CS391L: Machine Learning WB

Stochastic Optimization

Inderjit S. Dhillon
UT Austin

March 30, 2025

Stochastic Gradient Descent

Large-scale Problems

- Machine learning: usually minimize the training loss

$$\min_{\boldsymbol{w}} \{\frac{1}{N}\sum_{n=1}^{N} \ell(\boldsymbol{w}^T \boldsymbol{x}_n, y_n)\} := f(\boldsymbol{w}) \text{ (linear model)}$$

$$\min_{\boldsymbol{w}} \{\frac{1}{N}\sum_{n=1}^{N} \ell(h_{\boldsymbol{w}}(\boldsymbol{x}_n), y_n)\} := f(\boldsymbol{w}) \text{ (general hypothesis)}$$

  ℓ: loss function (e.g., $\ell(a, b) = (a - b)^2$)

- Gradient descent:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \underbrace{\nabla f(\boldsymbol{w})}_{\text{main computation}}$$

Large-scale Problems

- Machine learning: usually minimize the training loss

$$\min_{\boldsymbol{w}} \{\frac{1}{N}\sum_{n=1}^{N} \ell(\boldsymbol{w}^T \boldsymbol{x}_n, y_n)\} := f(\boldsymbol{w}) \text{ (linear model)}$$

$$\min_{\boldsymbol{w}} \{\frac{1}{N}\sum_{n=1}^{N} \ell(h_{\boldsymbol{w}}(\boldsymbol{x}_n), y_n)\} := f(\boldsymbol{w}) \text{ (general hypothesis)}$$

  ℓ: loss function (e.g., $\ell(a, b) = (a - b)^2$)

- Gradient descent:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \underbrace{\nabla f(\boldsymbol{w})}_{\text{main computation}}$$

- In general, $f(\boldsymbol{w}) = \frac{1}{N}\sum_{n=1}^{N} f_n(\boldsymbol{w})$,

  each $f_n(\boldsymbol{w})$ only depends on $(\boldsymbol{x}_n, y_n)$

Stochastic gradient

- Gradient:

$$\nabla f(\boldsymbol{w}) = \frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{w})$$

- Each gradient computation needs to go through all training samples

  slow when millions of samples

- Faster way to compute "approximate gradient"?

Stochastic gradient

- Gradient:

$$\nabla f(\boldsymbol{w}) = \frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{w})$$

- Each gradient computation needs to go through all training samples

  slow when millions of samples

- Faster way to compute "approximate gradient"?
- Use stochastic sampling:
  - Sample a small subset $B \subseteq \{1, \cdots, N\}$
  - Estimated gradient

$$\nabla f(\boldsymbol{w}) \approx \frac{1}{|B|}\sum_{n\in B}\nabla f_n(\boldsymbol{w})$$

  $|B|$: batch size

Stochastic gradient descent

Stochastic Gradient Descent (SGD)

- Input: training data $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$
- Initialize $\boldsymbol{w}$ (zero or random)
- For $t = 1, 2, \cdots$
  - Sample a small batch $B \subseteq \{1, \cdots, N\}$
  - Update parameter

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta^t \frac{1}{|B|} \sum_{n \in B} \nabla f_n(\boldsymbol{w})$$

Extreme case: $|B| = 1 \Rightarrow$ Sample one training data at a time

Logistic Regression by SGD

- Logistic regression:

$$\min_{\boldsymbol{w}} \frac{1}{N} \sum_{n=1}^{N} \underbrace{\log(1 + e^{-y_n \boldsymbol{w}^T \boldsymbol{x}_n})}_{f_n(\boldsymbol{w})}$$

SGD for Logistic Regression

- Input: training data $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$
- Initialize $\boldsymbol{w}$ (zero or random)
- For $t = 1, 2, \cdots$
  - Sample a batch $B \subseteq \{1, \cdots, N\}$
  - Update parameter

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta^t \frac{1}{|B|} \sum_{i \in B} \underbrace{\frac{-y_n \boldsymbol{x}_n}{1 + e^{y_n \boldsymbol{w}^T \boldsymbol{x}_n}}}_{\nabla f_n(\boldsymbol{w})}$$

Why SGD works?

