Randomized Algorithms

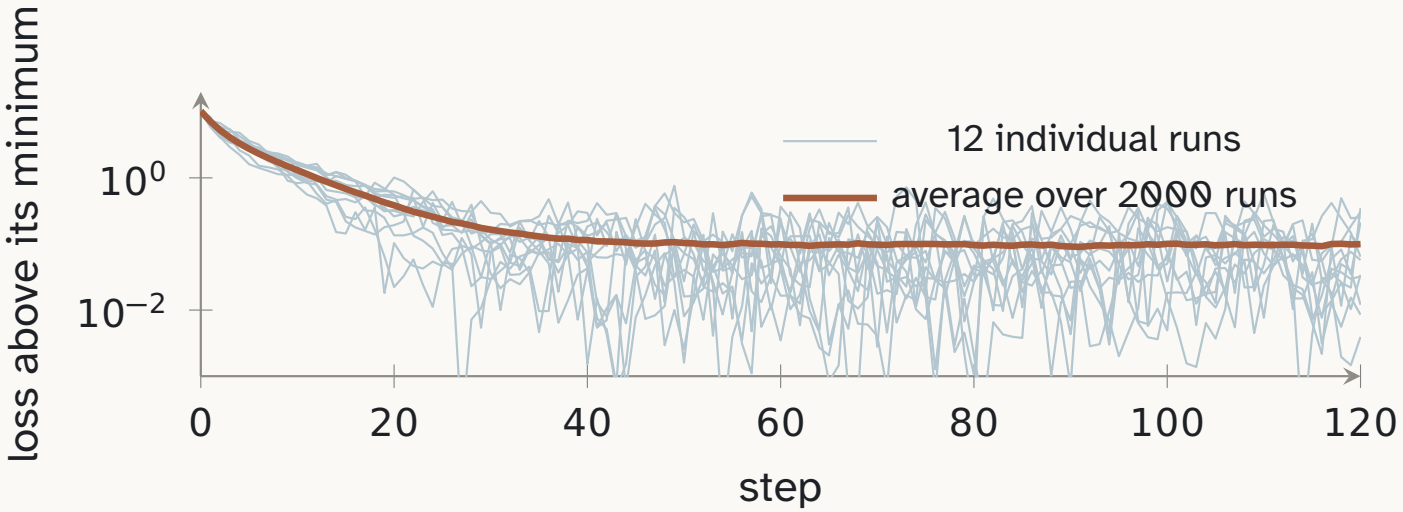
Lecture 20: Learning from random choices

Alex Gittens
Department of Computer Science, Rensselaer Polytechnic Institute

Fall 2026

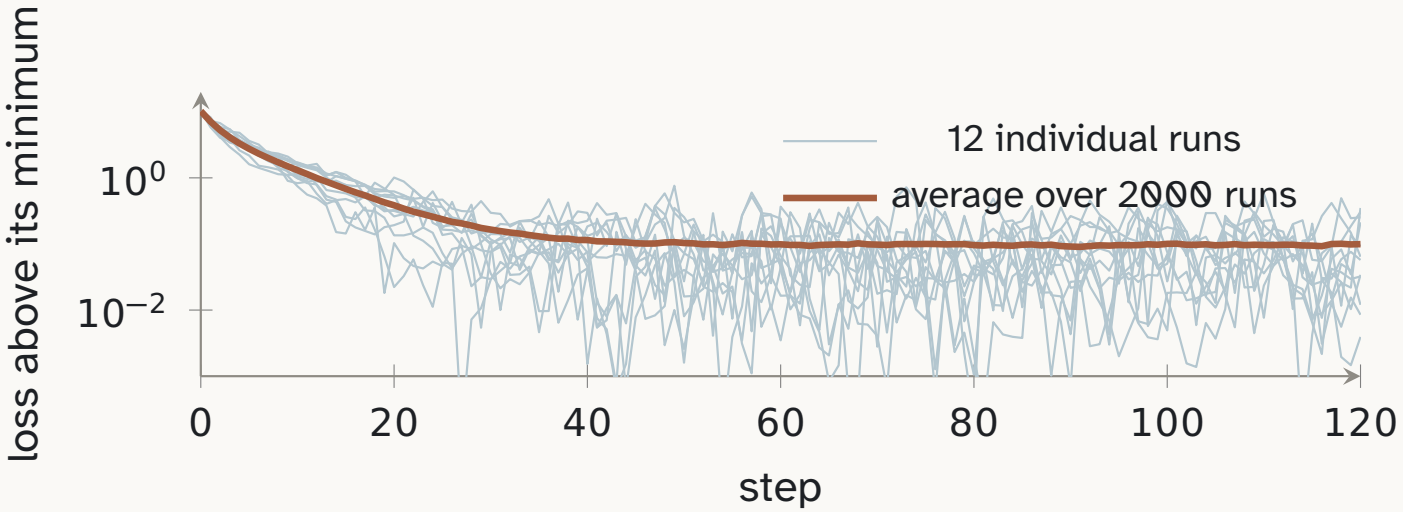
Good on average, but you only run it once

Many algorithms make a random choice at every step, and each choice changes what happens next.



Good on average, but you only run it once

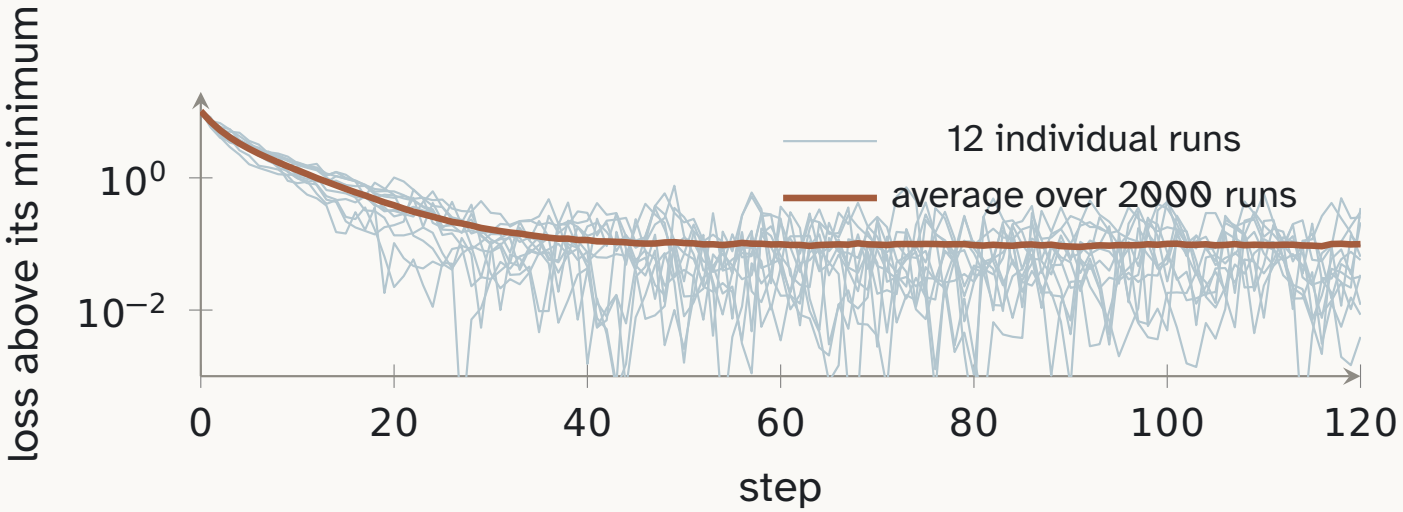
Many algorithms make a random choice at every step, and each choice changes what happens next.



A guarantee about the average does not tell you how *your* run went. **Today's question:** for two algorithms of this kind, will a single run be good?

Good on average, but you only run it once

Many algorithms make a random choice at every step, and each choice changes what happens next.



A guarantee about the average does not tell you how *your* run went. **Today's question:** for two algorithms of this kind, will a single run be good?

Choosing a route each day. Today's choice can change tomorrow's traffic.

Training a language model. Today's training sentence changes the parameters, and so tomorrow's update.

PART I

Choosing a route every morning

Each morning you choose a route to work. Afterward, a navigation app tells you how long every route took.

We assume nothing about traffic. What can we still guarantee?

The route problem

There are N routes. On each day $t = 1, \ldots, T$:

1. You choose a route I_t, using only what happened on earlier days.
2. The app reveals every route's travel time. Route i's time, rescaled to lie in $[0, 1]$, is its **cost** $\ell_{t,i}$.
3. You pay ℓ_{t,I_t}.

The route problem

There are N routes. On each day $t = 1, \ldots, T$:

1. You choose a route I_t, using only what happened on earlier days.
2. The app reveals every route's travel time. Route i's time, rescaled to lie in $[0, 1]$, is its **cost** $\ell_{t,i}$.
3. You pay ℓ_{t,I_t}.

We make *no* assumptions about traffic. Costs can be arbitrary, and can even react to our earlier choices (other drivers use the same app). We want a guarantee that holds for every cost sequence.

One rule of timing: today's costs are settled before today's choice is made.

The route problem

There are N routes. On each day $t = 1, \ldots, T$:

1. You choose a route I_t, using only what happened on earlier days.
2. The app reveals every route's travel time. Route i's time, rescaled to lie in $[0, 1]$, is its **cost** $\ell_{t,i}$.
3. You pay ℓ_{t,I_t}.

We make *no* assumptions about traffic. Costs can be arbitrary, and can even react to our earlier choices (other drivers use the same app). We want a guarantee that holds for every cost sequence.

One rule of timing: today's costs are settled before today's choice is made.

With no assumptions on the costs, what would count as doing well?

What counts as doing well?

Low total cost? If every route is jammed every day, every rule pays T. Total cost alone cannot separate good rules from bad ones.

What counts as doing well?

Low total cost? If every route is jammed every day, every rule pays T. Total cost alone cannot separate good rules from bad ones.

As well as the best route each day? Suppose that each day one route, chosen at random, is clear (cost 0) and the rest cost 1. Whatever we do, we find the clear route with probability $1/N$:

$$\mathbb{E}[\text{our cost per day}] = 1 - \tfrac{1}{N}, \qquad \text{best route each day} = 0.$$

The gap grows linearly in T for every rule. This target is impossible.

What counts as doing well?

Low total cost? If every route is jammed every day, every rule pays T. Total cost alone cannot separate good rules from bad ones.

As well as the best route each day? Suppose that each day one route, chosen at random, is clear (cost 0) and the rest cost 1. Whatever we do, we find the clear route with probability $1/N$:

$$\mathbb{E}[\text{our cost per day}] = 1 - \tfrac{1}{N}, \qquad \text{best route each day} = 0.$$

The gap grows linearly in T for every rule. This target is impossible.

As well as the best single route? Compare with the one route that would have been best had we taken it every day. We only learn which route that is at the end, but it is a fair target.

Regret

Regret

Let $C_{T,i} = \sum_{t=1}^{T} \ell_{t,i}$ be route i's total cost. Our **regret** is

$$\mathrm{Reg}_T = \sum_{t=1}^{T} \ell_{t,I_t} - \min_i C_{T,i}.$$

We want $\mathrm{Reg}_T / T \to 0$: averaged per day, we do as well as the best fixed route.

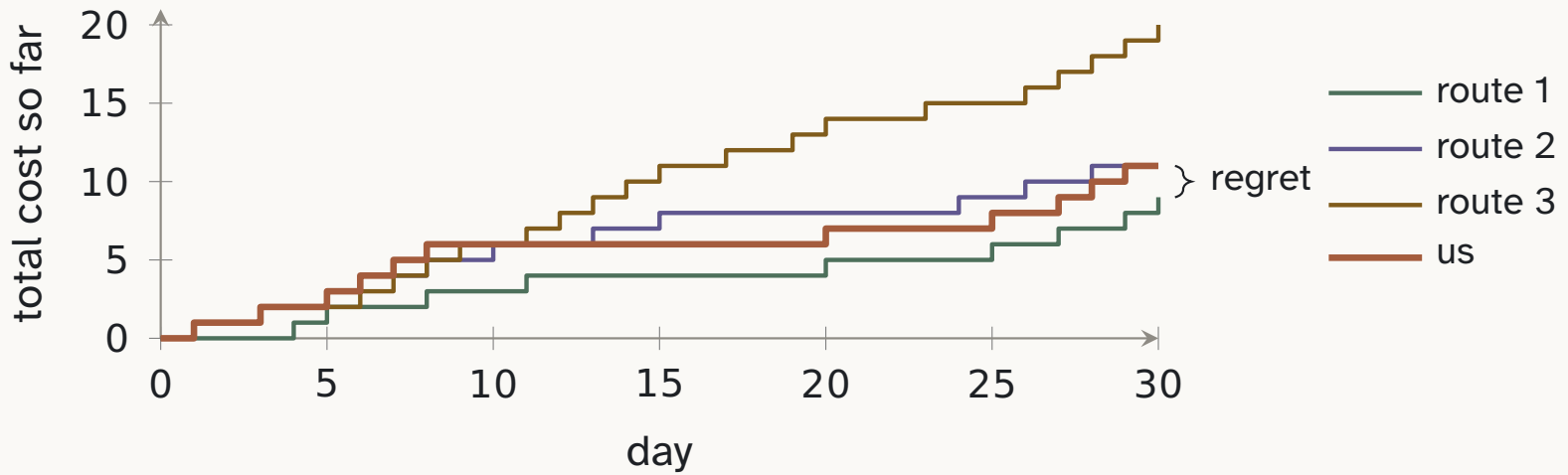
Regret

Regret

Let $C_{T,i} = \sum_{t=1}^{T} \ell_{t,i}$ be route i's total cost. Our **regret** is

$$\mathrm{Reg}_T = \sum_{t=1}^{T} \ell_{t,I_t} - \min_i C_{T,i}.$$

We want $\mathrm{Reg}_T / T \to 0$: averaged per day, we do as well as the best fixed route.



Why not follow the leader?

Follow the leader: each day, take the route with the smallest total cost so far.

Day	Route 1 cost	Route 2 cost	Leader that morning
1	1/2	0	tie
2	0	1	route 2
3	1	0	route 1
4	0	1	route 2

Why not follow the leader?

Follow the leader: each day, take the route with the smallest total cost so far.

Day	Route 1 cost	Route 2 cost	Leader that morning
1	1/2	0	tie
2	0	1	route 2
3	1	0	route 1
4	0	1	route 2

The leader always lands on the expensive route: it pays about T, while each route costs about $T/2$.

Why not follow the leader?

Follow the leader: each day, take the route with the smallest total cost so far.

Day	Route 1 cost	Route 2 cost	Leader that morning
1	1/2	0	tie
2	0	1	route 2
3	1	0	route 1
4	0	1	route 2

The leader always lands on the expensive route: it pays about T, while each route costs about $T/2$.

