

L09

Stochastic Gradient Descent

CS 169 / CS 268
Introduction to Optimization

Ioannis Panageas

Fall 2026

An average of many losses

We want to minimize an average of N differentiable losses:

$$f(x) = \frac{1}{N} \sum_{i=1}^N f_i(x), \quad f'(x) = \frac{1}{N} \sum_{i=1}^N f'_i(x).$$

Full gradient descent evaluates all N derivatives for each step.

SGD uses a random sample to estimate the derivative.

The SGD update

At step k , draw i_k uniformly from $\{1, \dots, N\}$, independently of earlier samples.

$$g_k = f'_{i_k}(x_k), \quad x_{k+1} = x_k - \eta_k g_k.$$

x_0 is fixed. Each step size $\eta_k > 0$ is chosen before drawing the new sample.

The iterates are random because the sampled indices are random.

An unbiased derivative

$\mathbb{E}_k$ means expectation over the new sample, with all earlier samples fixed. In particular, x_k is fixed.

$$\mathbb{E}_k[g_k] = \frac{1}{N} \sum_{i=1}^N f'_i(x_k) = f'(x_k).$$

This is **unbiasedness**. The average direction is correct.

$\mathbb{E}$ averages over all samples, including the earlier ones.

The two-loss example

$$f_1(x) = \frac{1}{2}(x-1)^2, \quad f_2(x) = \frac{1}{2}(x+1)^2.$$

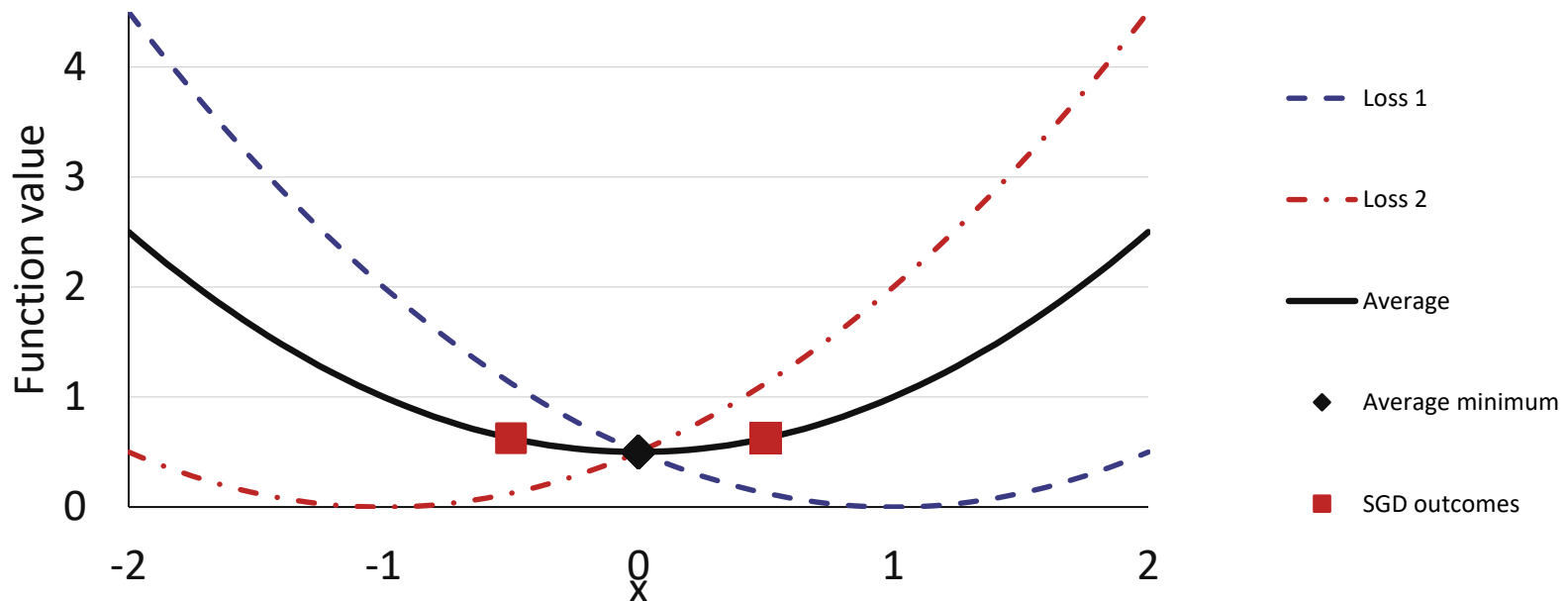
$$f(x) = \frac{f_1(x) + f_2(x)}{2} = \frac{x^2 + 1}{2}, \quad x^* = 0.$$

$$g_k = x_k - s_k, \quad s_k = \begin{cases} 1 & \text{with probability } 1/2, \\ -1 & \text{with probability } 1/2. \end{cases}$$

The signs $s_0, s_1, \dots$ are independent.