- Stochastic gradient is an unbiased estimator of full gradient:

$$E[\frac{1}{|B|}\sum_{n\in B}\nabla f_n(\boldsymbol{w})] = \frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{w})$$
$$= \nabla f(\boldsymbol{w})$$

- Each iteration updated by

$$\text{gradient} + \text{zero-mean noise}$$

Stochastic gradient descent

- In gradient descent, η (step size) is a fixed constant
- Can we use fixed step size for SGD?

Stochastic gradient descent

- In gradient descent, η (step size) is a fixed constant
- Can we use fixed step size for SGD?
- SGD with fixed step size cannot converge to global/local minimizers
- If $\boldsymbol{w}^*$ is the minimizer, $\nabla f(\boldsymbol{w}^*) = \frac{1}{N}\sum_{n=1}^{N} \nabla f_n(\boldsymbol{w}^*) = 0$,

$$\text{but} \quad \frac{1}{|B|}\sum_{n\in B} \nabla f_n(\boldsymbol{w}^*) \neq 0 \quad \text{if } B \text{ is a subset}$$

(Even if we got minimizer, SGD will move away from it)

Stochastic gradient descent, step size

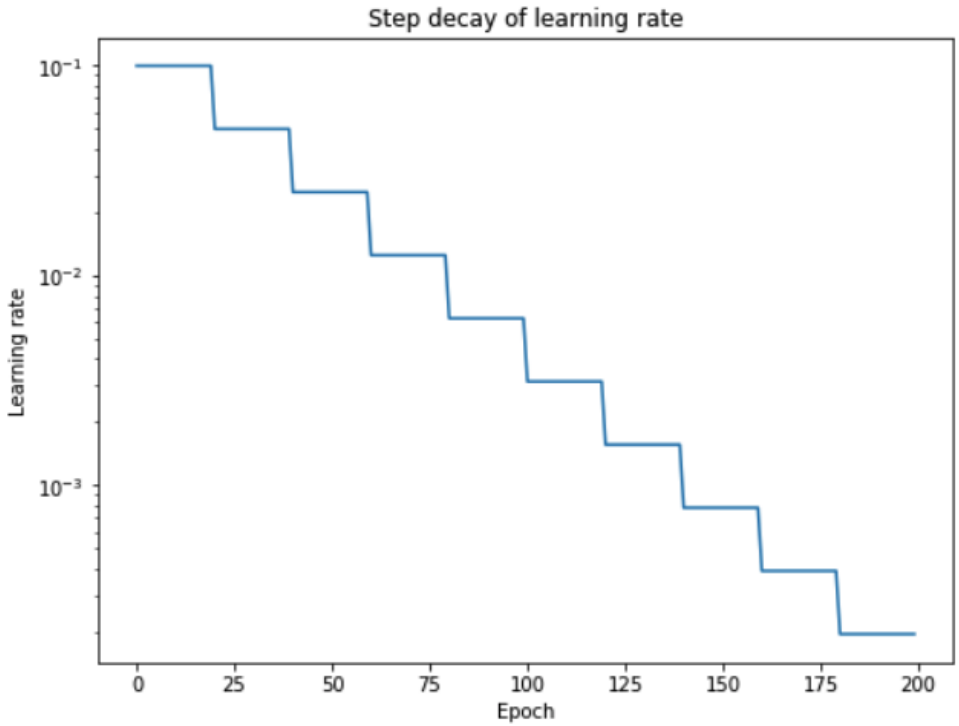
- To make SGD converge:

  Step size should decrease to 0

  $\eta^t \to 0$

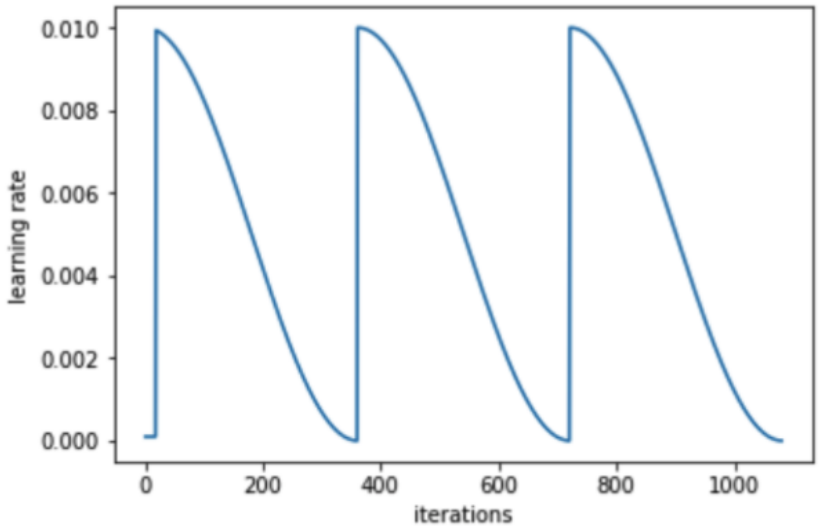
  Usually with polynomial rate: $\eta^t \approx t^{-a}$ with constant $a > 0$