This traps *every* deterministic rule: costs that may depend on our past can predict our choice, and put cost 1 on it and 0 on the other route.

Claim

For $N = 2$, every deterministic rule has $\mathrm{Reg}_T \geq T/2$ on some cost sequence.

Why not follow the leader?

Follow the leader: each day, take the route with the smallest total cost so far.

Day	Route 1 cost	Route 2 cost	Leader that morning
1	1/2	0	tie
2	0	1	route 2
3	1	0	route 1
4	0	1	route 2

The leader always lands on the expensive route: it pays about T, while each route costs about $T/2$.

This traps *every* deterministic rule: costs that may depend on our past can predict our choice, and put cost 1 on it and 0 on the other route.

Claim

For $N = 2$, every deterministic rule has $\mathrm{Reg}_T \geq T/2$ on some cost sequence.

So we choose at random: today's costs are settled before our coin flip.

Give better routes more probability

Keep a probability for each route, larger when its total cost so far is smaller:

$$p_{t,i} = \frac{e^{-\eta C_{t-1,i}}}{\sum_j e^{-\eta C_{t-1,j}}}, \qquad C_{t-1,i} = \sum_{s<t} \ell_{s,i}.$$

Give better routes more probability

Keep a probability for each route, larger when its total cost so far is smaller:

$$p_{t,i} = \frac{e^{-\eta C_{t-1,i}}}{\sum_j e^{-\eta C_{t-1,j}}}, \qquad C_{t-1,i} = \sum_{s<t} \ell_{s,i}.$$

Each extra unit of total cost divides a route's odds by e^{η}:

$$C_{t-1,j} - C_{t-1,i} = 1 \quad \Longrightarrow \quad \frac{p_{t,i}}{p_{t,j}} = e^{\eta}.$$

Give better routes more probability

Keep a probability for each route, larger when its total cost so far is smaller:

$$p_{t,i} = \frac{e^{-\eta C_{t-1,i}}}{\sum_j e^{-\eta C_{t-1,j}}}, \qquad C_{t-1,i} = \sum_{s<t} \ell_{s,i}.$$

Each extra unit of total cost divides a route's odds by e^{η}:

$$C_{t-1,j} - C_{t-1,i} = 1 \quad \Longrightarrow \quad \frac{p_{t,i}}{p_{t,j}} = e^{\eta}.$$

The parameter $\eta > 0$ sets how strongly we react. As $\eta \to 0$ we choose uniformly at random; as $\eta \to \infty$ we follow the leader. The analysis will tell us how to set it.

The exponential-weights algorithm

Exponential weights

Start with weights $w_{1,i} = 1$. For $t = 1, \ldots, T$:

1. Set $p_{t,i} = w_{t,i}/W_t$, where $W_t = \sum_i w_{t,i}$.
2. Draw $I_t \sim p_t$ and pay ℓ_{t,I_t}.
3. See every route's cost and update $w_{t+1,i} = w_{t,i}\, e^{-\eta \ell_{t,i}}$.

Unrolling the update gives $w_{t,i} = e^{-\eta C_{t-1,i}}$, so these are the probabilities from the previous slide. Each day costs $O(N)$ time and memory.[1]

[1] Avoiding underflow: Appendix A2.

One day of updates

Take $\eta = \log 2$, so a cost of 1 halves a route's weight.

	Route 1	Route 2	Route 3
Weights	1	1	1
Probabilities	1/3	1/3	1/3
Today's costs	0	1	1

One day of updates

Take $\eta = \log 2$, so a cost of 1 halves a route's weight.

	Route 1	Route 2	Route 3
Weights	1	1	1
Probabilities	1/3	1/3	1/3
Today's costs	0	1	1
New weights	1	1/2	1/2

One day of updates

Take $\eta = \log 2$, so a cost of 1 halves a route's weight.

	Route 1	Route 2	Route 3
Weights	1	1	1
Probabilities	1/3	1/3	1/3
Today's costs	0	1	1
New weights	1	1/2	1/2
New probabilities	1/2	1/4	1/4

One day of updates

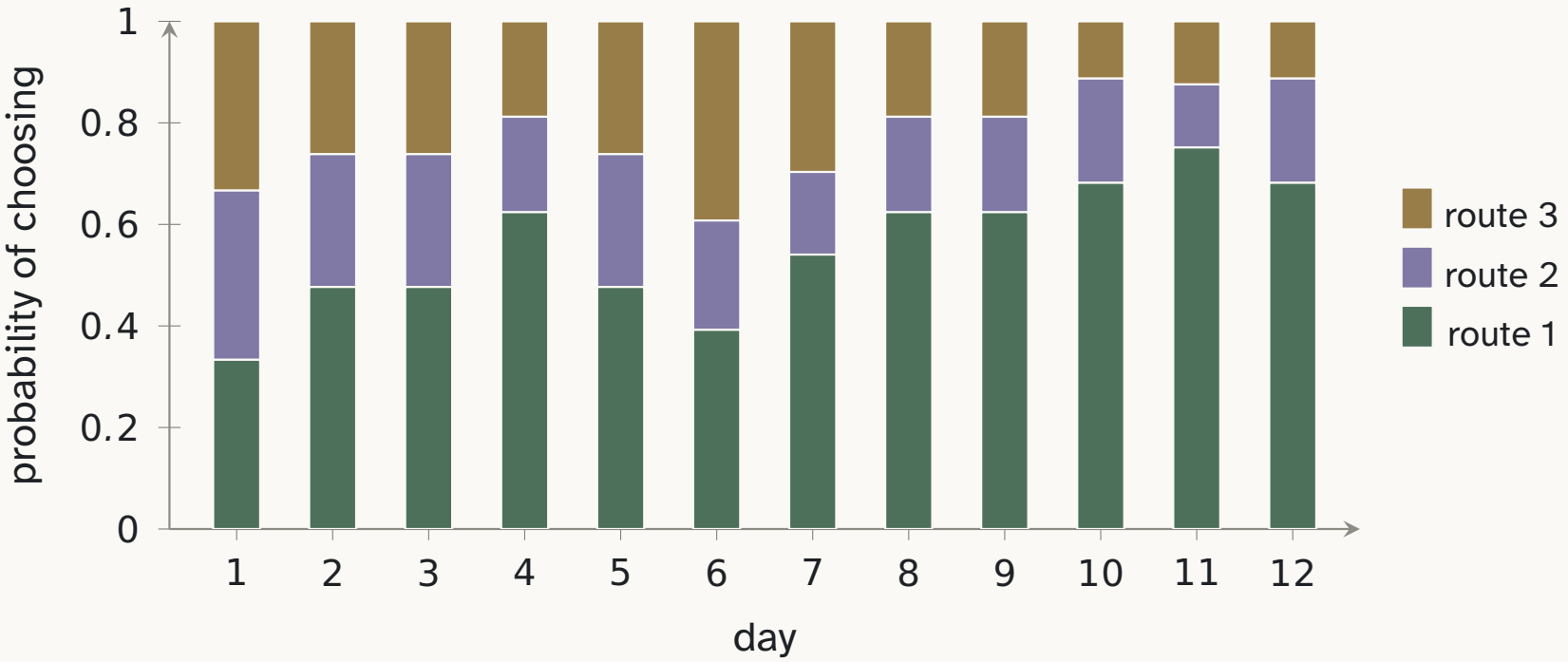
Take $\eta = \log 2$, so a cost of 1 halves a route's weight.

	Route 1	Route 2	Route 3
Weights	1	1	1
Probabilities	1/3	1/3	1/3
Today's costs	0	1	1
New weights	1	1/2	1/2
New probabilities	1/2	1/4	1/4

Today we paid 0 or 1, depending on the draw. Averaged over the draw we pay

$$p_t^\top \ell_t = \tfrac{1}{3}(0) + \tfrac{1}{3}(1) + \tfrac{1}{3}(1) = \tfrac{2}{3}.$$

Probability moves toward the better route



The first 12 days of the run in the regret plot ($\eta = 0.6$; average costs $0.35, 0.5, 0.65$). Route 1 gains probability, but an unlucky day still moves mass away from it.

Two sources of extra cost

Split the cost we pay into its average over today's draw and the luck of the draw:

$$\ell_{t,I_t} = p_t^\top \ell_t + D_t, \qquad D_t = \ell_{t,I_t} - p_t^\top \ell_t.$$

In the example, $D_t = -2/3$ with probability $1/3$ and $D_t = 1/3$ with probability $2/3$.

Two sources of extra cost

Split the cost we pay into its average over today's draw and the luck of the draw:

$$\ell_{t,I_t} = p_t^\top \ell_t + D_t, \qquad D_t = \ell_{t,I_t} - p_t^\top \ell_t.$$

In the example, $D_t = -2/3$ with probability $1/3$ and $D_t = 1/3$ with probability $2/3$.

Summing over the days,

$$\mathrm{Reg}_T = \Big(\sum_t p_t^\top \ell_t - \min_i C_{T,i}\Big) + \sum_t D_t.$$

No randomness once the costs are fixed

Luck of the draws

Two sources of extra cost

Split the cost we pay into its average over today's draw and the luck of the draw:

$$\ell_{t,I_t} = p_t^\top \ell_t + D_t, \qquad D_t = \ell_{t,I_t} - p_t^\top \ell_t.$$

In the example, $D_t = -2/3$ with probability $1/3$ and $D_t = 1/3$ with probability $2/3$.

Summing over the days,

$$\mathrm{Reg}_T = \Big(\sum_t p_t^\top \ell_t - \min_i C_{T,i}\Big) + \sum_t D_t.$$

No randomness once the costs are fixed — Luck of the draws

The blue term can be bounded by watching the weights. The red term is where luck enters, and to handle it we need a precise way to say "average over today's draw only".

Averaging over today's draw

Let $\mathcal{F}_{t-1}$ be everything settled before today's draw: earlier choices and today's costs. A quantity is **fixed given** $\mathcal{F}_{t-1}$ if $\mathcal{F}_{t-1}$ determines it; p_t and ℓ_t are. Then $\mathbb{E}[Z \mid \mathcal{F}_{t-1}]$ averages Z over today's draw only:

$$\mathbb{E}[\ell_{t,I_t} \mid \mathcal{F}_{t-1}] = \sum_i p_{t,i}\ell_{t,i} = p_t^\top \ell_t.$$

Averaging over today's draw

Let $\mathcal{F}_{t-1}$ be everything settled before today's draw: earlier choices and today's costs. A quantity is **fixed given** $\mathcal{F}_{t-1}$ if $\mathcal{F}_{t-1}$ determines it; p_t and ℓ_t are. Then $\mathbb{E}[\,Z \mid \mathcal{F}_{t-1}\,]$ averages Z over today's draw only:

$$\mathbb{E}[\,\ell_{t,I_t} \mid \mathcal{F}_{t-1}\,] = \sum_i p_{t,i}\ell_{t,i} = p_t^\top \ell_t.$$

Two rules follow from averaging over one draw at a time:

- **Pull out what is fixed:** if a is fixed given $\mathcal{F}_{t-1}$, then $\mathbb{E}[\,aZ \mid \mathcal{F}_{t-1}\,] = a\,\mathbb{E}[\,Z \mid \mathcal{F}_{t-1}\,]$.
- **Tower rule:** $\mathbb{E}\big[\mathbb{E}[\,Z \mid \mathcal{F}_{t-1}\,]\big] = \mathbb{E}Z$.

Averaging over today's draw

Let $\mathcal{F}_{t-1}$ be everything settled before today's draw: earlier choices and today's costs. A quantity is **fixed given** $\mathcal{F}_{t-1}$ if $\mathcal{F}_{t-1}$ determines it; p_t and ℓ_t are. Then $\mathbb{E}[Z \mid \mathcal{F}_{t-1}]$ averages Z over today's draw only:

$$\mathbb{E}[\ell_{t,I_t} \mid \mathcal{F}_{t-1}] = \sum_i p_{t,i}\ell_{t,i} = p_t^\top \ell_t.$$

Two rules follow from averaging over one draw at a time:

- **Pull out what is fixed:** if a is fixed given $\mathcal{F}_{t-1}$, then $\mathbb{E}[aZ \mid \mathcal{F}_{t-1}] = a\,\mathbb{E}[Z \mid \mathcal{F}_{t-1}]$.
- **Tower rule:** $\mathbb{E}\big[\mathbb{E}[Z \mid \mathcal{F}_{t-1}]\big] = \mathbb{E}Z$.

So $\mathbb{E}[D_t \mid \mathcal{F}_{t-1}] = p_t^\top \ell_t - p_t^\top \ell_t = 0$. Terms with conditional mean zero are **martingale differences**, and their running sums $M_t = \sum_{s \le t} D_s$ form a **martingale**.

Bounding the average costs

Lemma

For every cost sequence,

$$\sum_t p_t^\top \ell_t - \min_i C_{T,i} \leq \frac{\log N}{\eta} + \frac{\eta T}{8}.$$

Bounding the average costs

Lemma

For every cost sequence,

$$\sum_t p_t^\top \ell_t - \min_i C_{T,i} \le \frac{\log N}{\eta} + \frac{\eta T}{8}.$$

Idea. Watch the total weight $W_t = \sum_i w_{t,i}$.

- If some route is cheap, its weight stays large, so W_{T+1} cannot become too small.
- Each day W shrinks by a factor that depends on our average cost.

Comparing the two pins down our average cost.

A cheap route keeps the total weight large

Route i ends with weight $w_{T+1,i} = e^{-\eta C_{T,i}}$.

A cheap route keeps the total weight large

Route i ends with weight $w_{T+1,i} = e^{-\eta C_{T,i}}$.

The total includes route i, so

$$W_{T+1} = \sum_j w_{T+1,j} \geq e^{-\eta C_{T,i}}.$$

A cheap route keeps the total weight large

Route i ends with weight $w_{T+1,i} = e^{-\eta C_{T,i}}$.

The total includes route i, so

$$W_{T+1} = \sum_j w_{T+1,j} \geq e^{-\eta C_{T,i}}.$$

Take logs and use $W_1 = N$:

$$\log \frac{W_{T+1}}{W_1} \geq -\eta C_{T,i} - \log N \qquad \text{for every route } i.$$

Our average cost shrinks the total weight

Each day the total weight shrinks by

$$\frac{W_{t+1}}{W_t} = \sum_i p_{t,i} e^{-\eta \ell_{t,i}} = \mathbb{E}\left[e^{-\eta \ell_{t,I_t}} \,\middle|\, \mathcal{F}_{t-1} \right].$$

Our average cost shrinks the total weight

Each day the total weight shrinks by

$$\frac{W_{t+1}}{W_t} = \sum_i p_{t,i} e^{-\eta \ell_{t,i}} = \mathbb{E}\left[e^{-\eta \ell_{t,I_t}} \,\middle|\, \mathcal{F}_{t-1} \right].$$

This is the average of an exponential of a bounded random variable. Such averages are controlled by the following lemma.