A sample can move uphill

$$x_0 = 0, \quad \eta = \frac{1}{2} : \quad x_1 = \pm \frac{1}{2}, \quad f(x_1) = \frac{5}{8} > f(0) = \frac{1}{2}.$$



At the minimum of the average loss, the individual derivatives are still nonzero.

Noise in the derivative

$$g_k - f'(x_k) = -s_k.$$

$$\mathbb{E}_k[g_k] = x_k, \quad \mathbb{E}_k[(g_k - x_k)^2] = 1.$$

As x_k approaches 0, the full derivative shrinks, while the sampling variance stays equal to 1.

This noise explains why a fixed step may keep moving near the minimum.

Variance and second moment

Assume the estimator is unbiased and its conditional variance is bounded:

$$\mathbb{E}_k[g_k] = f'(x_k), \quad \mathbb{E}_k[(g_k - f'(x_k))^2] \leq \sigma^2.$$

$$\mathbb{E}_k[g_k^2] = [f'(x_k)]^2 + \mathbb{E}_k[(g_k - f'(x_k))^2].$$

Variance measures noise. The **second moment** also includes the true derivative.

Mini-batches

Draw b independent samples with replacement at the same point x_k , and average their derivatives:

$$g_k^{(b)} = \frac{1}{b} \sum_{j=1}^b f'_{i_{k,j}}(x_k).$$

$$\mathbb{E}_k[g_k^{(b)}] = f'(x_k), \quad \mathbb{E}_k[(g_k^{(b)} - f'(x_k))^2] \leq \frac{\sigma^2}{b}.$$

Larger batches reduce noise, but cost b derivative evaluations per step.

