

CS 182 / EECS 282A Official Course Notes

Fall 2026

Note 3: Gradient Descent and Stochastic Gradient Descent

Based on lectures by Prof. Gireeja Ranade and Prof. Anant Sahai

Prerequisites

Least-squares gradients, SVD coordinates, conditional expectation, and positive-semidefinite matrix inequalities. The row-space and minimum-norm facts from Note 2, *Least-Squares Optimization, Ridge, and Implicit Regularization* are used throughout.

Reading and coverage map

Prince correspondence (direct: Sections 6.2–6.3). Prince covers stochastic gradients, mini-batches, momentum, and their practical behavior.

JRS prerequisite review (pp. 104–108). JRS supplies concise background on gradient-based optimization and stochastic updates.

Beyond Prince. The course’s overparameterized constant-step contraction and its Lyapunov calculation, together with the RC-filter analogy for momentum, are developed explicitly here.

3.1 From full gradients to mini-batches

Suppose the empirical objective separates over examples:

$$f(\theta) = \frac{1}{n} \sum_{i=1}^n f_i(\theta), \quad \nabla f(\theta) = \frac{1}{n} \sum_{i=1}^n \nabla f_i(\theta). \quad (3.1)$$

Computing the full sum can be prohibitive when n is large. If B is sampled uniformly from all subsets of a fixed size b , the mini-batch estimator

$$g_B(\theta) = \frac{1}{b} \sum_{i \in B} \nabla f_i(\theta) \quad (3.2)$$

is unbiased:

$$\mathbb{E}_B[g_B(\theta)] = \nabla f(\theta). \quad (3.3)$$

Indeed, each example lies in B with probability b/n , so linearity of expectation gives Equation (3.3).

Stochastic gradient descent uses

$$\theta_{t+1} = \theta_t - \eta_t g_{B_t}(\theta_t). \quad (3.4)$$

SGD takes steps based on the gradient of only a few samples at a time. $b \ll n$ reduces per-step computation and activation memory. The noise can also move an iterate out of a shallow basin

or away from a saddle in a nonconvex landscape (Figure 3.1). That statement is possibility, not guarantee: the same noise can send the iterate toward a worse region. The direction is correct on average because in expectation the step is the same as the gradient, but we do not decrease the full loss on every step.

Early in training, a random initialization is usually far from useful behavior. At that scale, even a rough mini-batch direction can make meaningful progress without paying for the full gradient. This is why cheap early steps can be reasonable. It does not say that smaller batches are always better. As training approaches a solution, estimator variance can become the limiting effect.

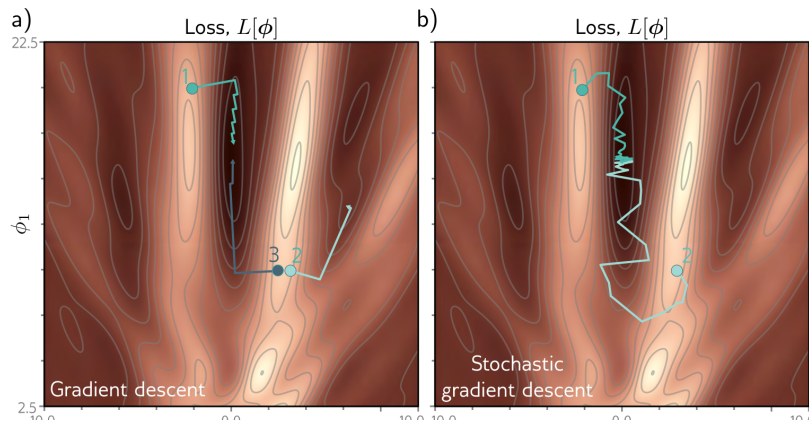


Figure 3.1: A nonconvex-landscape illustration adapted from Prince, Figure 6.5. Mini-batch noise can sometimes dislodge an iterate from a suboptimal valley, but the picture is an intuition rather than a guarantee that stochasticity finds a better minimum.