Hoeffding's lemma[1]

If $a \le X \le b$ and $\mathbb{E}X = 0$, then for every real λ,

$$\mathbb{E}e^{\lambda X} \le e^{\lambda^2 (b-a)^2/8}.$$

The same holds with $\mathbb{E}[\,\cdot \mid \mathcal{F}_{t-1}]$ in place of $\mathbb{E}$ when a and b are fixed given $\mathcal{F}_{t-1}$.

[1] Proof: Appendix A3.

Applying Hoeffding's lemma

Today's cost is $p_t^\top \ell_t + D_t$, where $D_t \in \left[-p_t^\top \ell_t,\ 1 - p_t^\top \ell_t\right]$ has mean zero given $\mathcal{F}_{t-1}$, and the endpoints are fixed given $\mathcal{F}_{t-1}$. The width is 1.

Applying Hoeffding's lemma

Today's cost is $p_t^\top \ell_t + D_t$, where $D_t \in \left[-p_t^\top \ell_t,\ 1 - p_t^\top \ell_t\right]$ has mean zero given $\mathcal{F}_{t-1}$, and the endpoints are fixed given $\mathcal{F}_{t-1}$. The width is 1.

Pull out the fixed factor, then apply the lemma with $\lambda = -\eta$:

$$\mathbb{E}\left[e^{-\eta \ell_{t,I_t}} \,\middle|\, \mathcal{F}_{t-1}\right] = e^{-\eta p_t^\top \ell_t}\, \mathbb{E}\left[e^{-\eta D_t} \,\middle|\, \mathcal{F}_{t-1}\right] \le e^{-\eta p_t^\top \ell_t + \eta^2/8}.$$

So $\log(W_{t+1}/W_t) \le -\eta\, p_t^\top \ell_t + \eta^2/8$.

Applying Hoeffding's lemma

Today's cost is $p_t^\top \ell_t + D_t$, where $D_t \in \left[-p_t^\top \ell_t,\ 1 - p_t^\top \ell_t\right]$ has mean zero given $\mathcal{F}_{t-1}$, and the endpoints are fixed given $\mathcal{F}_{t-1}$. The width is 1.

Pull out the fixed factor, then apply the lemma with $\lambda = -\eta$:

$$\mathbb{E}\left[e^{-\eta \ell_{t,I_t}} \,\middle|\, \mathcal{F}_{t-1}\right] = e^{-\eta p_t^\top \ell_t}\, \mathbb{E}\left[e^{-\eta D_t} \,\middle|\, \mathcal{F}_{t-1}\right] \le e^{-\eta p_t^\top \ell_t + \eta^2/8}.$$

So $\log(W_{t+1}/W_t) \le -\eta\, p_t^\top \ell_t + \eta^2/8$.

Add over $t = 1, \ldots, T$; the left side telescopes:

$$\log \frac{W_{T+1}}{W_1} \le -\eta \sum_t p_t^\top \ell_t + \frac{\eta^2 T}{8}.$$

Comparing the two bounds

For every route i,

$$-\eta C_{T,i} - \log N \;\le\; \log \frac{W_{T+1}}{W_1} \;\le\; -\eta \sum_t p_t^\top \ell_t + \frac{\eta^2 T}{8}.$$

Comparing the two bounds

For every route i,

$$-\eta C_{T,i} - \log N \;\le\; \log \frac{W_{T+1}}{W_1} \;\le\; -\eta \sum_t p_t^\top \ell_t + \frac{\eta^2 T}{8}.$$

Rearrange and take the best route i:

$$\sum_t p_t^\top \ell_t - \min_i C_{T,i} \le \frac{\log N}{\eta} + \frac{\eta T}{8}. \qquad \square$$

Comparing the two bounds

For every route i,

$$-\eta C_{T,i} - \log N \;\le\; \log \frac{W_{T+1}}{W_1} \;\le\; -\eta \sum_t p_t^\top \ell_t + \frac{\eta^2 T}{8}.$$

Rearrange and take the best route i:

$$\sum_t p_t^\top \ell_t - \min_i C_{T,i} \le \frac{\log N}{\eta} + \frac{\eta T}{8}. \qquad \square$$

Small η makes the first term large (we learn slowly); large η makes the second large (we overreact). Balancing them with $\eta = \sqrt{8 \log N / T}$ gives[1]

$$\frac{1}{T}\Big(\sum_t p_t^\top \ell_t - \min_i C_{T,i}\Big) \le \sqrt{\frac{\log N}{2T}}.$$

[1] When T is not known in advance: Appendix A5.

Sums of martingale differences

Azuma's inequality[1]

Suppose $\mathbb{E}[D_t \mid \mathcal{F}_{t-1}] = 0$ and $A_t \le D_t \le A_t + c_t$ for each t, where A_t is fixed given $\mathcal{F}_{t-1}$ and the width c_t is a fixed number. Then for every $s > 0$,

$$\mathbb{P}\Big(\sum_{t=1}^{T} D_t \ge s\Big) \le \exp\Big(-\frac{2s^2}{\sum_t c_t^2}\Big).$$

Equivalently, with probability at least $1 - \delta$,

$$\sum_t D_t \le \sqrt{\tfrac{1}{2}\textstyle\sum_t c_t^2 \, \log(1/\delta)}.$$

[1] Proof: Appendix A6.

Sums of martingale differences

Azuma's inequality[1]

Suppose $\mathbb{E}[D_t \mid \mathcal{F}_{t-1}] = 0$ and $A_t \le D_t \le A_t + c_t$ for each t, where A_t is fixed given $\mathcal{F}_{t-1}$ and the width c_t is a fixed number. Then for every $s > 0$,

$$\mathbb{P}\Big(\sum_{t=1}^{T} D_t \ge s\Big) \le \exp\Big(-\frac{2s^2}{\sum_t c_t^2}\Big).$$

Equivalently, with probability at least $1 - \delta$,

$$\sum_t D_t \le \sqrt{\tfrac{1}{2}\textstyle\sum_t c_t^2 \, \log(1/\delta)}.$$

The D_t need not be independent. What matters is mean zero given the past and a width known in advance.

[1] Proof: Appendix A6.

Sums of martingale differences

Azuma's inequality[1]

Suppose $\mathbb{E}[D_t \mid \mathcal{F}_{t-1}] = 0$ and $A_t \le D_t \le A_t + c_t$ for each t, where A_t is fixed given $\mathcal{F}_{t-1}$ and the width c_t is a fixed number. Then for every $s > 0$,

$$\mathbb{P}\Big(\sum_{t=1}^{T} D_t \ge s\Big) \le \exp\Big(-\frac{2s^2}{\sum_t c_t^2}\Big).$$

Equivalently, with probability at least $1 - \delta$,

$$\sum_t D_t \le \sqrt{\tfrac{1}{2}\textstyle\sum_t c_t^2 \, \log(1/\delta)}.$$

The D_t need not be independent. What matters is mean zero given the past and a width known in advance.

For a bound on $\left|\sum_t D_t\right|$, apply it to $-D_t$ too: $\log(1/\delta)$ becomes $\log(2/\delta)$.

[1] Proof: Appendix A6.

The luck of the draws

Tomorrow's costs may depend on today's choice, so the D_t are not independent.[1] Azuma does not need independence: D_t has mean zero given $\mathcal{F}_{t-1}$ and lies in

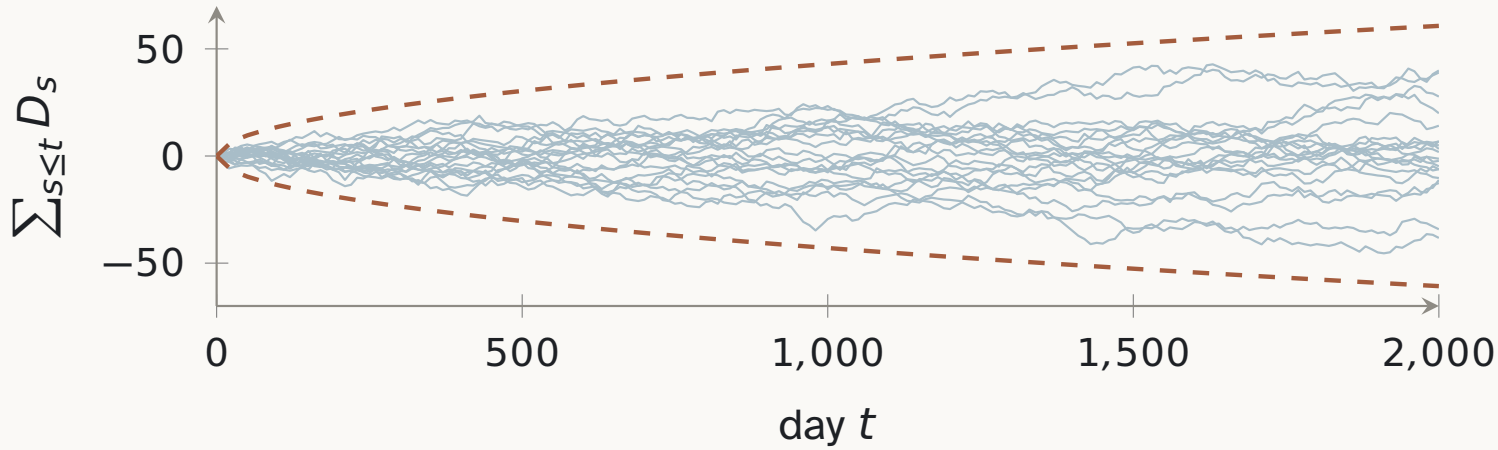
$$D_t \in \left[-p_t^\top \ell_t,\ 1 - p_t^\top \ell_t\right], \qquad \text{width } c_t = 1.$$

[1] A two-day example: Appendix A7.

The luck of the draws

Tomorrow's costs may depend on today's choice, so the D_t are not independent.[1] Azuma does not need independence: D_t has mean zero given $\mathcal{F}_{t-1}$ and lies in

$$D_t \in \left[-p_t^\top \ell_t,\ 1 - p_t^\top \ell_t\right], \qquad \text{width } c_t = 1.$$



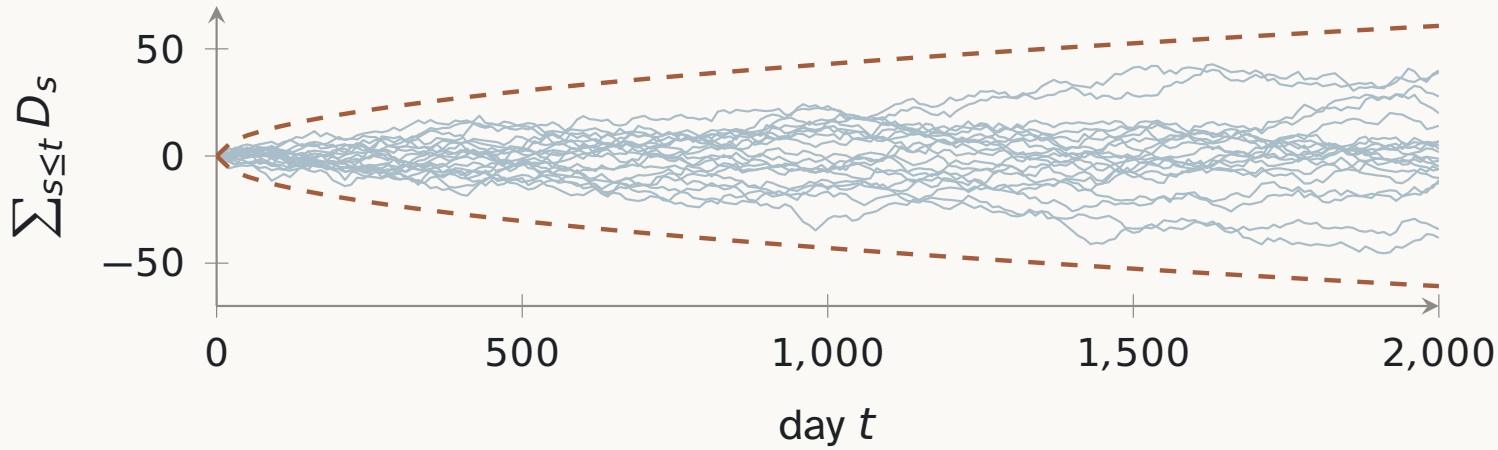
20 runs, $N = 3$; each day the route we took yesterday costs 1 and the others 0. Dashed: Azuma's two-sided 95% bound at each fixed day.

[1] A two-day example: Appendix A7.

The luck of the draws

Tomorrow's costs may depend on today's choice, so the D_t are not independent.[1] Azuma does not need independence: D_t has mean zero given $\mathcal{F}_{t-1}$ and lies in

$$D_t \in \left[-p_t^\top \ell_t,\ 1 - p_t^\top \ell_t\right], \qquad \text{width } c_t = 1.$$



20 runs, $N = 3$; each day the route we took yesterday costs 1 and the others 0. Dashed: Azuma's two-sided 95% bound at each fixed day.

Azuma with $c_t = 1$: with probability at least $1 - \delta$,

$$\sum_t D_t \le \sqrt{\frac{T}{2} \log(1/\delta)}.$$

[1] A two-day example: Appendix A7.

Exponential weights in a single run

Theorem 1 (exponential weights)

Take $\eta = \sqrt{8 \log N/T}$. For every way the costs may react to our choices, with probability at least $1 - \delta$,

$$\frac{\mathrm{Reg}_T}{T} \leq \sqrt{\frac{\log N}{2T}} + \sqrt{\frac{\log(1/\delta)}{2T}}.$$

Learning which route is good — Luck of the draws

Proof. Add the lemma and the Azuma bound, then divide by T. □

Exponential weights in a single run

Theorem 1 (exponential weights)

Take $\eta = \sqrt{8 \log N/T}$. For every way the costs may react to our choices, with probability at least $1 - \delta$,

$$\frac{\mathrm{Reg}_T}{T} \leq \sqrt{\frac{\log N}{2T}} + \sqrt{\frac{\log(1/\delta)}{2T}}.$$

Learning which route is good — Luck of the draws

Proof. Add the lemma and the Azuma bound, then divide by T. □

Regret per day falls like $1/\sqrt{T}$. The number of routes enters only through $\log N$: going from 2 routes to 1000 multiplies the days needed by about 10.