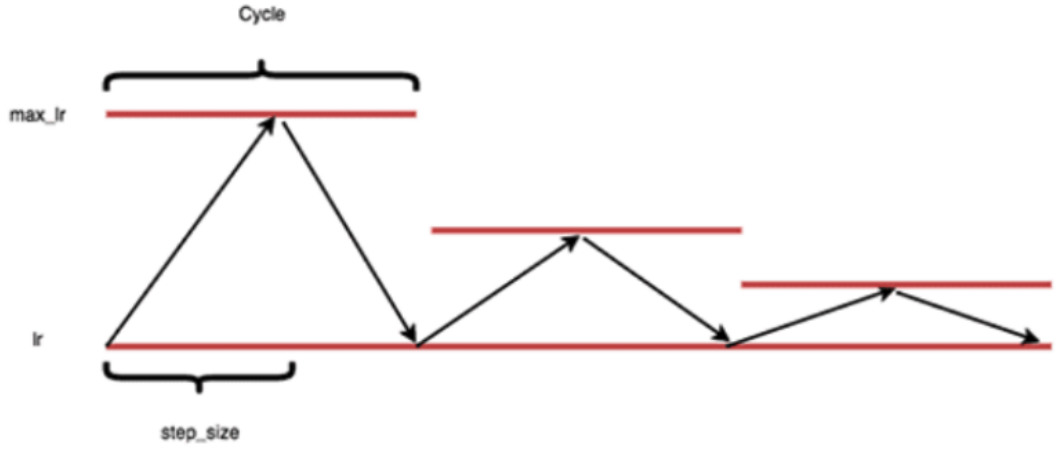
- Step decay of learning rate



Learning rate scheduling

- Other cyclic learning rate scheduling strategies

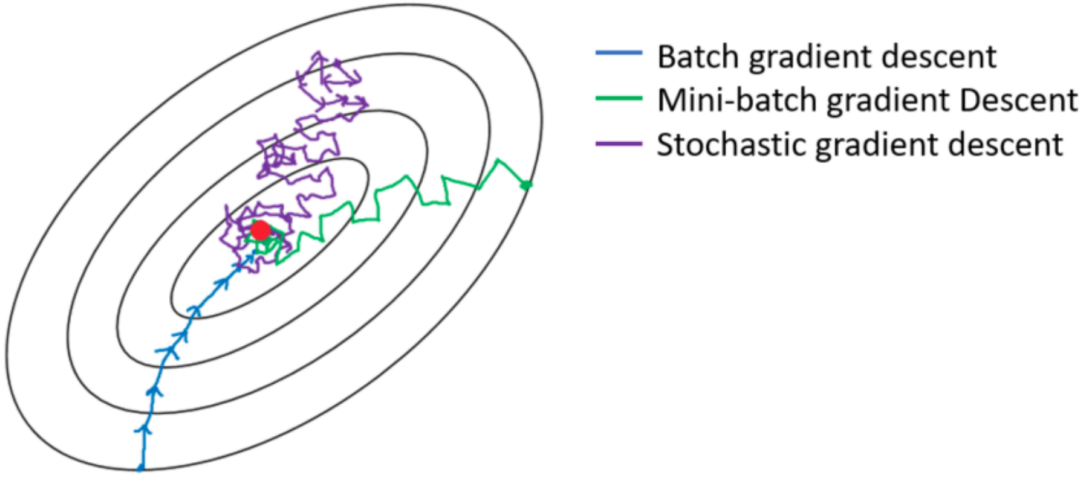




Stochastic gradient descent vs Gradient descent

Stochastic gradient descent:

- pros:
  - cheaper computation per iteration
  - faster convergence in the beginning
- cons:
  - less stable, slower final convergence
  - hard to tune step size



(Figure from `https://medium.com/@ImadPhd/gradient-descent-algorithm-and-its-variants-10f652806a3`)

Momentum

- Gradient descent: only using current gradient (local information)
- Momentum: use previous gradient information
- The momentum update rule:

$$\boldsymbol{v}_t = \beta \boldsymbol{v}_{t-1} + (1-\beta)\nabla f(\boldsymbol{w}_t)$$
$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \alpha \boldsymbol{v}_t$$

  $\beta \in [0, 1)$: discount factors, α: step size

- Equivalent to using moving average of gradient:

$$\boldsymbol{v}_t = (1-\beta)\nabla f(\boldsymbol{w}_t) + \beta(1-\beta)\nabla f(\boldsymbol{w}_{t-1}) + \beta^2(1-\beta)\nabla f(\boldsymbol{w}_{t-2}) + \cdots$$

Momentum gradient descent

Momentum gradient descent

- Initialize $\boldsymbol{w}_0, \boldsymbol{v}_0 = 0$
- For $t = 1, 2, \cdots$
  - Compute $\boldsymbol{v}_t \leftarrow \beta \boldsymbol{v}_{t-1} + (1 - \beta)\nabla f(\boldsymbol{w}_t)$
  - Update $\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \alpha \boldsymbol{v}_t$

α: learning rate

β: discount factor ($\beta = 0$ means no momentum)

Momentum stochastic gradient descent

Optimizing $f(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} f_i(\boldsymbol{w})$

Momentum stochastic gradient descent

- Initialize $\boldsymbol{w}_0, \boldsymbol{v}_0 = 0$
- For $t = 1, 2, \cdots$
  - Sample an $i \in \{1, \cdots, N\}$
  - Compute $\boldsymbol{v}_t \leftarrow \beta \boldsymbol{v}_{t-1} + (1-\beta)\nabla f_i(\boldsymbol{w}_t)$
  - Update $\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \alpha \boldsymbol{v}_t$

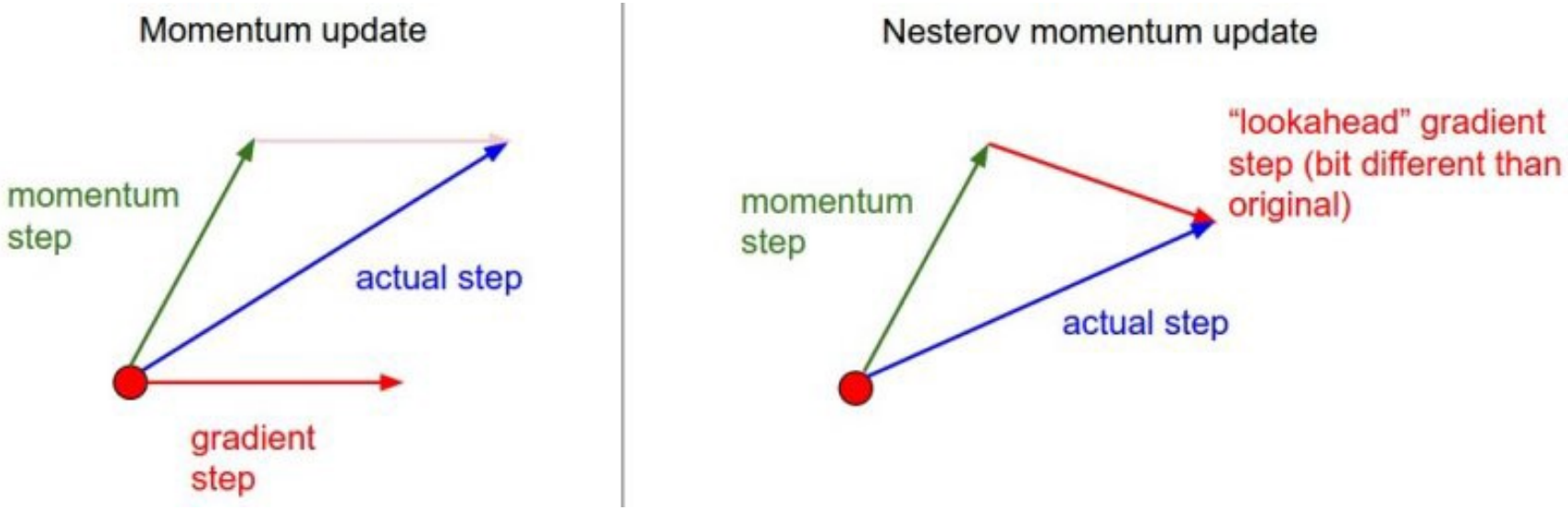
α: learning rate

β: discount factor ($\beta = 0$ means no momentum)

