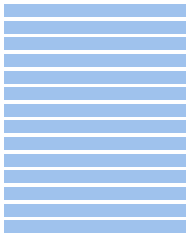
Stochastic gradient descent

208424 Lecture 7: SGD, momentum, Adam

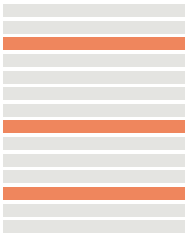
Rows used per step: GD versus SGD

gradient descent



every row, every step

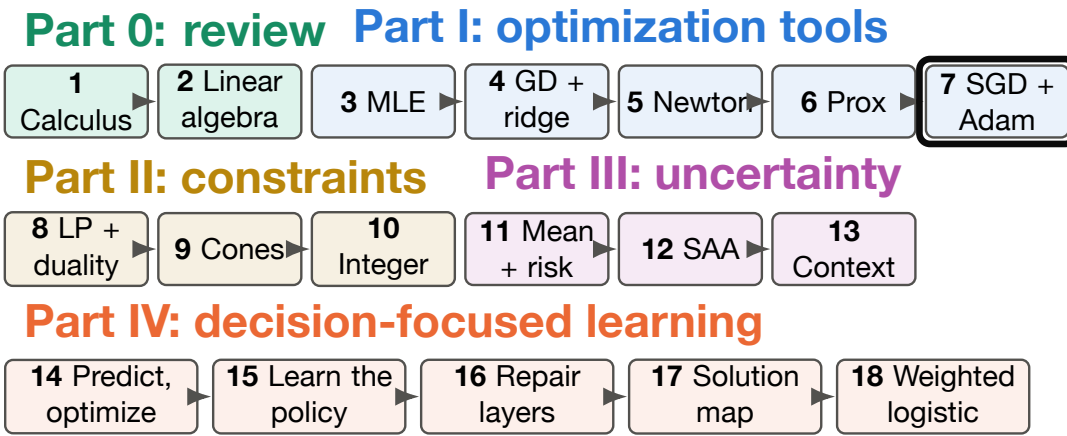
SGD



a few random rows

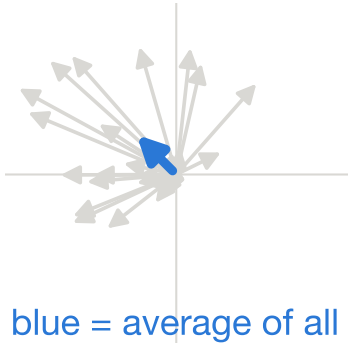
one GD step = one gradient per row

Course map


Part 0: review
Part I: optimization tools
1 Calculus
2 Linear algebra
3 MLE
4 GD + ridge
5 Newton
6 Prox
7 SGD + Adam
Part II: constraints
Part III: uncertainty
8 LP + duality
9 Cones
10 Integer
11 Mean + risk
12 SAA
13 Context
Part IV: decision-focused learning
14 Predict, optimize
15 Learn the policy
16 Repair layers
17 Solution map
18 Weighted logistic