Exponential weights in a single run

Theorem 1 (exponential weights)

Take $\eta = \sqrt{8 \log N / T}$. For every way the costs may react to our choices, with probability at least $1 - \delta$,

$$\frac{\mathrm{Reg}_T}{T} \le \sqrt{\frac{\log N}{2T}} + \sqrt{\frac{\log(1/\delta)}{2T}}.$$

Learning which route is good — Luck of the draws

Proof. Add the lemma and the Azuma bound, then divide by T. □

Regret per day falls like $1/\sqrt{T}$. The number of routes enters only through $\log N$: going from 2 routes to 1000 multiplies the days needed by about 10.

With "routes" replaced by experts, classifiers, or strategies, the same algorithm underlies boosting and fast algorithms for solving zero-sum games.

PART II

Training a language model

A language model learns by adjusting its parameters to make the text it has seen more probable. The training text is enormous.

Can we make progress while reading only one example per step?

Predicting the next token

She opened her laptop and started to _____

A model with parameters x assigns a probability $P_x(y \mid c)$ to each possible next token y after the context c. Here $y =$ write.

Predicting the next token

She opened her laptop and started to _____

A model with parameters x assigns a probability $P_x(y \mid c)$ to each possible next token y after the context c. Here $y =$ write.

	Model A	Model B
$P_x(\text{write} \mid c)$	0.10	0.50
$-\log P_x(\text{write} \mid c)$	2.30	0.69

Predicting the next token

She opened her laptop and started to _____

A model with parameters x assigns a probability $P_x(y \mid c)$ to each possible next token y after the context c. Here $y =$ write.

	Model A	Model B
$P_x(\text{write} \mid c)$	0.10	0.50
$-\log P_x(\text{write} \mid c)$	2.30	0.69

The better model has the smaller **log loss** $-\log P_x(y \mid c)$. Training means changing x to reduce the log loss of the tokens that actually appeared.

The training loss averages over the corpus

Context c_i	Next token y_i
She opened her laptop and started to	write
He put on his headphones and pressed	play
The final song brought the crowd to its	feet

$$f_i(x) = -\log P_x(y_i \mid c_i), \qquad f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x).$$

The training loss averages over the corpus

Context c_i	Next token y_i
She opened her laptop and started to	write
He put on his headphones and pressed	play
The final song brought the crowd to its	feet

$$f_i(x) = -\log P_x(y_i \mid c_i), \qquad f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x).$$

One parameter vector serves every example, and its gradient is an average too:

$$\nabla f(x) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x).$$

The training loss averages over the corpus

Context c_i	Next token y_i
She opened her laptop and started to	write
He put on his headphones and pressed	play
The final song brought the crowd to its	feet

$$f_i(x) = -\log P_x(y_i \mid c_i), \qquad f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x).$$

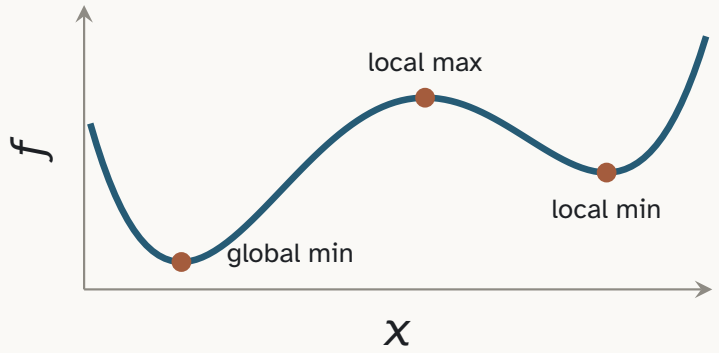
One parameter vector serves every example, and its gradient is an average too:

$$\nabla f(x) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x).$$

Modern models are trained on trillions of tokens. Computing ∇f *once* means a pass over all of them.

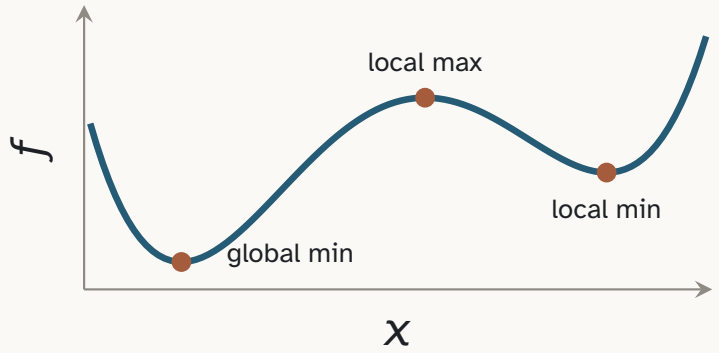
What can we hope to find?

We would like a minimizer of f. But for neural networks f can have many valleys and flat regions, and no efficient method is known that finds the lowest point of every such function.



What can we hope to find?

We would like a minimizer of f. But for neural networks f can have many valleys and flat regions, and no efficient method is known that finds the lowest point of every such function.



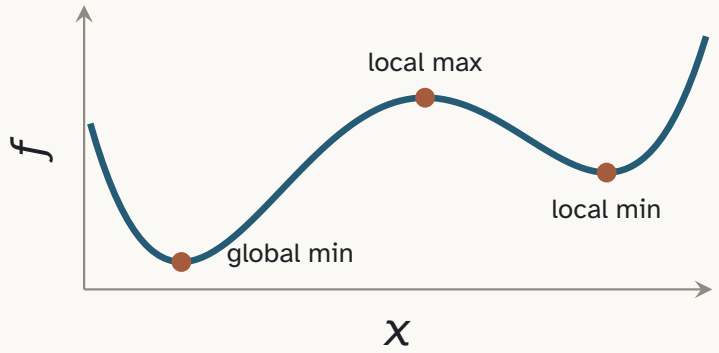
A weaker target: a point with a small gradient. For a step $h = -\eta \nabla f(x)$ with step size $\eta > 0$,

$$f(x+h) = f(x) - \eta \|\nabla f(x)\|^2 + \big(\text{terms that vanish faster than } \eta\big).$$

While the gradient is large, a small enough step downhill still helps.

What can we hope to find?

We would like a minimizer of f. But for neural networks f can have many valleys and flat regions, and no efficient method is known that finds the lowest point of every such function.



A weaker target: a point with a small gradient. For a step $h = -\eta \nabla f(x)$ with step size $\eta > 0$,

$$f(x+h) = f(x) - \eta \|\nabla f(x)\|^2 + \big(\text{terms that vanish faster than } \eta\big).$$

While the gradient is large, a small enough step downhill still helps.

ε-stationary point

A point x with $\|\nabla f(x)\| \le \varepsilon$. This does not promise low loss: in the picture, all three marked points have $\nabla f = 0$.

How large a step is safe?

The first-order picture only holds for small steps. To say *how* small, assume:

L-smoothness

$$\|\nabla f(x) - \nabla f(x')\| \le L\|x - x'\| \qquad \text{for all } x, x'.$$

How large a step is safe?

The first-order picture only holds for small steps. To say *how* small, assume:

L-smoothness

$$\|\nabla f(x) - \nabla f(x')\| \le L\|x - x'\| \qquad \text{for all } x, x'.$$

The gradient cannot change too fast, and this gives a quadratic upper bound:[1] for any step h,

$$f(x+h) \le f(x) + \langle \nabla f(x), h \rangle + \frac{L}{2}\|h\|^2.$$

[1] Proof: Appendix A8.

How large a step is safe?

The first-order picture only holds for small steps. To say *how* small, assume:

L-smoothness

$$\|\nabla f(x) - \nabla f(x')\| \le L\|x - x'\| \qquad \text{for all } x, x'.$$

The gradient cannot change too fast, and this gives a quadratic upper bound:[1] for any step h,

$$f(x+h) \le f(x) + \langle \nabla f(x), h\rangle + \frac{L}{2}\|h\|^2.$$

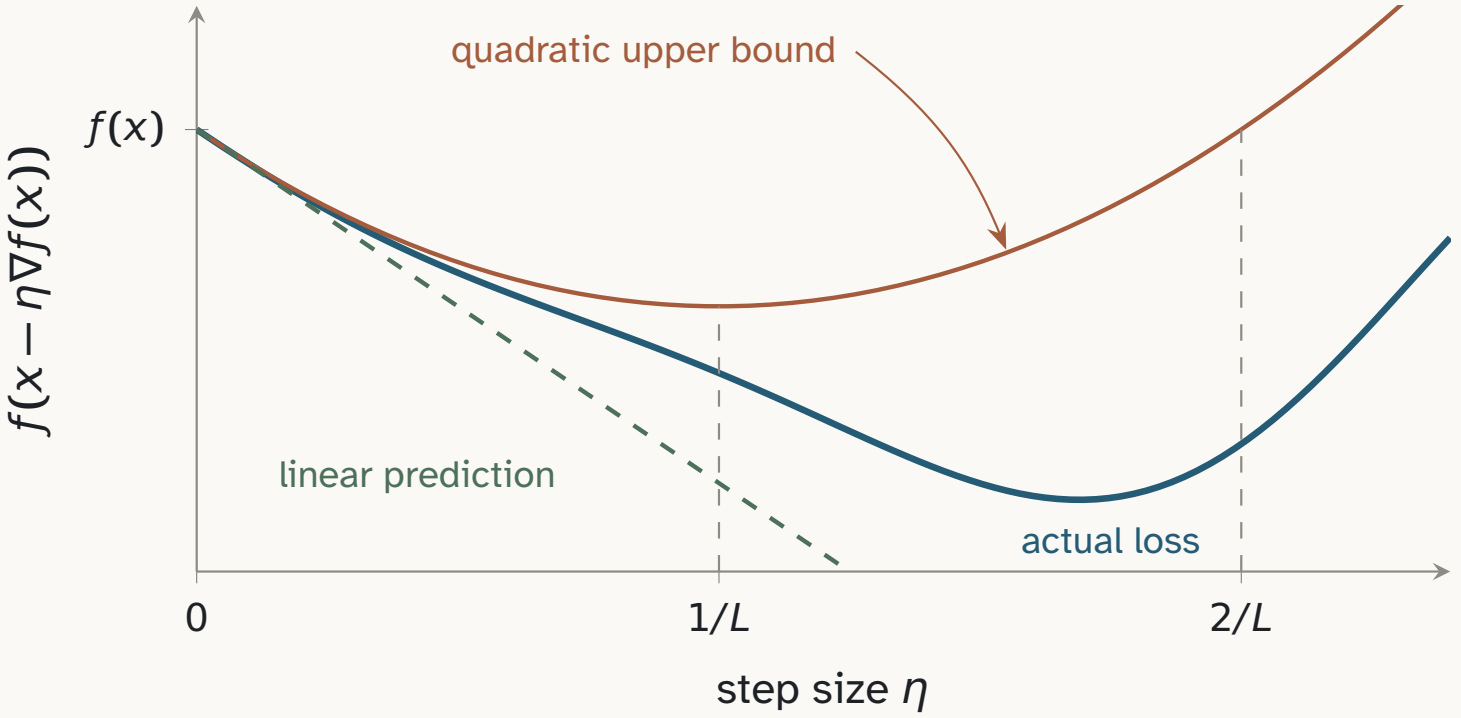
For the downhill step $h = -\eta \nabla f(x)$,

$$f(x+h) \le f(x) - \eta\left(1 - \tfrac{L\eta}{2}\right)\|\nabla f(x)\|^2,$$

a guaranteed decrease whenever $0 < \eta < 2/L$ and $\nabla f(x) \neq 0$.

[1] Proof: Appendix A8.

The upper bound picks the step size



The actual loss stays below the quadratic bound. The bound guarantees a decrease for any $\eta < 2/L$ and promises the most at $\eta = 1/L$, where it guarantees a drop of $\|\nabla f(x)\|^2/(2L)$.

Gradient descent: a decrease at every step

Gradient descent uses the step size the bound suggests:

$$x_{t+1} = x_t - \frac{1}{L}\nabla f(x_t).$$

Gradient descent: a decrease at every step

Gradient descent uses the step size the bound suggests:

$$x_{t+1} = x_t - \frac{1}{L}\nabla f(x_t).$$

Substitute $\eta = 1/L$ into the smoothness bound:

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|^2.$$

Gradient descent: a decrease at every step

Gradient descent uses the step size the bound suggests:

$$x_{t+1} = x_t - \frac{1}{L}\nabla f(x_t).$$

Substitute $\eta = 1/L$ into the smoothness bound:

$$f(x_{t+1}) \le f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|^2.$$

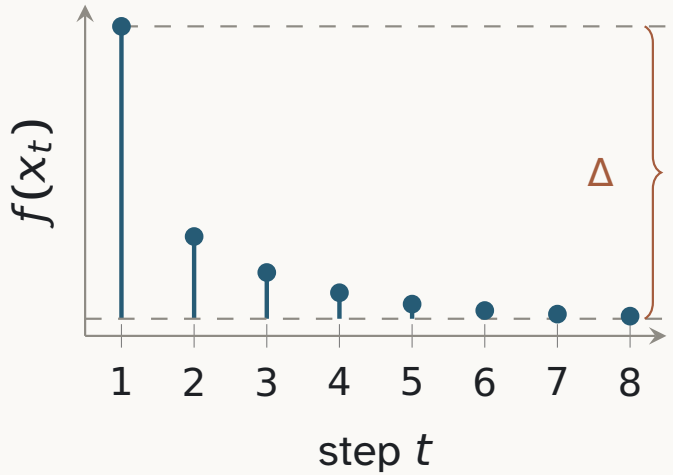
Suppose f is bounded below by f_{inf}, and let $\Delta = f(x_1) - f_{\text{inf}}$ be the total decrease available.

Every step with $\|\nabla f(x_t)\| > \varepsilon$ uses up more than $\varepsilon^2/2L$ of Δ. How many such steps can there be?