3.2 SGD convergence in an interpolating linear model

Remark

Analyzable special cases are debugging instruments. A complex training run can produce a plausible plot even when many lines of implementation are wrong; a case with a known qualitative and quantitative prediction lets one test the machinery. The same “guts-level” understanding is what supports designing the next optimizer rather than only calling a library routine.

Consider

$$X\vec{w} = \vec{y}, \quad X \in \mathbb{R}^{n \times d}, \quad d > n, \quad \text{rank}(X) = n. \quad (3.5)$$

The minimum-norm interpolator is

$$\vec{w}_* = X^\top (XX^\top)^{-1} \vec{y}. \quad (3.6)$$

Use batch size one, draw I_t uniformly from $\{1, \dots, n\}$, initialize $\vec{w}_0 = 0$, and update using the sampled squared loss:

$$\vec{w}_{t+1} = \vec{w}_t - 2\eta(x_{I_t}^\top \vec{w}_t - y_{I_t})x_{I_t}. \quad (3.7)$$

For $\vec{q}_t = \vec{w}_t - \vec{w}_*$, interpolation gives

$$\vec{q}_{t+1} = (I - 2\eta x_{I_t} x_{I_t}^\top) \vec{q}_t. \quad (3.8)$$

All iterates and errors remain in $\text{Row}(X)$ because the initialization and every update do.

3.2.1 Singular-value lower bound on the row space

Since $d > n$, the matrix $X^\top X$ has $d - n$ zero eigenvalues (the null-space directions) and n positive eigenvalues $\sigma_1^2 \geq \dots \geq \sigma_n^2 > 0$ (on $\text{Row}(X)$). For any $\vec{q} \in \text{Row}(X)$ the Rayleigh quotient sees only the positive eigenvalues, so

$$\|X\vec{q}\|_2^2 = \vec{q}^\top X^\top X \vec{q} \geq \sigma_{\min}^2 \|\vec{q}\|_2^2, \quad (3.9)$$

where $\sigma_{\min} = \sigma_n > 0$ is the smallest singular value of X .

3.2.2 Convergence proof

Let $\rho = \max_{1 \leq i \leq n} \|x_i\|_2$ be the largest row norm of X . If the learning rate satisfies

$$0 < \eta < \frac{1}{\rho^2}, \quad (3.10)$$

then the SGD iterates (3.7) satisfy

$$\mathbb{E} \|\vec{w}_t - \vec{w}_*\|_2^2 \leq \alpha^t \|\vec{w}_*\|_2^2,$$

where the contraction factor is

$$\alpha = 1 - \frac{4\eta(1 - \eta\rho^2)\sigma_{\min}^2}{n} \in [0, 1).$$

In particular, $\vec{w}_t \rightarrow \vec{w}_*$ in mean square.

To see this, we establish a one-step contraction conditionally on $\vec{q}_t$:

$$\mathbb{E}[\|\vec{q}_{t+1}\|_2^2 \mid \vec{q}_t] \leq \alpha \|\vec{q}_t\|_2^2.$$

Step 1: Expand the squared norm. From (3.8), expand $\|\vec{q}_{t+1}\|_2^2 = \|\vec{q}_t - 2\eta(x_{I_t}^\top \vec{q}_t)x_{I_t}\|_2^2$ using $\|\vec{a} - \vec{b}\|_2^2 = \|\vec{a}\|_2^2 - 2\vec{a}^\top \vec{b} + \|\vec{b}\|_2^2$:

$$\begin{aligned} \|\vec{q}_{t+1}\|_2^2 &= \|\vec{q}_t\|_2^2 - 4\eta(x_{I_t}^\top \vec{q}_t)^2 + 4\eta^2 \|x_{I_t}\|_2^2 (x_{I_t}^\top \vec{q}_t)^2 \\ &= \|\vec{q}_t\|_2^2 - 4\eta(1 - \eta\|x_{I_t}\|_2^2)(x_{I_t}^\top \vec{q}_t)^2. \end{aligned} \quad (3.11)$$

This is an identity—no approximation yet.