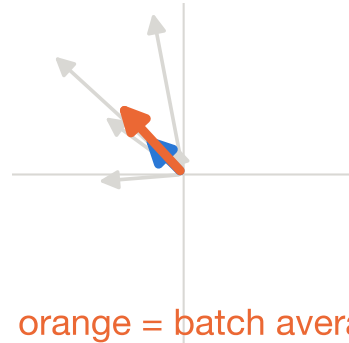
The gradient is an average

one arrow per customer



blue = average of all

5 random customers

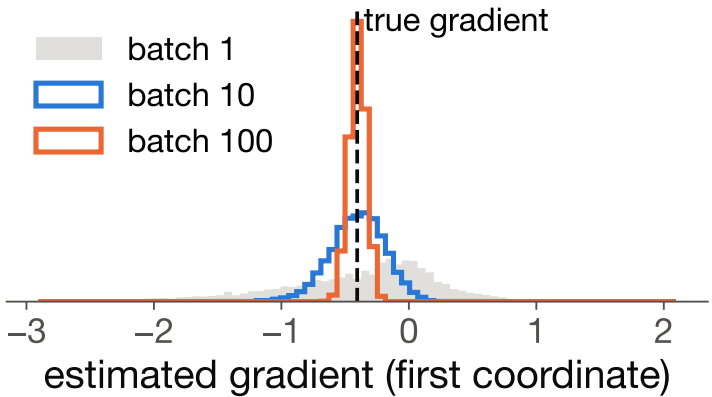


orange = batch average

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

f_i: the loss on row i

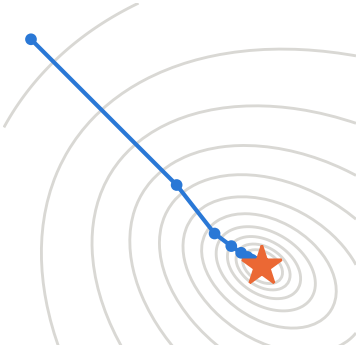
A batch gradient is unbiased, but noisy



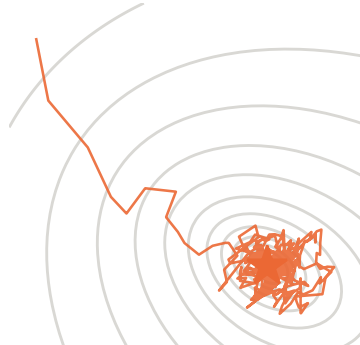
batch $\times 10$ $\Rightarrow$ SD (Lecture 5) $\div \sqrt{10}$, about 3

SGD: many cheap, noisy steps

GD: 30 steps = 30 passes



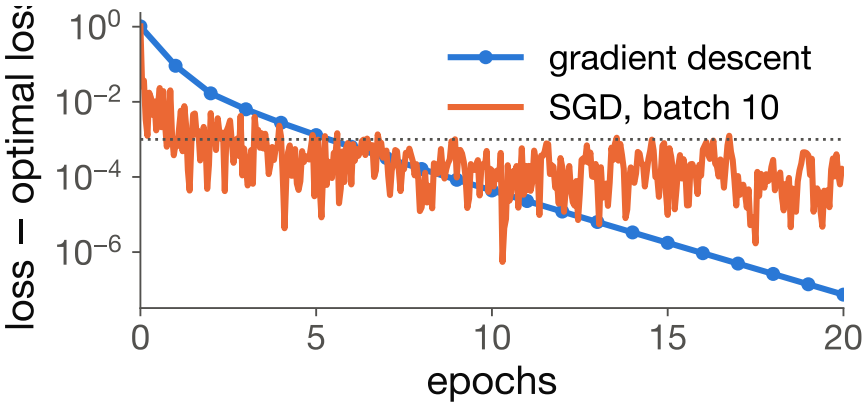
SGD: 300 steps = 0.3 passes



$$w \leftarrow w - \eta \nabla f_B(w), \quad B = \text{a random batch}$$

∇f_B: the average gradient of the rows in B

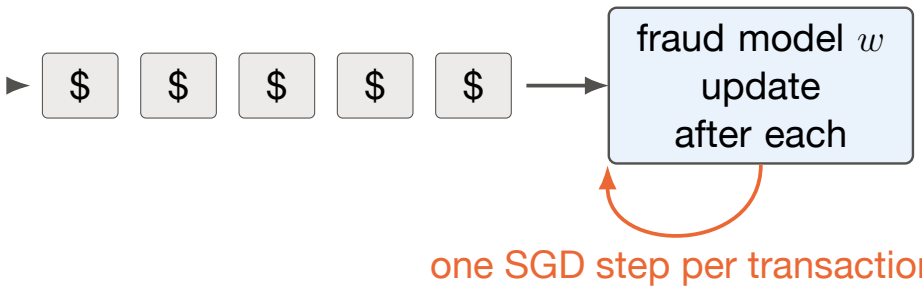
Progress per amount of data read



epoch = one pass over all the rows

first below 10^{-3}: GD 6 epochs, SGD 0.75

Online learning: SGD while data arrive



transactions, clicks, messages

Summary: stochastic gradient descent

- Let $f_i(w)$ be the loss on row i, and $f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$.
- Let B be a random batch of rows, $\nabla f_B(w)$ the average of their $\nabla f_i(w)$.
- We want: $\min_w f(w)$, but each GD step uses all n rows.
- SGD: $w \leftarrow w - \eta \nabla f_B(w)$ with step size η, using only the rows in B.
- Unbiased (right on average), but noisy: the noise SD shrinks like $1/\sqrt{\text{batch size}}$.

Checkpoint 1

$n = 1{,}000{,}000$ customers, batch size 100

How many SGD steps cost as much as one GD step?