Large gradients cannot last

Add the per-step decreases:

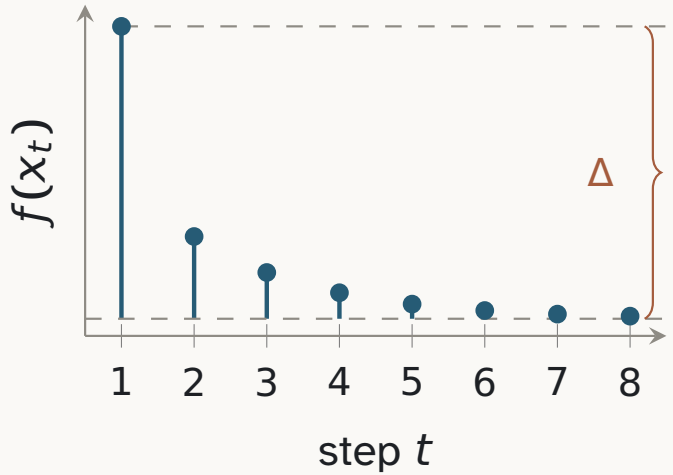
$$\frac{1}{2L}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2 \le f(x_1) - f(x_{T+1}) \le \Delta.$$



Large gradients cannot last

Add the per-step decreases:

$$\frac{1}{2L}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2 \le f(x_1) - f(x_{T+1}) \le \Delta.$$



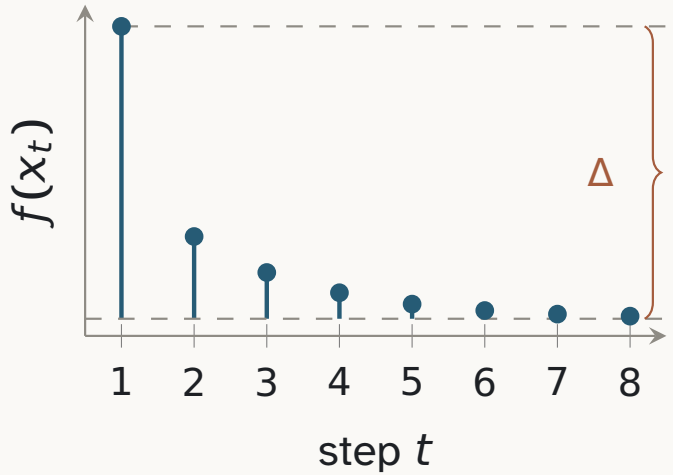
Gradient descent

$$\min_{t\le T}\|\nabla f(x_t)\|^2 \le \frac{2L\Delta}{T}.$$

Large gradients cannot last

Add the per-step decreases:

$$\frac{1}{2L}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2 \le f(x_1) - f(x_{T+1}) \le \Delta.$$



Gradient descent

$$\min_{t\le T}\|\nabla f(x_t)\|^2 \le \frac{2L\Delta}{T}.$$

To reach $\|\nabla f(x_t)\| \le \varepsilon$ takes $2L\Delta/\varepsilon^2$ steps, each reading all n examples: $2nL\Delta/\varepsilon^2$ **example gradients in total**.

Stochastic gradient descent

Gradient descent reads the whole corpus at every step. Instead, read one random example:

Stochastic gradient descent (SGD)

Start at x_1 with a step size $\eta > 0$. For $t = 1, \ldots, T$:

1. Draw I_t uniformly from $\{1, \ldots, n\}$, independently of the past.[1]
2. Compute $g_t = \nabla f_{I_t}(x_t)$.
3. Update $x_{t+1} = x_t - \eta g_t$.

[1] Why not one shuffled pass: Appendix A9.

Stochastic gradient descent

Gradient descent reads the whole corpus at every step. Instead, read one random example:

Stochastic gradient descent (SGD)

Start at x_1 with a step size $\eta > 0$. For $t = 1, \ldots, T$:

1. Draw I_t uniformly from $\{1, \ldots, n\}$, independently of the past.[1]
2. Compute $g_t = \nabla f_{I_t}(x_t)$.
3. Update $x_{t+1} = x_t - \eta g_t$.

Each step costs one example gradient instead of n.[2] The randomness picks a training pair from the corpus; the model is not generating text.

[1] Why not one shuffled pass: Appendix A9.

[2] Averaging several examples per step: Appendix A16.

The sampled gradient is right on average

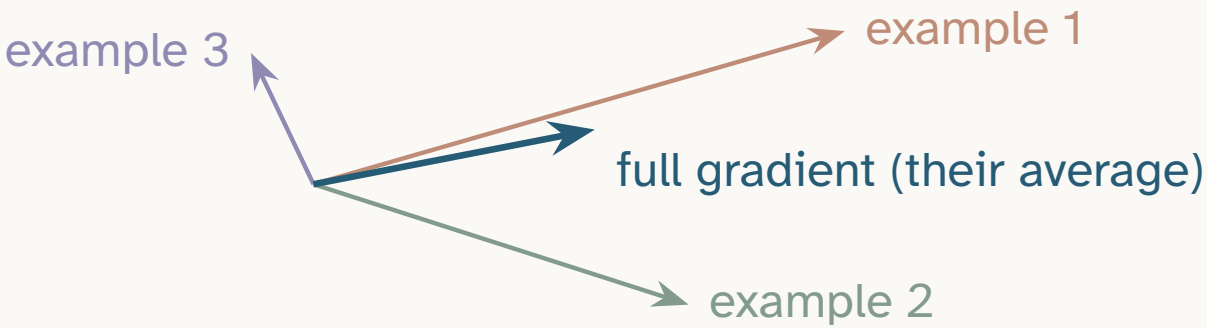
Now let $\mathcal{F}_{t-1}$ be the history of the first $t-1$ draws $I_1, \ldots, I_{t-1}$. Given $\mathcal{F}_{t-1}$, the point x_t is fixed and I_t is still uniform, so

$$\mathbb{E}[\, g_t \mid \mathcal{F}_{t-1}] = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_t) = \nabla f(x_t).$$

The sampled gradient is right on average

Now let $\mathcal{F}_{t-1}$ be the history of the first $t-1$ draws $I_1, \ldots, I_{t-1}$. Given $\mathcal{F}_{t-1}$, the point x_t is fixed and I_t is still uniform, so

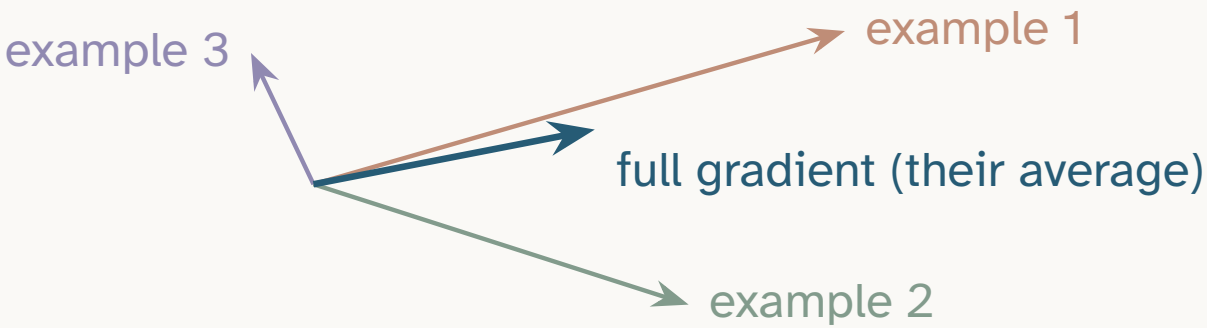
$$\mathbb{E}[\, g_t \mid \mathcal{F}_{t-1}] = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_t) = \nabla f(x_t).$$



The sampled gradient is right on average

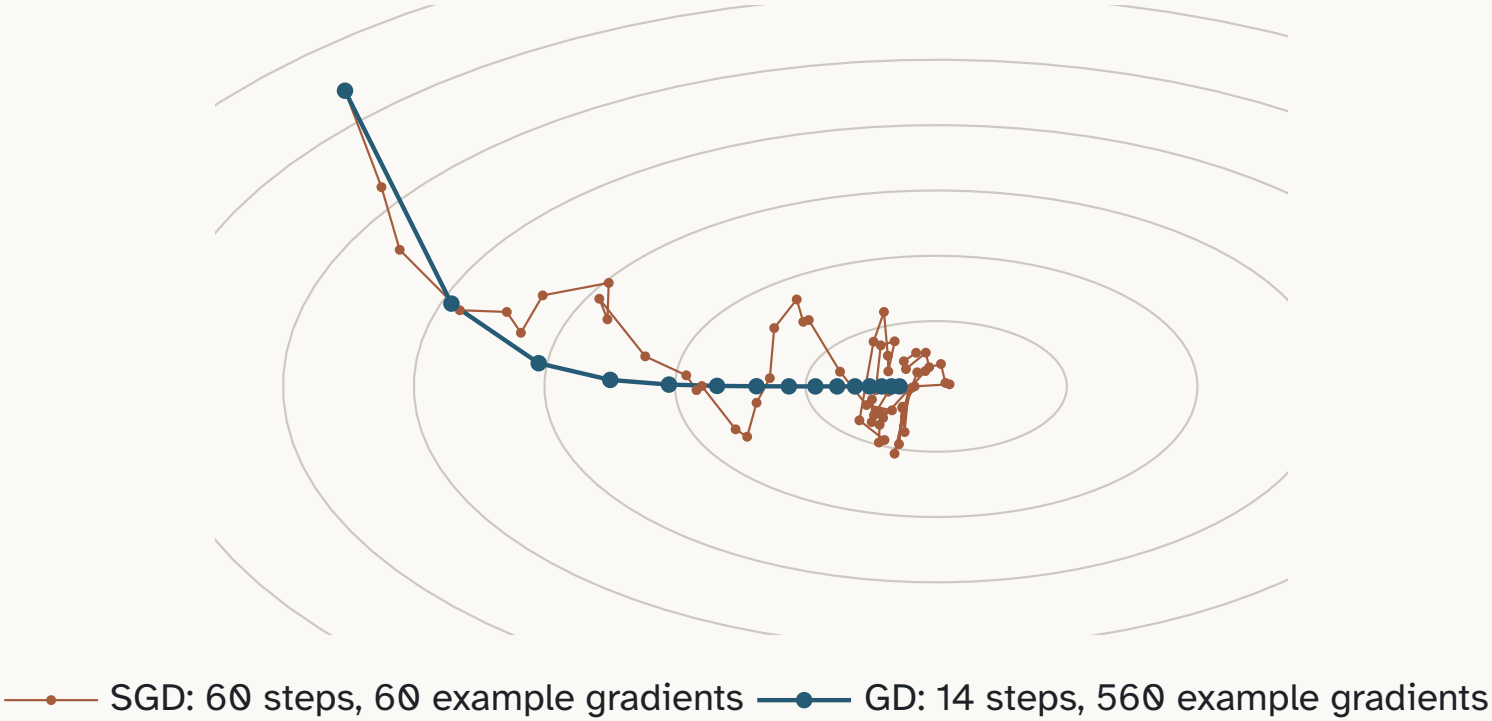
Now let $\mathcal{F}_{t-1}$ be the history of the first $t-1$ draws $I_1, \ldots, I_{t-1}$. Given $\mathcal{F}_{t-1}$, the point x_t is fixed and I_t is still uniform, so

$$\mathbb{E}[\, g_t \mid \mathcal{F}_{t-1}] = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_t) = \nabla f(x_t).$$



The gradients at different steps are *not* independent: earlier draws moved x_t. Only the fresh draw I_t is independent of the past.

What the paths look like



Least squares in two dimensions with $n = 40$ examples, contours of f in grey. SGD's path is noisy, but it heads downhill on average and reads far less data.

The assumptions for the analysis

Assumption	What it provides
$f(x) \geq f_{\text{inf}}$	At most $\Delta = f(x_1) - f_{\text{inf}}$ total decrease.
L-smoothness	Controls the error of a finite step.
$\mathbb{E}[\, g_t \mid \mathcal{F}_{t-1}] = \nabla f(x_t)$	The sampling error has mean zero.
$\|g_t\| \leq G$ always	Bounds each step and the spread of the error.

The assumptions for the analysis

Assumption	What it provides
$f(x) \geq f_{\text{inf}}$	At most $\Delta = f(x_1) - f_{\text{inf}}$ total decrease.
L-smoothness	Controls the error of a finite step.
$\mathbb{E}[g_t \mid \mathcal{F}_{t-1}] = \nabla f(x_t)$	The sampling error has mean zero.
$\|g_t\| \leq G$ always	Bounds each step and the spread of the error.

The log loss is nonnegative, so $f_{\text{inf}} = 0$ works, and uniform sampling gives unbiasedness. The constants L and G depend on the model,[1] and clipping gradients to force a bound G can change their mean.[2] Here is a model where they hold.

[1] A model where they fail: Appendix A10.
[2] Example: Appendix A11.

A model where the assumptions hold

Use fixed feature vectors $\phi(c, v)$ for each context c and token v, with $\|\phi(c, v)\| \le R$:

$$P_x(v \mid c) = \frac{e^{\langle x, \phi(c,v)\rangle}}{\sum_u e^{\langle x, \phi(c,u)\rangle}}.$$

A model where the assumptions hold

Use fixed feature vectors $\phi(c, v)$ for each context c and token v, with $\|\phi(c, v)\| \le R$:

$$P_x(v \mid c) = \frac{e^{\langle x, \phi(c,v)\rangle}}{\sum_u e^{\langle x, \phi(c,u)\rangle}}.$$

Property	Constant
$\|\nabla f_i(x)\| \le 2R$	$G = 2R$
∇f is R^2-Lipschitz	$L = R^2$
$f(x) \ge 0$	$f_{\inf} = 0$

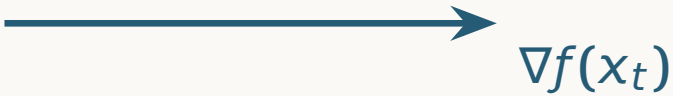
These bounds hold at every parameter vector x.[1]

[1] Proofs: Appendix A12.

Progress and noise in one step

Smoothness with $h = -\eta g_t$:

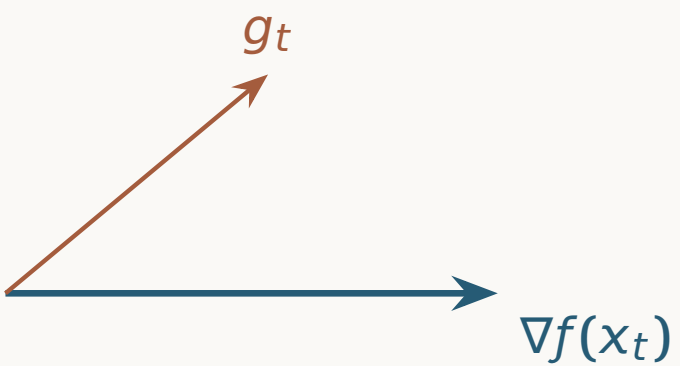
$$f(x_{t+1}) \le f(x_t) - \eta \langle \nabla f(x_t), g_t \rangle + \frac{L\eta^2}{2} \|g_t\|^2.$$



Progress and noise in one step

Smoothness with $h = -\eta g_t$:

$$f(x_{t+1}) \le f(x_t) - \eta \langle \nabla f(x_t), g_t \rangle + \frac{L\eta^2}{2} \|g_t\|^2.$$



Write $g_t = \nabla f(x_t) + e_t$. Then

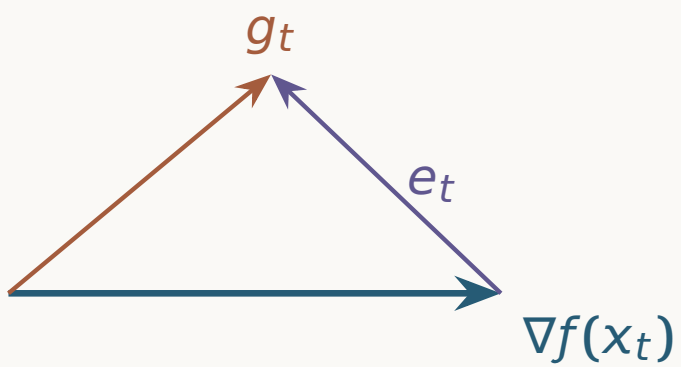
$$-\langle \nabla f(x_t), g_t \rangle = -\|\nabla f(x_t)\|^2 + D_t,$$

with $D_t = -\langle \nabla f(x_t), e_t \rangle$.