Nesterov accelerated gradient

- Using the "look-ahead" gradient

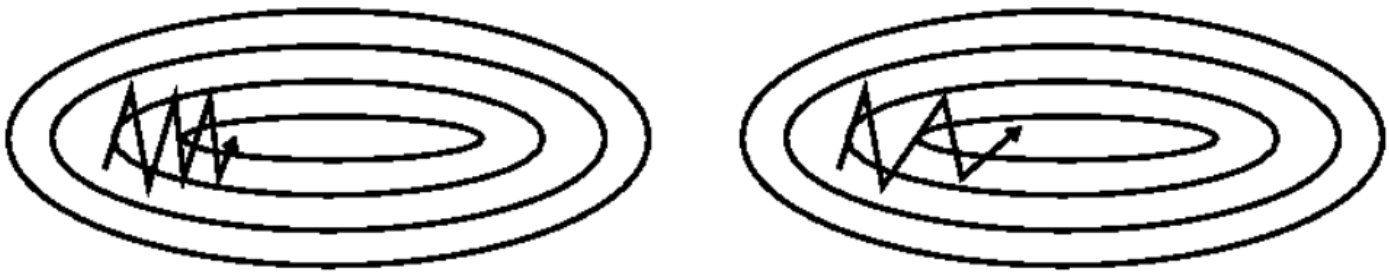
$$\boldsymbol{v}_t = \beta \boldsymbol{v}_{t-1} + \alpha \nabla f(\boldsymbol{w}_t - \beta \boldsymbol{v}_{t-1})$$
$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \boldsymbol{v}_t$$



(Figure from `https://towardsdatascience.com`)

Why momentum works?

- Reduce variance of gradient estimator for SGD
- Even for gradient descent, it's able to speed up convergence in some cases:



Left—SGD without momentum, right— SGD with momentum. (Source: Genevieve B. Orr)

Adagrad: Adaptive updates (2010)

- SGD update: same step size for all variables
- Adaptive algorithms: each dimension can have a different step size

Adagrad

- Initialize $\boldsymbol{w}_0$
- For $t = 1, 2, \cdots$
  - Sample an $i \in \{1, \cdots, N\}$
  - Compute $\boldsymbol{g}^t \leftarrow \nabla f_i(\boldsymbol{w}_t)$
  - $G_j^t \leftarrow G_j^{t-1} + (g_j^t)^2$ for all $j = 1, \ldots d$
  - Update $\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \frac{\eta}{\sqrt{G_i^t + \epsilon}} g_i^t$

η: step size (constant)

ϵ: small constant to avoid division by 0

Adagrad

- For each dimension i, we have observed T samples $g_i^1, \cdots, g_i^t$
- Standard deviation of g_i:

$$\sqrt{\frac{\sum_{t'}(g_i^{t'})^2}{t}} = \sqrt{\frac{G_i^t}{t}}$$

- Assume step size is $\eta/\sqrt{t}$, then the update becomes

$$w_i^{t+1} \leftarrow w_i^t - \frac{\eta}{\sqrt{t}} \frac{\sqrt{t}}{\sqrt{G_i^t}} g_i^t$$

Adam: Momentum + Adaptive updates (2015)