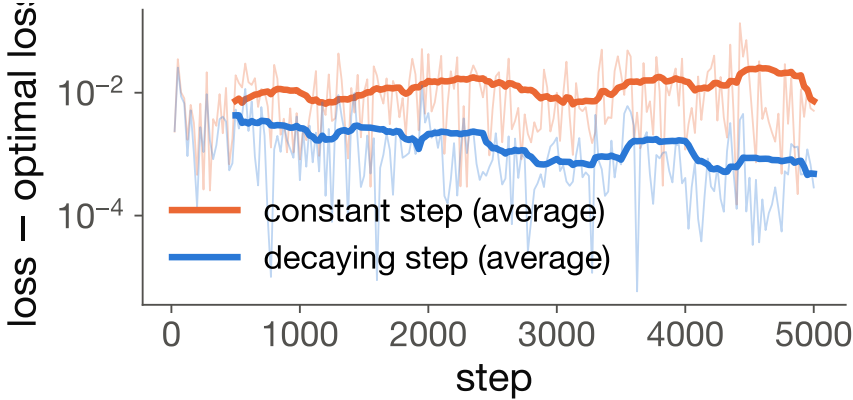
Checkpoint 1

$n = 1{,}000{,}000$ customers, batch size 100

How many SGD steps cost as much as one GD step?

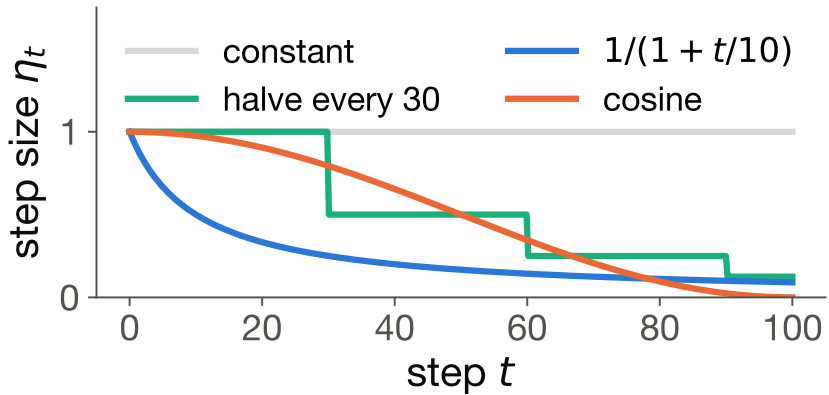
$$1{,}000{,}000/100 = 10{,}000$$

Constant versus decaying step size



shrink the step over time: noise averages out

Common step-size schedules



start large to move fast, end small to settle

Summary: step-size schedules

- Let η_t be the SGD step size at step t.
- We want: SGD to stay near the optimal weights.
- Constant η: new noise each step; shrinking η_t slowly: the noise averages out.
- Schedules: halve every 30 steps, $1/(1 + t/10)$, cosine.
- Bigger batch: less noise, but each step costs more.

Checkpoint 2

batch size $25 \rightarrow 100$

What happens to the noise (SD)?

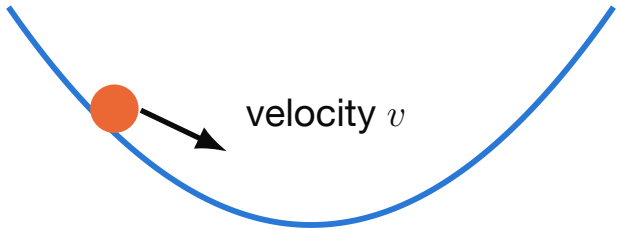
Checkpoint 2

batch size $25 \rightarrow 100$

What happens to the noise (SD)?

4 times the data: SD $\times \frac{1}{\sqrt{4}} = \frac{1}{2}$

Momentum: add up past gradients



$$v \leftarrow \beta\, v + \nabla f(w) \qquad w \leftarrow w - \eta\, v$$

same direction adds up: $\beta = 0.5$: $v = g,\ 1.5g,\ 1.75g, \cdots \rightarrow 2g$