Progress and noise in one step

Smoothness with $h = -\eta g_t$:

$$f(x_{t+1}) \le f(x_t) - \eta \langle \nabla f(x_t), g_t \rangle + \frac{L\eta^2}{2} \|g_t\|^2.$$



Write $g_t = \nabla f(x_t) + e_t$. Then

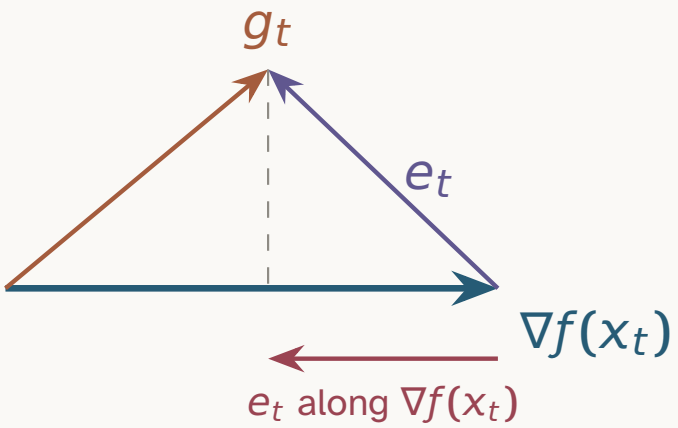
$$-\langle \nabla f(x_t), g_t \rangle = -\|\nabla f(x_t)\|^2 + D_t,$$

with $D_t = -\langle \nabla f(x_t), e_t \rangle$.

Progress and noise in one step

Smoothness with $h = -\eta g_t$:

$$f(x_{t+1}) \le f(x_t) - \eta \langle \nabla f(x_t), g_t \rangle + \frac{L\eta^2}{2} \|g_t\|^2.$$



Write $g_t = \nabla f(x_t) + e_t$. Then

$$-\langle \nabla f(x_t), g_t \rangle = -\|\nabla f(x_t)\|^2 + D_t,$$

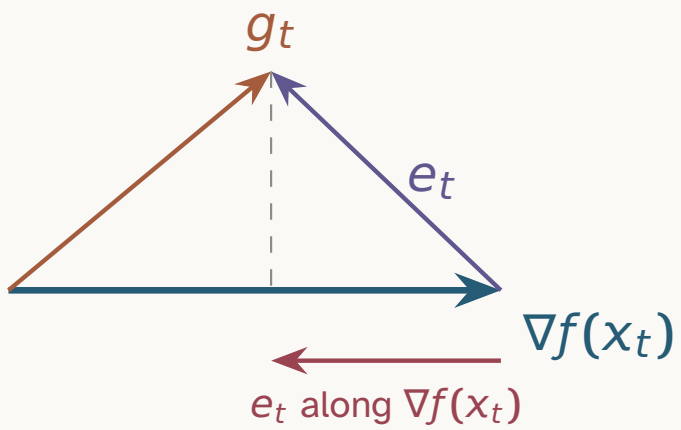
with $D_t = -\langle \nabla f(x_t), e_t \rangle$.

$D_t > 0$: this step makes less first-order progress than the full gradient would (as pictured). $D_t < 0$: more.

Progress and noise in one step

Smoothness with $h = -\eta g_t$:

$$f(x_{t+1}) \le f(x_t) - \eta \langle \nabla f(x_t), g_t \rangle + \frac{L\eta^2}{2} \|g_t\|^2.$$



Write $g_t = \nabla f(x_t) + e_t$. Then

$$-\langle \nabla f(x_t), g_t \rangle = -\|\nabla f(x_t)\|^2 + D_t,$$

with $D_t = -\langle \nabla f(x_t), e_t \rangle$.

$D_t > 0$: this step makes less first-order progress than the full gradient would (as pictured). $D_t < 0$: more.

The full gradient is fixed given $\mathcal{F}_{t-1}$, so pull it out; since $\mathbb{E}[\, e_t \mid \mathcal{F}_{t-1}] = 0$,

$$\mathbb{E}[\, D_t \mid \mathcal{F}_{t-1}] = -\langle \nabla f(x_t), \mathbb{E}[\, e_t \mid \mathcal{F}_{t-1}] \rangle = 0.$$

As in Part I, D_t is a mean-zero term created by our random choice.

One inequality for the whole run

Use $\|g_t\| \le G$ in the one-step bound:

$$\eta \|\nabla f(x_t)\|^2 \le f(x_t) - f(x_{t+1}) + \frac{L\eta^2 G^2}{2} + \eta D_t.$$

One inequality for the whole run

Use $\|g_t\| \le G$ in the one-step bound:

$$\eta \|\nabla f(x_t)\|^2 \le f(x_t) - f(x_{t+1}) + \frac{L\eta^2 G^2}{2} + \eta D_t.$$

Sum over t, telescope, and use $f(x_{T+1}) \ge f_{\inf}$:

$$\frac{1}{T}\sum_t \|\nabla f(x_t)\|^2 \le \frac{\Delta}{\eta T} + \frac{L\eta G^2}{2} + \frac{1}{T}\sum_t D_t.$$

One inequality for the whole run

Use $\|g_t\| \le G$ in the one-step bound:

$$\eta \|\nabla f(x_t)\|^2 \le f(x_t) - f(x_{t+1}) + \frac{L\eta^2 G^2}{2} + \eta D_t.$$

Sum over t, telescope, and use $f(x_{T+1}) \ge f_{\inf}$:

$$\frac{1}{T}\sum_t \|\nabla f(x_t)\|^2 \le \frac{\Delta}{\eta T} + \frac{L\eta G^2}{2} + \frac{1}{T}\sum_t D_t.$$

Every term but the last is fixed in advance. In expectation the last term is zero. In a single run it is random, and that is the subject of Part III.

SGD in expectation

Theorem 2 (SGD in expectation)[1]

Under the four assumptions, with $\eta = \sqrt{2\Delta/(LG^2T)}$,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(x_t)\|^2 \leq \sqrt{\frac{2\Delta LG^2}{T}}.$$

[1] A version using the variance of g_t instead of G: Appendix A14.

SGD in expectation

Theorem 2 (SGD in expectation)[1]

Under the four assumptions, with $\eta = \sqrt{2\Delta/(LG^2T)}$,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(x_t)\|^2 \le \sqrt{\frac{2\Delta LG^2}{T}}.$$

Proof. By the tower rule, $\mathbb{E}D_t = \mathbb{E}\big[\mathbb{E}(D_t \mid \mathcal{F}_{t-1})\big] = 0$. Take expectations in the inequality for the whole run:

$$\frac{1}{T}\sum_{t}\mathbb{E}\|\nabla f(x_t)\|^2 \le \frac{\Delta}{\eta T} + \frac{L\eta G^2}{2}.$$

The given η makes both terms equal to $\sqrt{\Delta LG^2/(2T)}$. □

[1] A version using the variance of g_t instead of G: Appendix A14.

Which point should SGD return?

Checking every $\|\nabla f(x_t)\|$ would cost n example gradients per step.
Instead, before training, draw U uniformly from $\{1, \ldots, T\}$ and save x_U.

Which point should SGD return?

Checking every $\|\nabla f(x_t)\|$ would cost n example gradients per step. Instead, before training, draw U uniformly from $\{1, \ldots, T\}$ and save x_U.

$$\mathbb{E}\|\nabla f(x_U)\|^2 = \frac{1}{T}\sum_t \mathbb{E}\|\nabla f(x_t)\|^2 \le \sqrt{\frac{2\Delta L G^2}{T}}.$$

Which point should SGD return?

Checking every $\|\nabla f(x_t)\|$ would cost n example gradients per step. Instead, before training, draw U uniformly from $\{1, \dots, T\}$ and save x_U.

$$\mathbb{E}\|\nabla f(x_U)\|^2 = \frac{1}{T}\sum_t \mathbb{E}\|\nabla f(x_t)\|^2 \le \sqrt{\frac{2\Delta L G^2}{T}}.$$

So

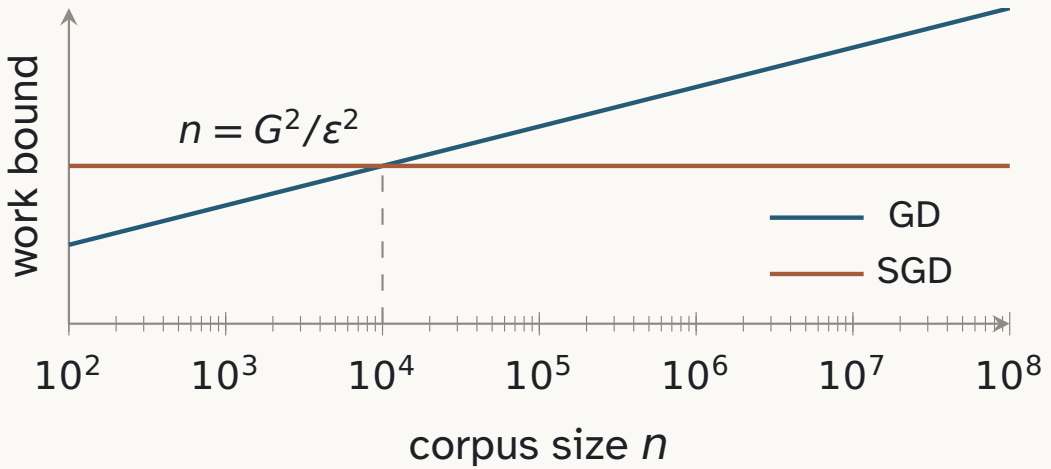
$$T \ge \frac{2\Delta L G^2}{\varepsilon^4} \quad \Longrightarrow \quad \mathbb{E}\|\nabla f(x_U)\|^2 \le \varepsilon^2.$$

More steps, less work

	GD	SGD
Example gradients per step	n	1
Steps sufficient for ε	$2L\Delta/\varepsilon^2$	$2\Delta LG^2/\varepsilon^4$
Total example gradients	$2nL\Delta/\varepsilon^2$	$2\Delta LG^2/\varepsilon^4$

More steps, less work

	GD	SGD
Example gradients per step	n	1
Steps sufficient for ε	$2L\Delta/\varepsilon^2$	$2\Delta LG^2/\varepsilon^4$
Total example gradients	$2nL\Delta/\varepsilon^2$	$2\Delta LG^2/\varepsilon^4$



SGD's bound is smaller once $n > G^2/\varepsilon^2$, and does not grow with the corpus at all.

PART III

Will the run you actually do be good?

Training a large model is expensive, so it is done once. A bound on the average over many runs does not tell us how this run went.

We want a bound that holds with high probability for a single run.

Only one term is random

Recall: in every run,

$$\frac{1}{T}\sum_t \|\nabla f(x_t)\|^2 \le \frac{\Delta}{\eta T} + \frac{L\eta G^2}{2} + \frac{1}{T}\sum_t D_t.$$

The D_t have conditional mean zero, so Azuma applies once we know their widths.

Only one term is random

Recall: in every run,

$$\frac{1}{T}\sum_t \|\nabla f(x_t)\|^2 \le \frac{\Delta}{\eta T} + \frac{L\eta G^2}{2} + \frac{1}{T}\sum_t D_t.$$

The D_t have conditional mean zero, so Azuma applies once we know their widths.

Given $\mathcal{F}_{t-1}$, Cauchy–Schwarz bounds $-\langle \nabla f(x_t), g_t\rangle$ in an interval of width $2G\|\nabla f(x_t)\|$. Since $D_t = \|\nabla f(x_t)\|^2 - \langle \nabla f(x_t), g_t\rangle$ only shifts it,

$$c_t = 2G\|\nabla f(x_t)\|.$$

Only one term is random

Recall: in every run,

$$\frac{1}{T}\sum_t \|\nabla f(x_t)\|^2 \le \frac{\Delta}{\eta T} + \frac{L\eta G^2}{2} + \frac{1}{T}\sum_t D_t.$$

The D_t have conditional mean zero, so Azuma applies once we know their widths.

Given $\mathcal{F}_{t-1}$, Cauchy–Schwarz bounds $-\langle \nabla f(x_t), g_t\rangle$ in an interval of width $2G\|\nabla f(x_t)\|$. Since $D_t = \|\nabla f(x_t)\|^2 - \langle \nabla f(x_t), g_t\rangle$ only shifts it,

$$c_t = 2G\|\nabla f(x_t)\|.$$

Since g_t can be any of the n example gradients at x_t, each has norm at most G, and so does their average $\nabla f(x_t)$ (triangle inequality). So $c_t \le 2G^2$. Start with this worst case.

A first one-run bound

Azuma with widths $2G^2$: with probability at least $1 - \delta$,

$$\frac{1}{T}\sum_t D_t \le G^2\sqrt{\frac{2\log(1/\delta)}{T}}.$$

A first one-run bound

Azuma with widths $2G^2$: with probability at least $1-\delta$,

$$\frac{1}{T}\sum_t D_t \le G^2\sqrt{\frac{2\log(1/\delta)}{T}}.$$

First one-run bound

With probability at least $1-\delta$,

$$\frac{1}{T}\sum_t \|\nabla f(x_t)\|^2 \le \frac{\Delta}{\eta T} + \frac{L\eta G^2}{2} + G^2\sqrt{\frac{2\log(1/\delta)}{T}}.$$

A first one-run bound

Azuma with widths $2G^2$: with probability at least $1-\delta$,

$$\frac{1}{T}\sum_t D_t \le G^2\sqrt{\frac{2\log(1/\delta)}{T}}.$$

First one-run bound

With probability at least $1-\delta$,

$$\frac{1}{T}\sum_t \|\nabla f(x_t)\|^2 \le \frac{\Delta}{\eta T} + \frac{L\eta G^2}{2} + G^2\sqrt{\frac{2\log(1/\delta)}{T}}.$$

If $\nabla f(x_t) = 0$, what is D_t?

