)$ of Theorem 1.5, which rests on (A1): $\langle \nabla L(\theta), e\rangle \geq \mu\|e\|^2$. For a general field this reads $\langle F(\theta), \theta - \theta^*\rangle \geq \mu\|\theta - \theta^*\|^2$ (strong monotonicity), a condition on the symmetric part $\frac{1}{2}(J + J^\top)$ near θ^*, not on the eigenvalues of J.

(b) Let $\tilde{e} = Q^{-1}e$, $\tilde{\xi} = Q^{-1}\xi$, $\tilde{r} = Q^{-1}r$. Then (5) becomes $\tilde{e}_{k+1} = (I - \eta_k\Lambda)\tilde{e}_k - \eta_k\tilde{\xi}_{k+1} - \eta_k\tilde{r}_k$ with Λ real diagonal, $\lambda_i \geq \mu$. Now the weights $\prod(I - \eta_j\Lambda)$ are diagonal, hence symmetric, and (1)–(3) hold verbatim with h_i replaced by λ_i and u_i by the standard basis. For (4), the Lecture 16 argument is run in the coordinates $\tilde{e}$: $\langle \Lambda\tilde{e}, \tilde{e}\rangle \geq \mu\|\tilde{e}\|^2$ is strong monotonicity of the linear part, and (A2)–(A5) transfer to $\tilde{F}(\tilde{\theta}) = Q^{-1}F(Q\tilde{\theta})$ with constants multiplied by powers of $\|Q\|_{\mathrm{op}}\|Q^{-1}\|_{\mathrm{op}}$ (the remainder $\|\tilde{r}\| \leq \frac{\gamma}{2}\|Q^{-1}\|\|Q\|^2\|\tilde{e}\|^2$, the noise $\tilde{C} = Q^{-1}CQ^{-\top}$). The conclusion is $\sqrt{n}\,\tilde{e}_n \Rightarrow \mathcal{N}(0, \tilde{\Sigma}_\eta)$ with $(\tilde{\Sigma}_\eta)_{ij} = \eta^2(Q^{-1}\Sigma Q^{-\top})_{ij}/(\eta(\lambda_i + \lambda_j) - 1)$, provided $\eta\lambda_{\min} > \frac{1}{2}$, and transforming back,

$$\sqrt{n}\,e_n \Rightarrow \mathcal{N}\big(0,\ Q\tilde{\Sigma}_\eta Q^\top\big), \qquad \text{equivalently} \qquad (\eta J - \tfrac{1}{2}I)\Sigma_\eta + \Sigma_\eta(\eta J - \tfrac{1}{2}I)^\top = \eta^2\Sigma,$$

the Lyapunov equation with the transpose in its proper place; it has a unique solution because all eigenvalues of $\eta J - \frac{1}{2}I$ have positive real part. With averaging (Theorem 2.7) the limit is $J^{-1}\Sigma J^{-\top}$: the proof reads the recursion backwards, $Je_k = (e_k - e_{k+1})/\eta_k - \xi_{k+1} - r_k$, and never diagonalizes anything, so only the rate (7), i.e. strong monotonicity, is needed from J.

(c) If $\lambda = a + ib$ with $a > 0$, the scalar weights become complex: $|1 - \eta_j\lambda|^2 = 1 - 2\eta_j a + \eta_j^2|\lambda|^2$, so $\log|1 - \eta_j\lambda| = -\eta_j a + O(\eta_j^2)$ and $|P_{n,k}^{(i)}| = (m/N)^{\eta a}(1 + O(1/m))$: Lemma 2.6(i) holds with ηh replaced by $\eta\,\mathrm{Re}\,\lambda$. The weights also rotate, $w_{n,k}^{(i)} \approx \frac{\eta}{m}(m/N)^{\eta a}e^{-i\eta b\log(m/N)}$, which does not affect (iii)–(iv), where only moduli enter, and in (ii) gives $n\sum_k w_{n,k}^{(i)}\overline{w_{n,k}^{(j)}} \to \eta^2/(\eta(\lambda_i + \bar{\lambda}_j) - 1)$, the complex Lyapunov equation (whose solution is a real covariance after transforming back, since the complex eigenvalues come in conjugate pairs). The condition $\eta\lambda_{\min} > \frac{1}{2}$ becomes

$$\eta\ \min_i \mathrm{Re}\,\lambda_i\ >\ \tfrac{1}{2},$$

and the admissibility condition $\eta_j|\lambda_i| \leq \frac{1}{2}$ fixes k_0. The imaginary parts make the iterates spiral into θ^* rather than approach it along straight lines; the mean-square rate depends only on how fast the spiral contracts, i.e. on the real parts.