gradients that flip sign cancel

Momentum on an ill-conditioned loss

gradient descent: 327 steps

momentum: 45 steps



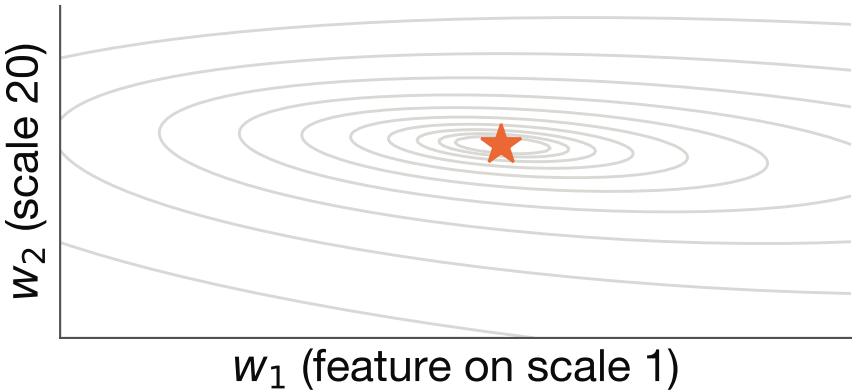
distance 10^{-6}: GD 327 steps ($\eta = 1/L$), 173 (best fixed η)

momentum, tuned to this loss: 45 steps

Summary: momentum

- Let v be the velocity, $0 < \beta < 1$ the share of v kept.
- We want: $\min_w f(w)$ when f is ill-conditioned, where GD is slow (Lecture 4).
- Momentum: $v \leftarrow \beta\, v + \nabla f(w), \quad w \leftarrow w - \eta\, v$, step size η.
- Gradients in one direction add up; gradients that flip sign cancel.
- Ill-conditioned, to 10^{-6}: momentum 45 steps; GD 327 ($\eta = 1/L$), 173 at its best η.

Features on different scales



one feature 20 times larger: ill-conditioned, $\kappa = 394$

Adam: a step size for each weight

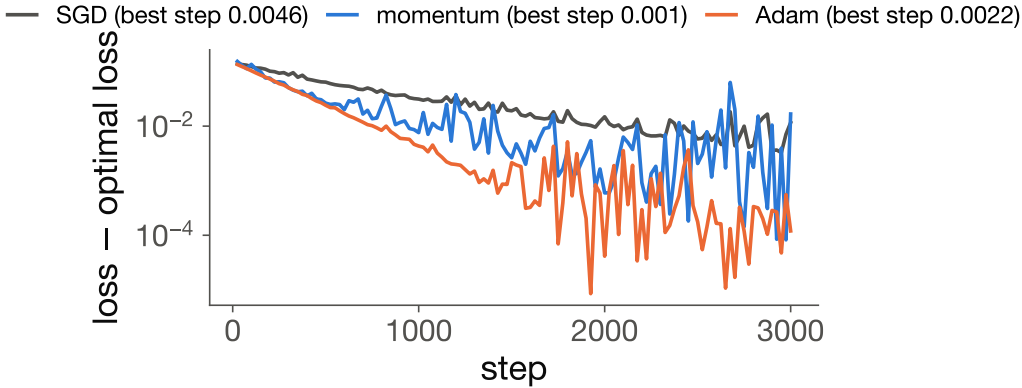
$m_j \leftarrow \beta_1 m_j + (1-\beta_1)\, g_j$: average of recent gradients g_j

$s_j \leftarrow \beta_2 s_j + (1-\beta_2)\, g_j^2$: of recent *squared* gradients

$$w_j \leftarrow w_j - \eta \frac{m_j}{\sqrt{s_j}}$$ for each weight j

defaults: $\beta_1 = 0.9,\ \beta_2 = 0.999,\ \eta = 10^{-3}$

Adam on the badly scaled problem



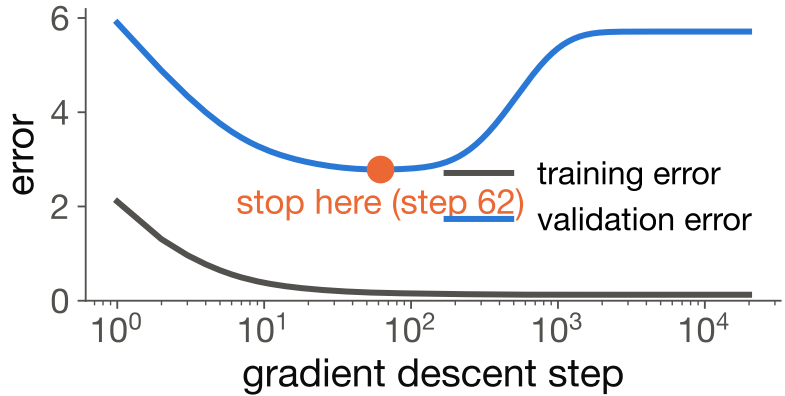
loss − optimal loss, each method at its best step, one run