A first one-run bound

Azuma with widths $2G^2$: with probability at least $1-\delta$,

$$\frac{1}{T}\sum_t D_t \le G^2\sqrt{\frac{2\log(1/\delta)}{T}}.$$

First one-run bound

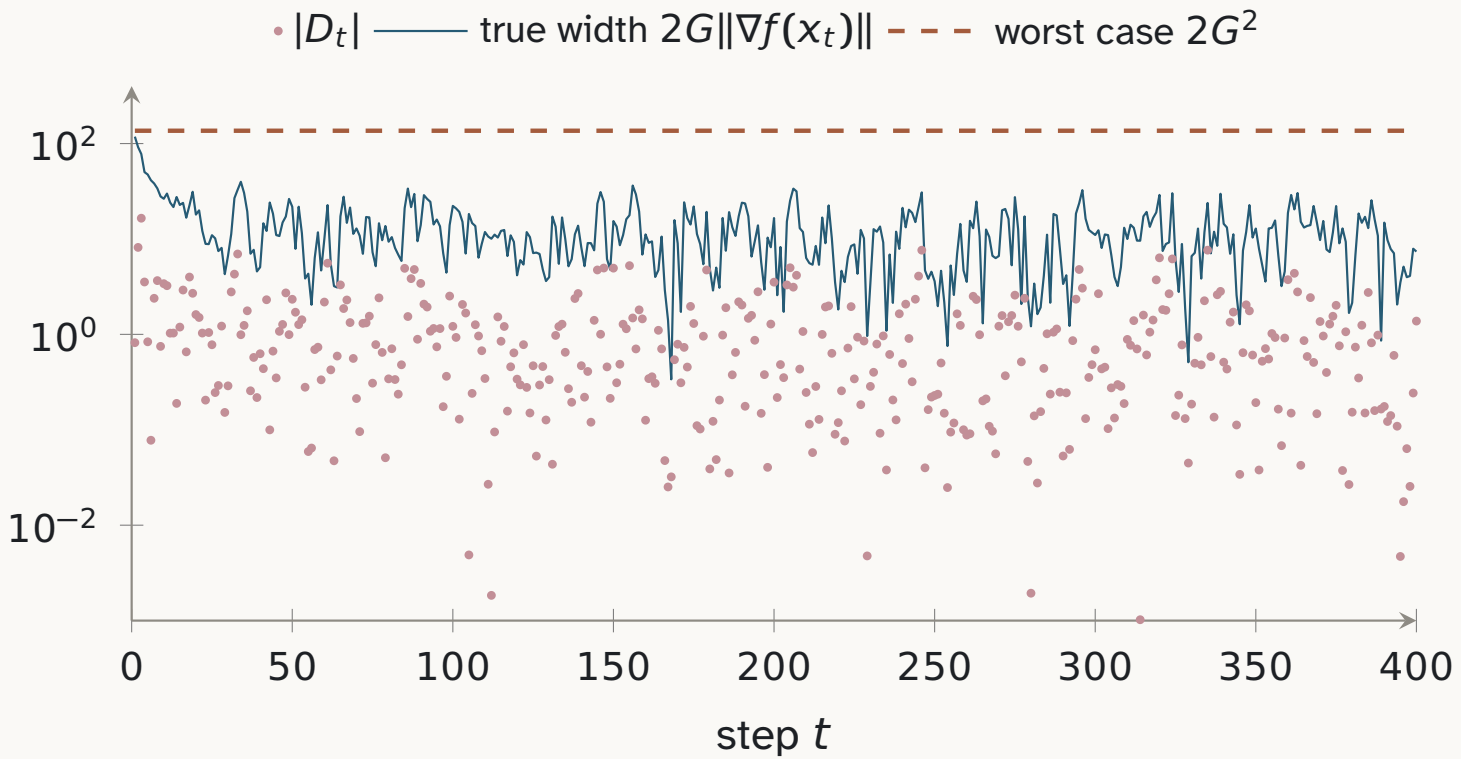
With probability at least $1-\delta$,

$$\frac{1}{T}\sum_t \|\nabla f(x_t)\|^2 \le \frac{\Delta}{\eta T} + \frac{L\eta G^2}{2} + G^2\sqrt{\frac{2\log(1/\delta)}{T}}.$$

If $\nabla f(x_t) = 0$, what is D_t?

$D_t = -\langle \nabla f(x_t), e_t\rangle = 0$. Near a stationary point the true width $2G\|\nabla f(x_t)\|$ is far below $2G^2$, and the bound ignores this.

The noise term shrinks with the gradient



One SGD run on the least-squares example. After the first few steps the true width is usually 5 to 40 times smaller than the worst case that the first bound charges at every step.

What a sharper bound must look like

Write $M_T = \sum_t D_t$ and $S_T = \sum_t \|\nabla f(x_t)\|^2$. The inequality for the whole run reads

$$S_T \le \frac{\Delta}{\eta} + \frac{L\eta G^2 T}{2} + M_T.$$

What a sharper bound must look like

Write $M_T = \sum_t D_t$ and $S_T = \sum_t \|\nabla f(x_t)\|^2$. The inequality for the whole run reads

$$S_T \le \frac{\Delta}{\eta} + \frac{L\eta G^2 T}{2} + M_T.$$

Goal

With probability at least $1 - \delta$, $\quad M_T \le \frac{1}{2} S_T + G^2 \log(1/\delta)$.

If the noise is at most half the progress, we can move it to the left side.

What a sharper bound must look like

Write $M_T = \sum_t D_t$ and $S_T = \sum_t \|\nabla f(x_t)\|^2$. The inequality for the whole run reads

$$S_T \le \frac{\Delta}{\eta} + \frac{L\eta G^2 T}{2} + M_T.$$

Goal

With probability at least $1-\delta$, $\quad M_T \le \frac{1}{2} S_T + G^2 \log(1/\delta)$.

If the noise is at most half the progress, we can move it to the left side.

Azuma cannot give this: it needs widths fixed in advance, and the true widths $2G\|\nabla f(x_t)\|$ depend on the run. So we prove the goal directly, in three steps: bound one step's exponential average with Hoeffding's lemma, multiply over the steps, and apply Markov's inequality.

Step 1: Hoeffding's lemma for one step

Hoeffding's lemma, applied given $\mathcal{F}_{t-1}$ with width $2G\|\nabla f(x_t)\|$:

$$\mathbb{E}\big[e^{\lambda D_t} \,\big|\, \mathcal{F}_{t-1}\big] \le \exp\Big(\tfrac{\lambda^2 G^2}{2}\|\nabla f(x_t)\|^2\Big).$$

Step 1: Hoeffding's lemma for one step

Hoeffding's lemma, applied given $\mathcal{F}_{t-1}$ with width $2G\|\nabla f(x_t)\|$:

$$\mathbb{E}\Big[e^{\lambda D_t} \,\Big|\, \mathcal{F}_{t-1}\Big] \le \exp\Big(\tfrac{\lambda^2 G^2}{2}\|\nabla f(x_t)\|^2\Big).$$

The right side is fixed given $\mathcal{F}_{t-1}$, so it can move inside:

$$\mathbb{E}\Big[\exp\Big(\lambda D_t - \tfrac{\lambda^2 G^2}{2}\|\nabla f(x_t)\|^2\Big) \,\Big|\, \mathcal{F}_{t-1}\Big] \le 1.$$

Step 2: a supermartingale

Multiply the per-step factors:

$$Z_t = \exp\Big(\lambda M_t - \frac{\lambda^2 G^2}{2} S_t\Big), \qquad Z_0 = 1,$$

where M_t and S_t are the sums up to step t.

Step 2: a supermartingale

Multiply the per-step factors:

$$Z_t = \exp\Big(\lambda M_t - \frac{\lambda^2 G^2}{2} S_t\Big), \qquad Z_0 = 1,$$

where M_t and S_t are the sums up to step t. Z_{t-1} is fixed given $\mathcal{F}_{t-1}$; pull it out and use step 1:

$$\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = Z_{t-1}\,\mathbb{E}\Big[\exp\Big(\lambda D_t - \tfrac{\lambda^2 G^2}{2}\|\nabla f(x_t)\|^2\Big)\,\Big|\,\mathcal{F}_{t-1}\Big] \le Z_{t-1}.$$

Step 2: a supermartingale

Multiply the per-step factors:

$$Z_t = \exp\Big(\lambda M_t - \frac{\lambda^2 G^2}{2} S_t\Big), \qquad Z_0 = 1,$$

where M_t and S_t are the sums up to step t. Z_{t-1} is fixed given $\mathcal{F}_{t-1}$; pull it out and use step 1:

$$\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = Z_{t-1}\,\mathbb{E}\Big[\exp\Big(\lambda D_t - \tfrac{\lambda^2 G^2}{2}\|\nabla f(x_t)\|^2\Big)\,\Big|\,\mathcal{F}_{t-1}\Big] \le Z_{t-1}.$$

Supermartingale

A sequence $Z_0, Z_1, \ldots$, with each Z_t fixed given $\mathcal{F}_t$, such that

$$\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] \le Z_{t-1} \qquad \text{for every } t.$$

On average it never goes up.

By the tower rule, $\mathbb{E}Z_t \le \mathbb{E}Z_{t-1}$ for every t, so $\mathbb{E}Z_T \le Z_0 = 1$.

Step 3: Markov, and a convenient λ

Markov: $\mathbb{P}(Z_T \geq 1/\delta) \leq \delta$. Taking logs, with probability at least $1 - \delta$,

$$M_T \leq \frac{\lambda G^2}{2} S_T + \frac{\log(1/\delta)}{\lambda}.$$

Step 3: Markov, and a convenient λ

Markov: $\mathbb{P}(Z_T \geq 1/\delta) \leq \delta$. Taking logs, with probability at least $1 - \delta$,

$$M_T \leq \frac{\lambda G^2}{2} S_T + \frac{\log(1/\delta)}{\lambda}.$$

We want the coefficient of S_T to be $\frac{1}{2}$. Which λ?

Step 3: Markov, and a convenient λ

Markov: $\mathbb{P}(Z_T \geq 1/\delta) \leq \delta$. Taking logs, with probability at least $1 - \delta$,

$$M_T \leq \frac{\lambda G^2}{2} S_T + \frac{\log(1/\delta)}{\lambda}.$$

We want the coefficient of S_T to be $\frac{1}{2}$. Which λ?

$\lambda = 1/G^2$ gives

$$M_T \leq \tfrac{1}{2} S_T + G^2 \log(1/\delta),$$

which is the goal. □

With fixed widths c_t in place of $2G\|\nabla f(x_t)\|$, the same three steps prove Azuma's inequality.[1]

[1] Details: Appendix A6.

SGD in a single run

Theorem 3 (SGD, one run)

Under the four assumptions, with probability at least $1 - \delta$,

$$\frac{1}{T}\sum_{t}\|\nabla f(x_t)\|^2 \le \frac{2\Delta}{\eta T} + L\eta G^2 + \frac{2G^2\log(1/\delta)}{T}.$$

SGD in a single run

Theorem 3 (SGD, one run)

Under the four assumptions, with probability at least $1 - \delta$,

$$\frac{1}{T}\sum_{t}\|\nabla f(x_t)\|^2 \le \frac{2\Delta}{\eta T} + L\eta G^2 + \frac{2G^2\log(1/\delta)}{T}.$$

Proof. Put the goal into $S_T \le \Delta/\eta + L\eta G^2 T/2 + M_T$, move $\frac{1}{2}S_T$ to the left, and multiply by $2/T$. □

SGD in a single run

Theorem 3 (SGD, one run)

Under the four assumptions, with probability at least $1 - \delta$,

$$\frac{1}{T}\sum_t \|\nabla f(x_t)\|^2 \le \frac{2\Delta}{\eta T} + L\eta G^2 + \frac{2G^2 \log(1/\delta)}{T}.$$

Proof. Put the goal into $S_T \le \Delta/\eta + L\eta G^2 T/2 + M_T$, move $\frac{1}{2}S_T$ to the left, and multiply by $2/T$. □

With $\eta = \sqrt{2\Delta/(LG^2T)}$:

$$\frac{1}{T}\sum_t \|\nabla f(x_t)\|^2 \le 2\sqrt{\frac{2\Delta L G^2}{T}} + \frac{2G^2 \log(1/\delta)}{T}.$$

So this run contains a point with small gradient. Finding which one takes ~~extra work.~~[1]

[1] How: Appendix A17.

What the sharper bound buys

Steps sufficient for an average squared gradient of at most ε^2, with probability $1 - \delta$:

	Optimization term	Confidence term
First bound	$8\Delta L G^2/\varepsilon^4$	$8G^4 \log(1/\delta)/\varepsilon^4$
Sharper bound	$32\Delta L G^2/\varepsilon^4$	$4G^2 \log(1/\delta)/\varepsilon^2$

What the sharper bound buys

Steps sufficient for an average squared gradient of at most ε^2, with probability $1 - \delta$:

	Optimization term	Confidence term
First bound	$8\Delta L G^2/\varepsilon^4$	$8G^4 \log(1/\delta)/\varepsilon^4$
Sharper bound	$32\Delta L G^2/\varepsilon^4$	$4G^2 \log(1/\delta)/\varepsilon^2$

For example, $\Delta L = G^2 = 1$ and $\varepsilon = 0.01$:

δ	First bound	Sharper bound
10^{-2}	3.7×10^9	3.2×10^9
10^{-6}	1.1×10^{10}	3.2×10^9
10^{-12}	2.2×10^{10}	3.2×10^9

What the sharper bound buys

Steps sufficient for an average squared gradient of at most ε^2, with probability $1 - \delta$:

	Optimization term	Confidence term
First bound	$8\Delta L G^2/\varepsilon^4$	$8G^4 \log(1/\delta)/\varepsilon^4$
Sharper bound	$32\Delta L G^2/\varepsilon^4$	$4G^2 \log(1/\delta)/\varepsilon^2$

For example, $\Delta L = G^2 = 1$ and $\varepsilon = 0.01$:

δ	First bound	Sharper bound
10^{-2}	3.7×10^9	3.2×10^9
10^{-6}	1.1×10^{10}	3.2×10^9
10^{-12}	2.2×10^{10}	3.2×10^9

With the sharper bound, high confidence is nearly free: its cost scales like $\log(1/\delta)/\varepsilon^2$ instead of $\log(1/\delta)/\varepsilon^4$.

Where the martingales were

Both analyses had the same shape:

$$\text{performance} \;\le\; \text{terms controlled in every run} \;+\; \sum_t D_t,$$

where $\mathbb{E}[\,D_t \mid \mathcal{F}_{t-1}\,] = 0$.

Where the martingales were

Both analyses had the same shape:

$$\text{performance} \;\le\; \text{terms controlled in every run} \;+\; \sum_t D_t,$$

where $\mathbb{E}[\,D_t \mid \mathcal{F}_{t-1}\,] = 0$.