Adam

- Initialize $\boldsymbol{w}_0, \boldsymbol{m}_0 = 0, \boldsymbol{v}_0 = 0,$
- For $t = 1, 2, \cdots$
  - Sample an $i \in \{1, \cdots, N\}$
  - Compute $\boldsymbol{g}_t \leftarrow \nabla f_i(\boldsymbol{w}_t)$
  - $\boldsymbol{m}_t \leftarrow \beta_1 \boldsymbol{m}_{t-1} + (1 - \beta_1)\boldsymbol{g}_t$
  - $\boldsymbol{v}_t \leftarrow \beta_2 \boldsymbol{v}_{t-1} + (1 - \beta_2)\boldsymbol{g}_t^2$
  - $\hat{\boldsymbol{m}}_t \leftarrow \boldsymbol{m}_t / (1 - \beta_1^t)$
  - $\hat{\boldsymbol{v}}_t \leftarrow \boldsymbol{v}_t / (1 - \beta_2^t)$
  - Update $\boldsymbol{w}_t \leftarrow \boldsymbol{w}_{t-1} - \alpha \cdot \hat{\boldsymbol{m}}_t / (\sqrt{\hat{\boldsymbol{v}}_t} + \epsilon)$

Batch size selection

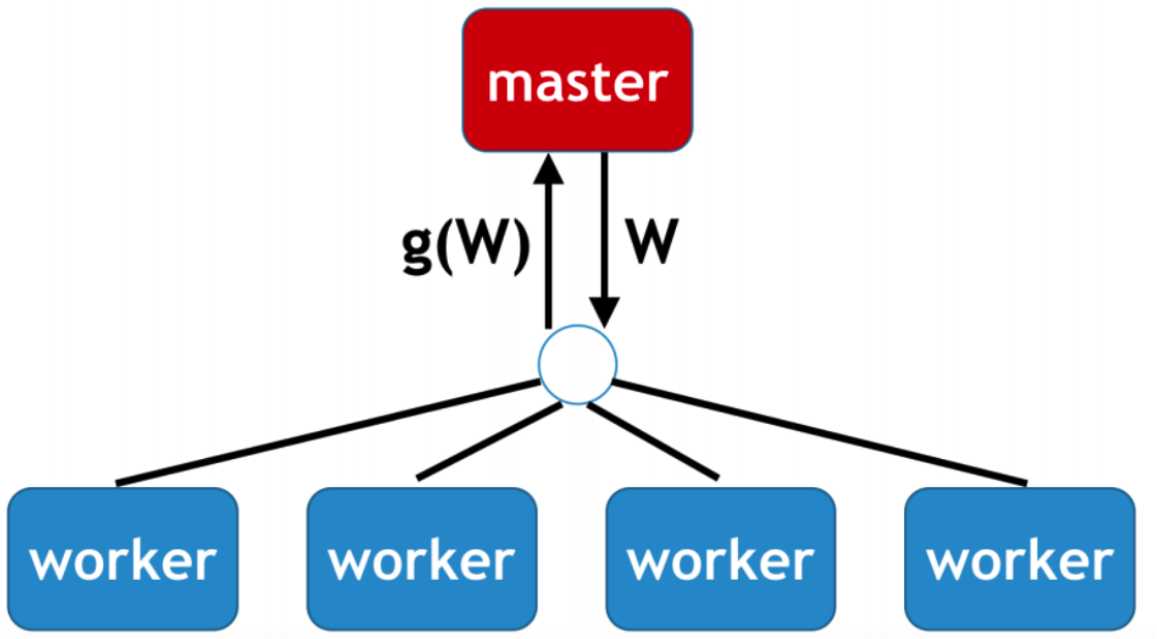
- Larger batch size $\Rightarrow$ more computation per update, less noise
- Usually, choose batch size large enough that
  - Fits in (GPU) memory
  - Can fully utilize the computation resource
- For example, 512 batch size for ImageNet training on a standard GPU.

Large batch training

- What if we want to train a model with hundreds or thousands of GPUs
- Data parallel distributed computing:

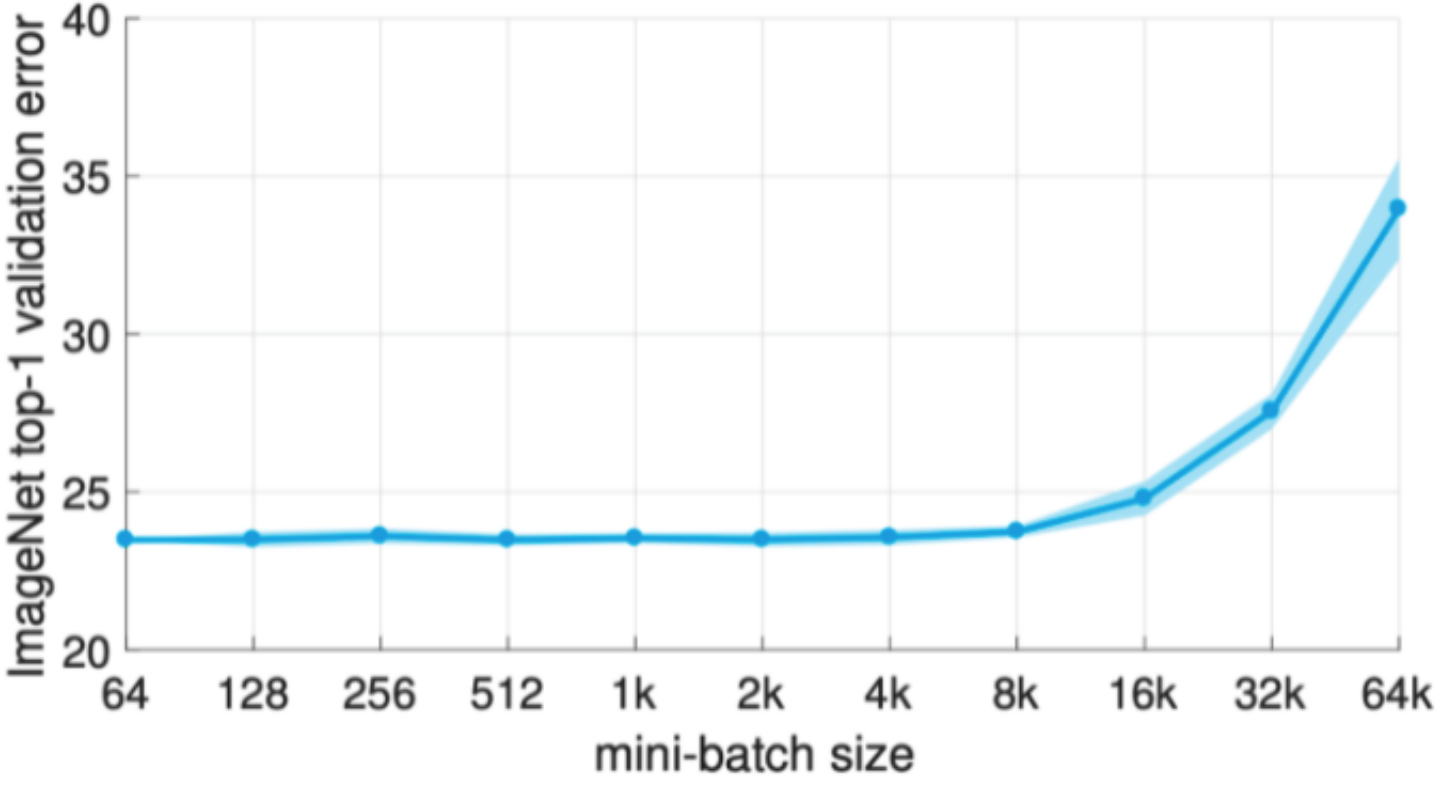
  Increasing the batch size linearly with number of devices.

  $\Rightarrow$ Large batch training



Problem of large batch training

- Tend to converge to models with lower test accuracy when using very large batch size



(Goyal et al., 2017) Training Imagenet in 1 hour

Sharp vs wide local minimum

- Large-batch SGD/Adam:
  - Usually converge to a sharp local minimum (not enough inherent noise in SGD)
  - Harder to generalize to test data