Adam 2.4×10^{-4} · SGD 8.9×10^{-3}

Summary: Adam

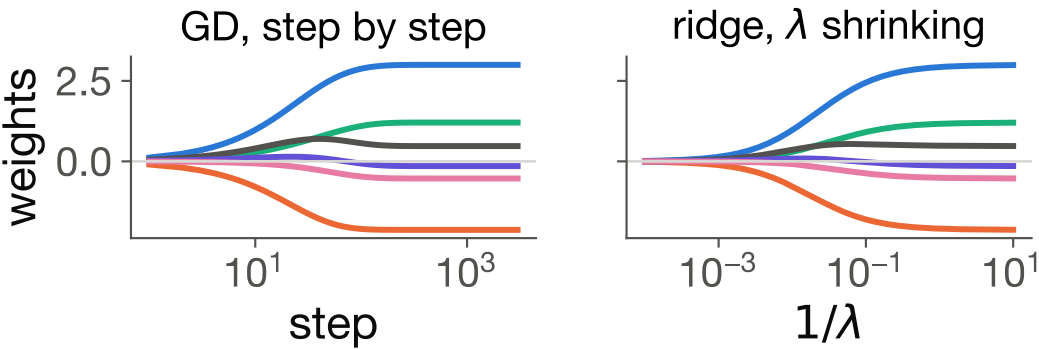
- Let m_j, s_j be the averages of recent gradients of w_j and of their squares (weights β_1, β_2).
- Let η be the step size.
- We want: $\min_w f(w)$, features on very different scales.
- Adam: $w_j \leftarrow w_j - \eta\, m_j / \sqrt{s_j}$ for each weight j.
- Badly scaled data, one run, best step each: Adam about 37 times closer to the optimal loss than SGD.

Stopping early prevents overfitting



keep the weights with the smallest *validation* error

Stopping early acts like ridge (Lecture 4)



least squares: $\|w\|$ grows, as when λ shrinks

Summary: early stopping

- Let training error be the error on the rows we fit, validation error on new rows.
- GD from $w = 0$: training error falls; validation error falls, then rises.
- We want: the weights w with the smallest *validation* error.
- Early stopping: keep w from that step (here step 62).
- Like ridge: $\|w\|$ grows as the ridge penalty λ shrinks.

Checkpoint 3

training error keeps falling for 1000 steps

validation error is lowest at step 200, then rises

Which weights do you keep?

Checkpoint 3

training error keeps falling for 1000 steps

validation error is lowest at step 200, then rises

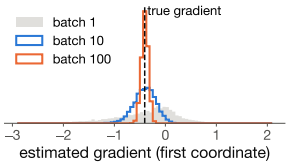
Which weights do you keep?

the weights from step 200

A practical recipe

1. standardize features, shuffle the rows
2. batches of 32 to 256
3. Adam ($\eta = 10^{-3}$) or SGD with momentum ($\beta = 0.9$)
4. lower η over time
5. watch validation error, keep the best weights

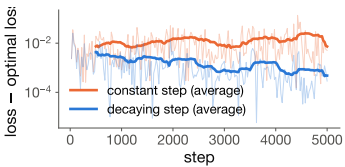
Lecture summary



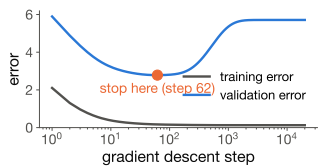
batches: unbiased, noisy



momentum: fewer steps

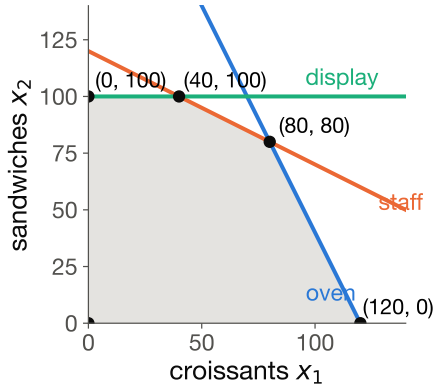


shrink the step over time



stop at the best validation error

Next lecture: decisions with constraints



the optimal plan when oven, staff and display set constraints

Exercise

1. $n = 500{,}000$, batch 50: how many SGD steps per GD step?
2. Batch $16 \to 64$: what happens to the noise SD?
3. Momentum $\beta = 0.9$ and a constant gradient g: what does v approach?
4. Validation error is lowest at step 80 of 1000. Which weights do you keep?