	Exponential weights	SGD
We tracked	total weight W_t	training loss $f(x_t)$
D_t is	cost paid minus average cost	shortfall in first-order progress
Width of D_t	1	$2G\|\nabla f(x_t)\|$

Where the martingales were

Both analyses had the same shape:

$$\text{performance} \;\le\; \text{terms controlled in every run} \;+\; \sum_t D_t,$$

where $\mathbb{E}[\,D_t \mid \mathcal{F}_{t-1}\,] = 0$.

	Exponential weights	SGD
We tracked	total weight W_t	training loss $f(x_t)$
D_t is	cost paid minus average cost	shortfall in first-order progress
Width of D_t	1	$2G\|\nabla f(x_t)\|$

Azuma turned the expected guarantee into a single-run guarantee for exponential weights. For SGD, keeping the true width made high confidence almost free.

Appendix

- Storing weights on a logarithmic scale
- Proof of Hoeffding’s lemma
- Exponential weights with an unknown horizon
- Proof of Azuma’s inequality
- Centered route costs can be dependent
- Proof of the smoothness inequality
- Why sampling with replacement matters
- What can fail when the model learns its features
- Clipping can change the mean
- Bounded gradients and curvature for the fixed-feature model
- A bound using the variance of the sampled gradient
- Minibatches reduce variance
- Selecting an output from the SGD run
- Reading

Storing weights on a logarithmic scale

Repeated multiplication can underflow. Store $a_{t,i} = \log w_{t,i}$:

$$a_{t+1,i} = a_{t,i} - \eta \ell_{t,i}.$$

Let $m_t = \max_i a_{t,i}$ and normalize after subtracting it:

$$\frac{e^{a_{t,i} - m_t}}{\sum_j e^{a_{t,j} - m_t}} = \frac{e^{-m_t} e^{a_{t,i}}}{e^{-m_t} \sum_j e^{a_{t,j}}} = \frac{w_{t,i}}{\sum_j w_{t,j}}.$$

The distribution is unchanged, and every exponent is now at most 0.

Used on slide 9.

Proof of Hoeffding's lemma

If $a = b$ then $X = 0$; assume $a < b$. Since $\mathbb{E}X = 0$, $a \le 0 \le b$. The curve $x \mapsto e^{\lambda x}$ bends upward, so on $[a, b]$ it lies below its chord:

$$e^{\lambda x} \le \frac{b - x}{b - a}\, e^{\lambda a} + \frac{x - a}{b - a}\, e^{\lambda b} \qquad (a \le x \le b).$$

Put $x = X$ and average, using $\mathbb{E}X = 0$:

$$\mathbb{E}e^{\lambda X} \le \frac{b\, e^{\lambda a} - a\, e^{\lambda b}}{b - a} = e^{\varphi(u)},$$

where $u = \lambda(b - a)$, $p = -a/(b - a) \in [0, 1]$, and $\varphi(u) = -pu + \log\bigl(1 - p + pe^{u}\bigr)$.

Used on slide 16.

Proof of Hoeffding's lemma, continued

Let $q(u) = pe^u/(1-p+pe^u)$, a number in $[0,1]$. Then

$$\varphi(0) = 0, \qquad \varphi'(u) = -p + q(u) \;\text{ so }\; \varphi'(0) = 0,$$

$$\varphi''(u) = q(u)\big(1-q(u)\big) \le \tfrac{1}{4}.$$

By Taylor's theorem with remainder, for some ξ between 0 and u,

$$\varphi(u) = \varphi(0) + \varphi'(0)u + \tfrac{1}{2}\varphi''(\xi)u^2 \le \frac{u^2}{8}.$$

Hence $\mathbb{E}e^{\lambda X} \le e^{\lambda^2(b-a)^2/8}$. □

Given $\mathcal{F}_{t-1}$: if a and b are fixed given $\mathcal{F}_{t-1}$, the chord inequality holds for every outcome. Averaging over the draw, the coefficients pull out, and the rest is unchanged.

Used on slide 16.

Exponential weights with an unknown horizon

Restart in phases of lengths $1, 2, 4, \ldots$, tuning η for each phase length. In phase j the average-cost regret is $O(\sqrt{2^j \log N})$, including a truncated final phase.

With $C_i^{(j)}$ route i's cost in phase j,

$$\min_i \sum_j C_i^{(j)} \geq \sum_j \min_i C_i^{(j)},$$

so regret against the best route overall is at most the sum of the phase regrets. Since

$$\sum_{j=0}^{m} 2^{j/2} \leq \frac{2^{m/2}}{1 - 2^{-1/2}},$$

the total average-cost regret is $O(\sqrt{T \log N})$. □
Apply Azuma once across all phases for a bound at a fixed time T.

Used on slide 18.

Proof of Azuma's inequality

Fix $\lambda > 0$. Hoeffding's lemma given $\mathcal{F}_{t-1}$, with endpoints A_t and $A_t + c_t$:

$$\mathbb{E}\big[e^{\lambda D_t} \,\big|\, \mathcal{F}_{t-1}\big] \le e^{\lambda^2 c_t^2/8}.$$

Let $M_t = \sum_{s \le t} D_s$ and

$$Z_t = \exp\Big(\lambda M_t - \frac{\lambda^2}{8}\sum_{s \le t} c_s^2\Big), \qquad Z_0 = 1.$$

Pulling out Z_{t-1} gives $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] \le Z_{t-1}$: a supermartingale, so $\mathbb{E}Z_T \le 1$. By Markov's inequality, for $s > 0$,

$$\mathbb{P}(M_T \ge s) = \mathbb{P}\Big(Z_T \ge e^{\lambda s - \frac{\lambda^2}{8}\sum_t c_t^2}\Big) \le e^{-\lambda s + \frac{\lambda^2}{8}\sum_t c_t^2}.$$

Choose $\lambda = 4s/\sum_t c_t^2$; the exponent becomes $-2s^2/\sum_t c_t^2$. □

Used on slide 19 and slide 47.

Centered route costs can be dependent

Two routes. Day 1: $p_1 = (1/2, 1/2)$ and costs $(0, 1)$, so $D_1 = \pm 1/2$, revealing I_1. Let day 2's costs depend on day 1's choice:

I_1	ℓ_2	D_2
1	$(0, 0)$	0
2	$(0, 1)$	$\neq 0$

So $D_1 = -1/2$ forces $D_2 = 0$, while $D_1 = 1/2$ makes $D_2 \neq 0$: the two are dependent. Still $\mathbb{E}[D_2 \mid \mathcal{F}_1] = 0$, which is all Azuma needs.

Used on slide 20.

Proof of the smoothness inequality

Let $h = x' - x$. Integrate the derivative of $s \mapsto f(x + sh)$ from 0 to 1:

$$f(x') - f(x) = \int_0^1 \langle \nabla f(x + sh), h \rangle \, ds$$
$$= \langle \nabla f(x), h \rangle + \int_0^1 \langle \nabla f(x + sh) - \nabla f(x), h \rangle \, ds.$$

By Cauchy–Schwarz and smoothness, the last integral is at most

$$\int_0^1 L s \|h\|^2 \, ds = \frac{L}{2} \|h\|^2. \qquad \square$$

Used on slide 26.

Why sampling with replacement matters

Fresh uniform draw:

$$\mathbb{E}[\nabla f_{I_t}(x_t) \mid \mathcal{F}_{t-1}] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x_t) = \nabla f(x_t).$$

One shuffled pass: let J_t be the indices not yet used.

$$\mathbb{E}[\nabla f_{I_t}(x_t) \mid \mathcal{F}_{t-1}] = \frac{1}{|J_t|}\sum_{i \in J_t} \nabla f_i(x_t),$$

which need not equal the full gradient. Overlapping contexts in the corpus are fixed data; the only randomness in our proof is the freshly drawn index.

Used on slide 30.

When the model learns its features

Two parameters (a, b) determine a token probability:

$$P(y) = \frac{1}{1 + e^{-ab}}, \qquad \ell(a, b) = \log(1 + e^{-ab}).$$

At $a = 0$	Value	As $\lvert b\rvert \to \infty$
Gradient entry $\partial_a \ell$	$-b/2$	unbounded
Curvature entry $\partial_{aa} \ell$	$b^2/4$	unbounded

So no single G or L works for all parameters: differentiability alone supplies neither.

Used on slide 33.

Clipping can change the mean

Clip a scalar estimate g to $[-1, 1]$:

Probability	1/3	2/3
Original g	2	-1
Clipped $\tilde{g}$	1	-1

$$\mathbb{E}g = \tfrac{1}{3}(2) + \tfrac{2}{3}(-1) = 0, \qquad \mathbb{E}\tilde{g} = \tfrac{1}{3}(1) + \tfrac{2}{3}(-1) = -\tfrac{1}{3}.$$

Clipping bounds the norm but can destroy unbiasedness.

Used on slide 33.

The gradient of a fixed-feature token loss

For $\phi_v = \phi(c_i, v)$, the observed token y_i has loss

$$f_i(x) = \log \sum_v e^{\langle x, \phi_v \rangle} - \langle x, \phi_{y_i} \rangle.$$

Differentiating,

$$\nabla f_i(x) = \sum_v P_x(v \mid c_i) \phi_v - \phi_{y_i}.$$

Every feature vector has norm at most R, so

$$\|\nabla f_i(x)\| \le \sum_v P_x(v \mid c_i) \|\phi_v\| + \|\phi_{y_i}\| \le 2R.$$

Hence $G = 2R$ works at every parameter vector. □

Used on slide 34.

The curvature of a fixed-feature token loss

Let V have probabilities $P_x(v \mid c_i)$ and let $\mu = \mathbb{E}\phi_V$. Then

$$\nabla^2 f_i(x) = \mathbb{E}[\phi_V \phi_V^\top] - \mu\mu^\top.$$

For every unit vector u,

$$0 \le u^\top \nabla^2 f_i(x) u = \mathrm{Var}(\langle u, \phi_V \rangle) \le \mathbb{E}(\langle u, \phi_V \rangle)^2 \le R^2.$$

So $\|\nabla^2 f_i(x)\|_{\mathrm{op}} \le R^2$, each ∇f_i is R^2-Lipschitz, and averaging keeps the bound: $L = R^2$. □

Used on slide 34.

A bound that separates signal from noise

Suppose instead that $e_t = g_t - \nabla f(x_t)$ has mean zero given $\mathcal{F}_{t-1}$ and

$$\mathbb{E}[\,\|e_t\|^2 \mid \mathcal{F}_{t-1}] \le \sigma^2.$$

Then

$$\mathbb{E}[\,\|g_t\|^2 \mid \mathcal{F}_{t-1}] \le \|\nabla f(x_t)\|^2 + \sigma^2.$$

Put this into the one-step smoothness bound and average given $\mathcal{F}_{t-1}$:

$$\mathbb{E}[f(x_{t+1}) \mid \mathcal{F}_{t-1}] \le f(x_t) - \eta(1 - L\eta/2)\|\nabla f(x_t)\|^2 + L\eta^2\sigma^2/2.$$

For $0 < \eta \le 1/L$, the coefficient of $\|\nabla f(x_t)\|^2$ is at least $\eta/2$.

Used on slide 37.

The variance-based expectation bound

For $0 < \eta \le 1/L$, sum the expected decreases:

$$\frac{\eta}{2} \sum_t \mathbb{E}\|\nabla f(x_t)\|^2 \le \Delta + \frac{L\eta^2\sigma^2 T}{2},$$

so

$$\frac{1}{T} \sum_t \mathbb{E}\|\nabla f(x_t)\|^2 \le \frac{2\Delta}{\eta T} + L\eta\sigma^2.$$

With $\sigma = 0$ and $\eta = 1/L$ this is gradient descent's bound $2L\Delta/T$. Finite variance alone does not give the fixed widths Azuma needs.

Used on slide 37.

Minibatches reduce variance

At fixed x_t, draw k independent uniform indices and average:

$$g_t = \frac{1}{k}\sum_{j=1}^{k} \nabla f_{I_{t,j}}(x_t), \qquad e_{t,j} = \nabla f_{I_{t,j}}(x_t) - \nabla f(x_t).$$

Suppose each $\mathbb{E}[\,\|e_{t,j}\|^2 \mid \mathcal{F}_{t-1}] \le \sigma^2$. The $e_{t,j}$ have mean zero and are independent given $\mathcal{F}_{t-1}$, so the cross terms vanish:

$$\mathbb{E}[\,\|g_t - \nabla f(x_t)\|^2 \mid \mathcal{F}_{t-1}] = \frac{1}{k^2}\sum_{j} \mathbb{E}[\,\|e_{t,j}\|^2 \mid \mathcal{F}_{t-1}] \le \sigma^2/k.$$

The variance drops by a factor k, and each step costs k example gradients.

Used on slide 30.

An average bound does not identify a point

Suppose a run satisfies

$$\frac{1}{T}\sum_{t}\|\nabla f(x_t)\|^2 \le A.$$

For U uniform over $\{1,\dots,T\}$, Markov's inequality given the run gives

$$\mathbb{P}_U\big(\|\nabla f(x_U)\|^2 > 2A \mid \text{run}\big) \le \tfrac{1}{2}.$$

So at least half the points in the run satisfy $\|\nabla f(x_t)\|^2 \le 2A$, and

$$K \text{ independent uniform picks} \quad\Longrightarrow\quad \mathbb{P}(\text{all miss that half}) \le 2^{-K}.$$

Used on slide 48.

Selecting and checking candidates

Let A be Theorem 3's bound with failure probability $\delta/2$.

1. Before training, draw $K = \lceil \log_2(2/\delta) \rceil$ uniform indices and save those iterates.
2. After training, compute the full gradient at each one.
3. Return the candidate with the smallest gradient norm.

$$\mathbb{P}(\text{output fails}) \le \underbrace{\delta/2}_{\text{run fails}} + \underbrace{2^{-K}}_{\text{all candidates miss}} \le \delta.$$

So $\|\nabla f(x_{\text{out}})\|^2 \le 2A$ with probability at least $1 - \delta$, for nK extra example gradients and $O(Kd)$ extra memory. □

Used on slide 48.

Reading

Martingales and concentration: Mitzenmacher and Upfal, *Probability and Computing*, the chapter on martingales and the Azuma–Hoeffding inequality.

Online learning: Cesa-Bianchi and Lugosi, *Prediction, Learning, and Games*, Chapter 2, for exponential weights and regret.

Language modeling: Bengio, Ducharme, Vincent, and Jauvin, *A Neural Probabilistic Language Model*, JMLR 3 (2003), 1137–1155. jmlr.org/papers/v3/bengio03a.html

SGD concentration: Liu, Nguyen, Nguyen, Ene, and Nguyen, *High Probability Convergence of Stochastic Gradient Methods*, ICML 2023, Section 4.1. proceedings.mlr.press/v202/liu23aa.html