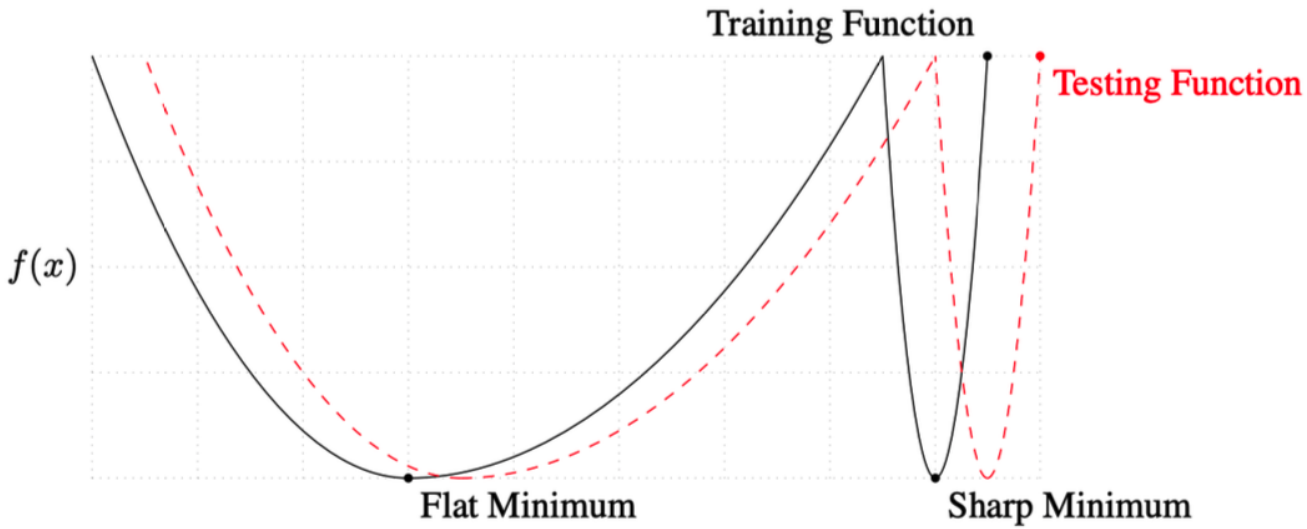


Figure from (Keskar et al., 2017)

A simple but practical solution: learning rate scaling

- $\mathrm{Var}[\frac{1}{|B|}\sum_{n\in B}\boldsymbol{g}_n] \approx O(\frac{1}{|B|})$
- Batch size $\uparrow$, learning rate $\uparrow$
  - LR scaling: LR as $O(\sqrt{|B|})$ or $O(|B|)$

A simple but practical solution: learning rate scaling

- $\mathrm{Var}[\frac{1}{|B|}\sum_{n\in B}\boldsymbol{g}_n] \approx O(\frac{1}{|B|})$
- Batch size $\uparrow$, learning rate $\uparrow$
  - LR scaling: LR as $O(\sqrt{|B|})$ or $O(|B|)$
- However, LR has to be bounded for convergence (as for GD)

$$LR \leq O(1/L) \qquad L : \text{Lipchitz constant}$$

- Can't increase LR without limit

Non-uniform updates between different layers

- Some layer becomes the bottleneck of LR

Layers	$\|\|w\|\|_2$	$\|\|\nabla w\|\|_2$	$\|\|w\|\|_2/\|\|\nabla w\|\|_2$
fc8.0	20.24	0.078445	258
fc8.1	0.316	0.006147	51
fc7.0	20.48	0.110949	184
fc7.1	6.400	0.004939	1296
fc6.0	30.72	0.097996	314
fc6.1	6.400	0.001734	**3690**
conv5.0	6.644	0.034447	193
conv5.1	0.160	0.000961	166
conv4.0	8.149	0.039939	204
conv4.1	0.196	0.000486	403
conv3.0	9.404	0.049182	191
conv3.1	0.196	0.000511	384
conv2.0	5.545	0.057997	96
conv2.1	0.160	0.000649	247
conv1.0	1.866	0.071503	26
conv1.1	0.098	0.004909	**20**

Another solution

- Use **Layer-wise Adaptive LR Scaling**

$$\boldsymbol{w}^{(i)} \leftarrow \boldsymbol{w}^{(i)} - \eta \frac{\|\boldsymbol{w}^{(i)}\|}{\|\boldsymbol{g}^{(i)}\|} \boldsymbol{g}^{(i)},$$

  where $(\cdot)^{(i)}$ means the i-th layer of neural network
- LARS: $\boldsymbol{g}$ is the stochastic gradient

  LAMB: $\boldsymbol{g}$ is the Adam update

  ("Large Batch Optimization for Deep Learning: Training BERT in 76 minutes, ICLR 2020")
- Google and Nvidia both use LAMB to train BERT within 1 minute (4096 TPU / 2048 GPU)

Conclusions

- Stochastic gradient descent
- Variants of SGD used in neural network training.

Questions?