Squared error in the example

$$x_{k+1} = (1 - \eta)x_k + \eta s_k, \quad 0 < \eta \leq 1.$$

$$x_{k+1}^2 = (1 - \eta)^2 x_k^2 + 2\eta(1 - \eta)x_k s_k + \eta^2 s_k^2,$$

$$\mathbb{E}_k[x_{k+1}^2] = (1 - \eta)^2 x_k^2 + \eta^2.$$

$$\boxed{\mathbb{E}[x_{k+1}^2] = (1 - \eta)^2 \mathbb{E}[x_k^2] + \eta^2.}$$

A constant step leaves an error

Unrolling the recurrence gives, for $k \geq 1$,

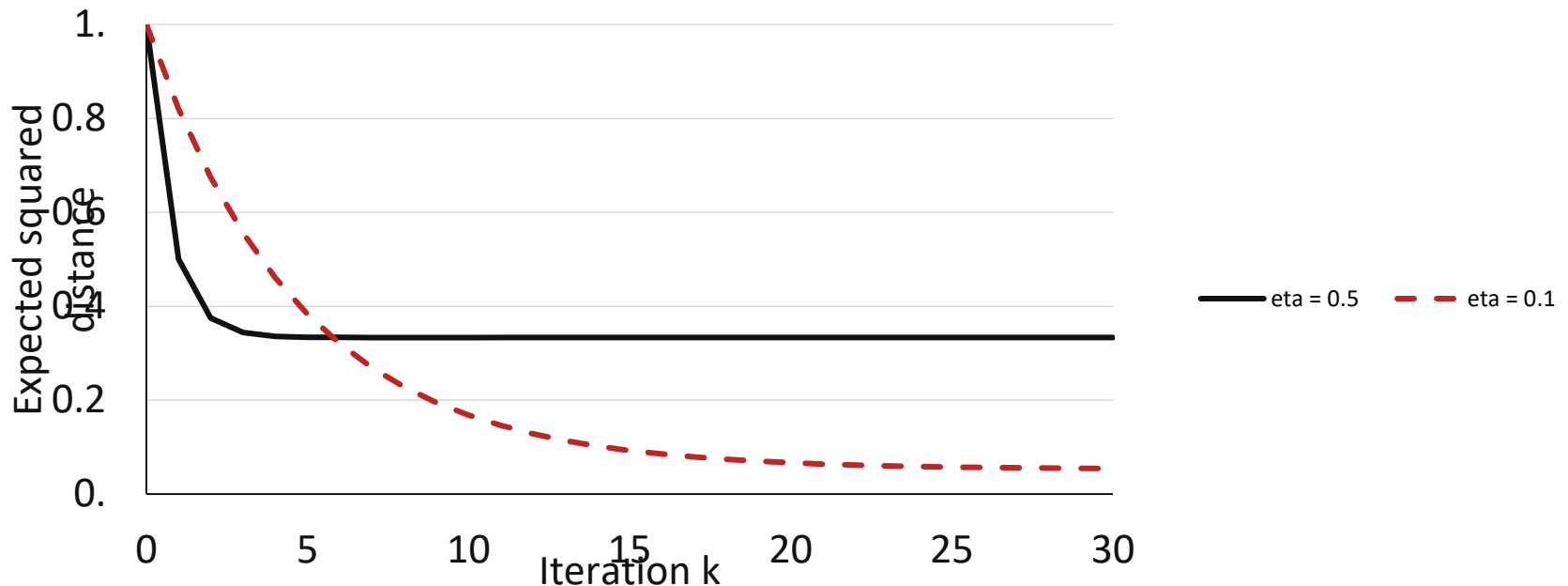
$$\mathbb{E}[x_k^2] = (1 - \eta)^{2k} x_0^2 + \frac{\eta}{2 - \eta} \left(1 - (1 - \eta)^{2k} \right).$$

$$\lim_{k \rightarrow \infty} \mathbb{E}[x_k^2] = \frac{\eta}{2 - \eta}, \quad \lim_{k \rightarrow \infty} \mathbb{E}[f(x_k) - f(0)] = \frac{\eta}{2(2 - \eta)}.$$

This positive limiting error is called a **noise floor**. A smaller step lowers it.

Two constant step sizes

$$x_0 = 1, \quad \eta = 0.5 \text{ or } 0.1.$$



A smaller step reduces the long-run error, but forgets the initial point more slowly.

A decreasing step size

$$\eta_k = \frac{1}{k+1}, \quad x_{k+1} = \frac{k}{k+1}x_k + \frac{s_k}{k+1}.$$

$$x_k = \frac{1}{k} \sum_{j=0}^{k-1} s_j \quad (k \geq 1).$$

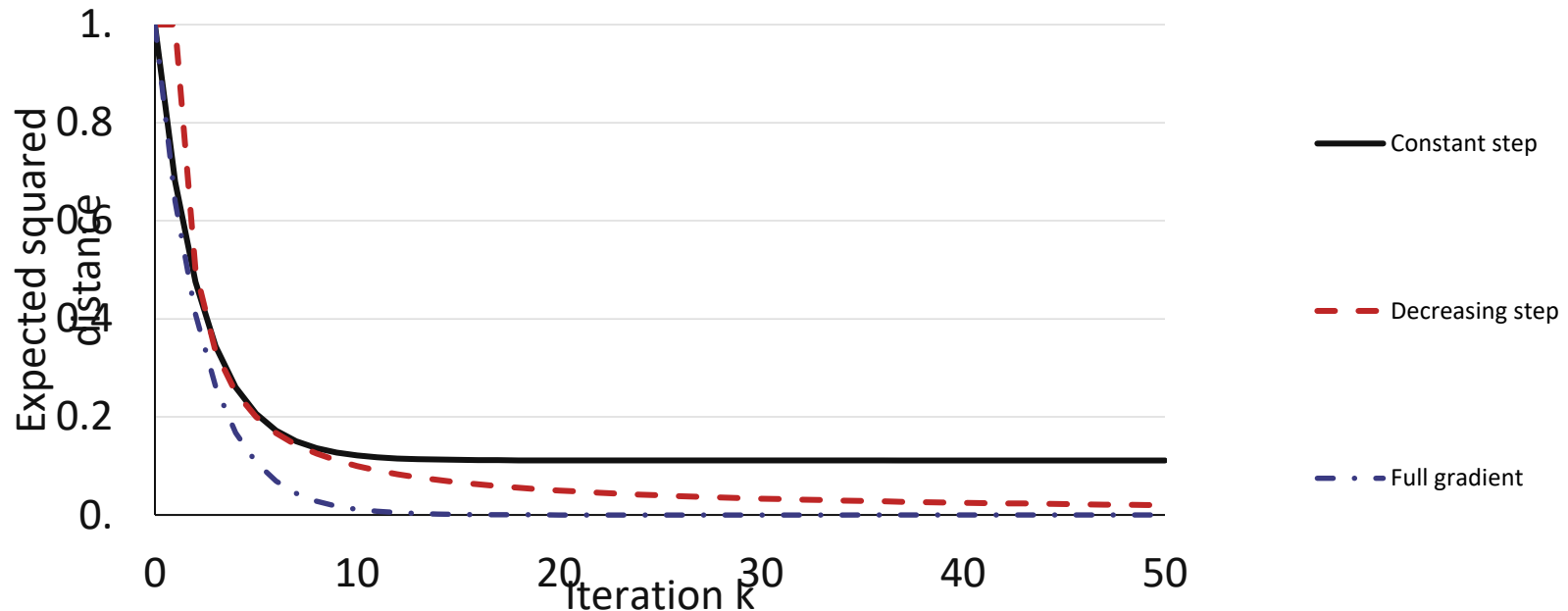
$$\mathbb{E}[x_k] = 0, \quad \mathbb{E}[x_k^2] = \frac{1}{k}, \quad \mathbb{E}[f(x_k) - f(0)] = \frac{1}{2k}.$$