Step 2: Pointwise upper bound. The term being subtracted in (3.11) is $4\eta(1 - \eta\|x_{I_t}\|_2^2)(x_{I_t}^\top \vec{q}_t)^2$. We want an *upper* bound on $\|\vec{q}_{t+1}\|_2^2$, so we need a *lower* bound on this subtracted quantity.

Since $\|x_{I_t}\|_2 \leq \rho$ and $\eta > 0$, we have $\eta\|x_{I_t}\|_2^2 \leq \eta\rho^2 < 1$, so

$$1 - \eta\|x_{I_t}\|_2^2 \geq 1 - \eta\rho^2 > 0.$$

Therefore the subtracted term is nonnegative (each step *does* reduce the squared error), and replacing $(1 - \eta\|x_{I_t}\|_2^2)$ by the smaller quantity $(1 - \eta\rho^2)$ makes the subtracted term smaller, giving the upper bound:

$$\|\vec{q}_{t+1}\|_2^2 \leq \|\vec{q}_t\|_2^2 - 4\eta(1 - \eta\rho^2)(x_{I_t}^\top \vec{q}_t)^2. \quad (3.12)$$

Step 3: Conditional expectation. Conditioning on $\vec{q}_t$ means that $\vec{q}_t$ is treated as a fixed (nonrandom) vector; the only randomness in (3.12) comes from the uniform draw $I_t \in \{1, \dots, n\}$. The constant prefactor $4\eta(1 - \eta\rho^2)$ and the vector $\vec{q}_t$ pass through the expectation, leaving

$$\begin{aligned} \mathbb{E}[\|\vec{q}_{t+1}\|_2^2 \mid \vec{q}_t] &\leq \|\vec{q}_t\|_2^2 - 4\eta(1 - \eta\rho^2) \mathbb{E}[(x_{I_t}^\top \vec{q}_t)^2 \mid \vec{q}_t] \\ &= \|\vec{q}_t\|_2^2 - 4\eta(1 - \eta\rho^2) \frac{1}{n} \sum_{i=1}^n (x_i^\top \vec{q}_t)^2, \end{aligned} \quad (3.13)$$

where the second line uses $\Pr(I_t = i) = 1/n$ for each i .

Now recognize the sum as the squared norm of $X\vec{q}_t$: the i -th component of the vector $X\vec{q}_t \in \mathbb{R}^n$ is $x_i^\top \vec{q}_t$, so

$$\sum_{i=1}^n (x_i^\top \vec{q}_t)^2 = \|X\vec{q}_t\|_2^2. \quad (3.14)$$

Step 4: Singular-value lower bound. Since $\vec{q}_t \in \text{Row}(X)$ (by the row-space invariance above), inequality (3.9) gives $\|X\vec{q}_t\|_2^2 \geq \sigma_{\min}^2 \|\vec{q}_t\|_2^2$. Substituting into (3.13):

$$\mathbb{E}[\|\vec{q}_{t+1}\|_2^2 \mid \vec{q}_t] \leq \underbrace{\left(1 - \frac{4\eta(1 - \eta\rho^2) \sigma_{\min}^2}{n}\right)}_{\alpha} \|\vec{q}_t\|_2^2. \quad (3.15)$$

$\alpha < 1$: The quantity subtracted from 1 is strictly positive ($\eta > 0$, $1 - \eta\rho^2 > 0$ by the learning-rate bound, and $\sigma_{\min}^2 > 0$ since X has full row rank), so $\alpha < 1$.

$\alpha \geq 0$: The product $\eta(1 - \eta\rho^2)$ is a downward parabola in η , maximized at $\eta = 1/(2\rho^2)$ with value $1/(4\rho^2)$. Since $\sigma_{\min}^2 = \lambda_{\min}(XX^\top) \leq e_i^\top XX^\top e_i = \|x_i\|_2^2 \leq \rho^2$,

$$\frac{4\eta(1 - \eta\rho^2) \sigma_{\min}^2}{n} \leq \frac{4 \cdot \frac{1}{4\rho^2} \cdot \rho^2}{n} = \frac{1}{n} \leq 1,$$

confirming $\alpha \geq 0$.

Taking total (unconditional) expectations of both sides of the conditional contraction and using the tower property $\mathbb{E}[\mathbb{E}[Y \mid Z]] = \mathbb{E}[Y]$ gives

$$\mathbb{E}\|\vec{q}_{t+1}\|_2^2 \leq \alpha \mathbb{E}\|\vec{q}_t\|_2^2.$$

Iterating from $t = 0$:

$$\mathbb{E}\|\vec{q}_t\|_2^2 \leq \alpha \mathbb{E}\|\vec{q}_{t-1}\|_2^2 \leq \alpha^2 \mathbb{E}\|\vec{q}_{t-2}\|_2^2 \leq \dots \leq \alpha^t \mathbb{E}\|\vec{q}_0\|_2^2.$$

Since $\vec{w}_0 = 0$, we have $\vec{q}_0 = \vec{w}_0 - \vec{w}_* = -\vec{w}_*$, so $\|\vec{q}_0\|_2^2 = \|\vec{w}_*\|_2^2$. Because $\alpha \in [0, 1)$,

$$\mathbb{E}\|\vec{w}_t - \vec{w}_*\|_2^2 = \mathbb{E}\|\vec{q}_t\|_2^2 \leq \alpha^t \|\vec{w}_*\|_2^2 \longrightarrow 0 \quad \text{as } t \rightarrow \infty.$$

SGD therefore converges in mean square to the minimum-norm interpolator with a constant step under these assumptions.

Pitfall

Does strict decrease force convergence to zero? No. A nonnegative, strictly decreasing sequence may approach a positive limit. What closes this proof is the quantitative geometric

bound $\mathbb{E}[\|\vec{q}_{t+1}\|_2^2 \mid \vec{q}_t] \leq \alpha \|\vec{q}_t\|_2^2$ with $\alpha < 1$, not monotonicity by itself.

Pitfall

The theorem is not a blanket constant-step guarantee for stochastic optimization. It uses exact interpolation, full row rank, uniform independent row sampling, zero initialization, squared loss, and the explicit learning-rate bound. If sampled gradients remain noisy at every minimizer, a nonvanishing step generally leaves a stationary error neighborhood.

Recap

- Uniform fixed-size mini-batches give unbiased gradients, but an unbiased step need not decrease the loss on every draw.
- Momentum stores an exponentially weighted history in one vector and acts as a low-pass filter on noisy or oscillatory gradients.
- Acceleration guarantees require more structure and tuning than low-pass filtering alone.
- In an interpolating full-row-rank linear system, the sampled gradient noise vanishes at the solution.
- A quadratic Lyapunov function then contracts in conditional expectation under Equation (3.10), yielding mean-square convergence to the minimum-norm interpolator.

Further reading

- Ma et al. (2018). Explains why interpolation changes the stochastic-gradient story: near a shared zero-loss solution, mini-batch SGD can converge at a rate comparable to full-gradient descent.
- Polyak (1964). The classical heavy-ball method and its quadratic acceleration perspective.
- Bottou et al. (2018). A modern synthesis of optimization methods for large-scale machine learning, including stochastic-gradient tradeoffs and learning-rate schedules.

References and Further Reading

- Bottou, L., F. E. Curtis, and J. Nocedal (2018). “Optimization Methods for Large-Scale Machine Learning”. In: *SIAM Review* 60.2, pp. 223–311. DOI: [10.1137/16M1080173](https://doi.org/10.1137/16M1080173).
- Ma, S., R. Bassily, and M. Belkin (2018). “The Power of Interpolation: Understanding the Effectiveness of SGD in Modern Over-Parametrized Learning”. In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 3325–3334. URL: <https://proceedings.mlr.press/v80/ma18a.html>.
- Polyak, B. T. (1964). “Some Methods of Speeding Up the Convergence of Iteration Methods”. In: *USSR Computational Mathematics and Mathematical Physics* 4.5, pp. 1–17. DOI: [10.1016/0041-5553\(64\)90137-5](https://doi.org/10.1016/0041-5553(64)90137-5).