These are exact identities for this quadratic example.

Constant and decreasing steps

Constant SGD and full GD: $\eta = 1/5$, $x_0 = 1$.

Decreasing SGD: $\eta_k = 1/(k+1)$.



Decreasing steps average away the noise in this example. Full gradient descent uses the exact derivative.

Convex functions: assumptions

Now let f be any differentiable convex function with an unconstrained minimizer x^* .

$$x_{k+1} = x_k - \eta g_k, \quad \eta > 0, \quad \mathbb{E}_k[g_k] = f'(x_k).$$

$$\mathbb{E}_k[g_k^2] \leq G^2, \quad |x_0 - x^*| \leq D.$$

The second-moment bound is an assumption along the iterates. Smoothness is not required for this proof.

The squared-distance identity

$$\begin{aligned}(x_{k+1} - x^*)^2 &= (x_k - x^* - \eta g_k)^2 \\ &= (x_k - x^*)^2 - 2\eta g_k(x_k - x^*) + \eta^2 g_k^2.\end{aligned}$$

Rearranging the cross term gives:

$$g_k(x_k - x^*) = \frac{(x_k - x^*)^2 - (x_{k+1} - x^*)^2}{2\eta} + \frac{\eta}{2} g_k^2.$$

Expected progress for convex f

Convexity gives $f(x_k) - f(x^*) \leq f'(x_k)(x_k - x^*)$.

$$\begin{aligned} f(x_k) - f(x^*) &\leq \mathbb{E}_k[g_k(x_k - x^*)] \\ &\leq \frac{(x_k - x^*)^2 - \mathbb{E}_k[(x_{k+1} - x^*)^2]}{2\eta} + \frac{\eta G^2}{2}. \end{aligned}$$

Now take expectation over the earlier samples as well.

Summing over the iterations

$$\begin{aligned}\sum_{k=0}^{K-1} \mathbb{E}[f(x_k) - f(x^*)] &\leq \frac{(x_0 - x^*)^2 - \mathbb{E}[(x_K - x^*)^2]}{2\eta} + \frac{K\eta G^2}{2} \\ &\leq \frac{D^2}{2\eta} + \frac{K\eta G^2}{2}.\end{aligned}$$

The squared distances **telescope**. Dividing by K gives:

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[f(x_k) - f(x^*)] \leq \frac{D^2}{2\eta K} + \frac{\eta G^2}{2}.$$

The averaged iterate

$$\bar{x}_K = \frac{1}{K} \sum_{k=0}^{K-1} x_k.$$

Convexity implies that the value at the average is no larger than the average of the values:

$$f(\bar{x}_K) \leq \frac{1}{K} \sum_{k=0}^{K-1} f(x_k).$$

$$\mathbb{E}[f(\bar{x}_K) - f(x^*)] \leq \frac{D^2}{2\eta K} + \frac{\eta G^2}{2}.$$

Choosing the step size

$$\mathbb{E}[f(\bar{x}_K) - f(x^*)] \leq \underbrace{\frac{D^2}{2\eta K}}_{\text{initial distance}} + \underbrace{\frac{\eta G^2}{2}}_{\text{stochastic term}}.$$

For a planned number K of steps and $D, G > 0$, balance the terms:

$$\eta = \frac{D}{G\sqrt{K}}, \quad \mathbb{E}[f(\bar{x}_K) - f(x^*)] \leq \frac{DG}{\sqrt{K}}.$$

This step is **constant during the run**; it depends on the planned horizon K .

Several variables

$$x_{k+1} = x_k - \eta g_k, \quad \mathbb{E}_k[g_k] = \nabla f(x_k).$$

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= \|x_k - x^*\|^2 - 2\eta g_k^T (x_k - x^*) \\ &\quad + \eta^2 \|g_k\|^2. \end{aligned}$$

With $\mathbb{E}_k \|g_k\|^2 \leq G^2$ and $\|x_0 - x^*\| \leq D$, the same convex proof gives

$$\mathbb{E}[f(\bar{x}_K) - f(x^*)] \leq \frac{D^2}{2\eta K} + \frac{\eta G^2}{2}.$$

Nonconvex functions: assumptions

Now assume f is L -smooth, with $L > 0$, and bounded below by f_{low} . Convexity is not assumed.

$$\mathbb{E}_k[g_k] = f'(x_k), \quad \mathbb{E}_k[(g_k - f'(x_k))^2] \leq \sigma^2.$$

$$x_{k+1} = x_k - \eta g_k, \quad 0 < \eta \leq \frac{1}{L}.$$

The goal is a **small derivative**, as in Lecture 7.

Smoothness in expectation

$$f(x_{k+1}) \leq f(x_k) - \eta f'(x_k)g_k + \frac{L\eta^2}{2}g_k^2.$$

$$\mathbb{E}_k[f(x_{k+1})] \leq f(x_k) - \eta \left(1 - \frac{L\eta}{2}\right) [f'(x_k)]^2 + \frac{L\eta^2\sigma^2}{2}$$

$$\leq f(x_k) - \frac{\eta}{2} [f'(x_k)]^2 + \frac{L\eta^2\sigma^2}{2}.$$

The last term is the contribution of **sampling noise**. It vanishes for an exact derivative.

A small expected squared derivative

Let $\Delta = f(x_0) - f_{\text{low}}$. Sum the previous bound over K steps:

$$\frac{\eta}{2} \sum_{k=0}^{K-1} \mathbb{E} [[f'(x_k)]^2] \leq \Delta + \frac{KL\eta^2\sigma^2}{2}.$$

$$\min_{0 \leq k < K} \mathbb{E} [[f'(x_k)]^2] \leq \frac{2\Delta}{\eta K} + L\eta\sigma^2.$$

Choosing $\eta = 1/(L\sqrt{K})$ gives an upper bound $(2L\Delta + \sigma^2)/\sqrt{K}$. This measures **stationarity**.

Vectors and constraints

For unconstrained $x \in \mathbb{R}^n$, the nonconvex proof replaces $[f'(x_k)]^2$ by $\|\nabla f(x_k)\|^2$.

For a nonempty closed convex feasible set C , project each stochastic step:

$$x_{k+1} = \Pi_C(x_k - \eta g_k), \quad \|\Pi_C(z) - x^*\| \leq \|z - x^*\| \quad (x^* \in C).$$

The convex averaged-iterate proof still works. At a boundary optimum, the ordinary gradient may be nonzero.

Reading

Chong, Lu, and Zak, An Introduction to Optimization, fifth edition.

Section 27.1: Stochastic Gradient Descent Algorithm.

Use these slides for the variance calculations, the averaged-iterate proof, and the nonconvex convergence bound.

Review Lecture 7 for the squared-distance identity and the smoothness inequality.

Conceptual questions

1. Can the expected iterate approach the minimizer while the expected objective gap stays positive? Explain using the two-loss example.
2. How do a smaller step size and a larger mini-batch affect noise, progress, and the cost of an iteration?