Minibatch Stochastic Gradient Descent

Outline

- General neural network process
- (Batch) gradient descent
- Stochastic gradient descent (SGD)
- Minibatch stochastic gradient descent (Minibatch SGD)
- Discuss tradeoffs

Recap: Automatic Differentiation

- Take five minutes to draw
  - Whatever will help you remember (no correct or incorrect drawings)
  - You’ll keep a running drawing log the rest of the semester

What do MSE and MAE look like in code?

	Symbolic	Vectorized	Vectorized Tensors code
MSE:	$\frac{1}{N}\sum_{i=1}^{N}(\hat{y}^{(i)} - y^{(i)})^2$	$\|\|(\hat{y} - y)^2\|\|$	((yhat - y)**2).mean(
MAE	$\frac{1}{N}\sum_{i=1}^{N}\|\hat{y}^{(i)} - y^{(i)}\|$	$\|\|\hat{Y} - Y\|\|_1$	(yhat - y).abs().mean() abs(yhat - y)

Mean

Matrix.py
(compute graph)

$$\frac{a+b+c\ldots+z}{26}$$

Abs (Matrix)

mean (Matrix)

```
class Matrix:
    ...

    def mean(self):
        r = Matrix(List2D(1, 1, self.d.mean()), children=[self])

        def _grad():  # r, self
            self.grad += List2D(self.nrow, self.ncol, 1/n) . r.grad.vals[0][0]

        return r
```

List2D(self.nrow, self.ncol, 1/n): from mean gradient

r.grad.vals[0][0]: back prop

list2d.py
(linear algebra)

```
def mean(self):
    sum = 0
    for i in self.vals
        for j in i
            sum += j

    return sum/n
```

Mean

```python
def mean(self) -> Matrix:
    """Return the mean of all values across both dimensions."""
    result = Matrix(List2D(1, 1, self.data.mean()), children=(self,))

    def _gradient() -> None:
        info(f"Gradient of mean. Shape: {self.shape}")
        n = self.nrow * self.ncol
        self.grad += List2D(self.nrow, self.ncol, result.grad.vals[0][0] / n)

    result._gradient = _gradient
    return result
```

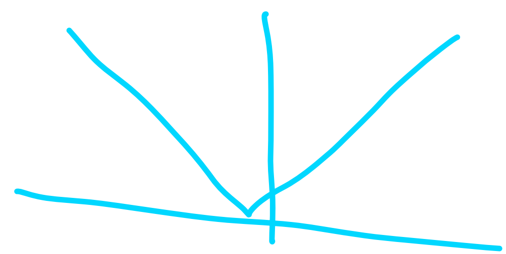
```python
def mean(self) -> float:
    """Compute the mean of all values in the matrix."""
    return self.sum() / (self.nrow * self.ncol)
```

Absolute Value

```
def   abs(self)
    r = Matrix(...)

    def _grad():                          ?
        self.grad += (self.data > 0) * r.grad
        self.grad += -(self.data <= 0) * r.grad
                                          backprop

    return  r
```

```
T = match.randn(r,c)
A = T.abs()
A = abs(T)
A.backward()
```

$$\frac{\partial A}{\partial t_{i,j}} = \begin{cases} 1 & \text{if } t_{i,j} \geq \\ -1 & \text{otherwise} \end{cases}$$



Absolute Value

```python
def abs(self) -> Matrix:
    """Return the element-wise absolute matrix values."""
    result = Matrix(self.data.abs(), children=(self,))

    def _gradient() -> None:
        self.grad += (self.data > 0) * result.grad
        self.grad += -(self.data < 0) * result.grad

    result._gradient = _gradient
    return result

def __abs__(self) -> Matrix:
    """Return the element-wise absolute matrix values."""
    return self.abs()
```

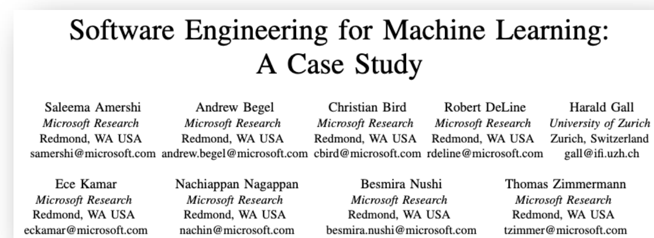
$$\underbrace{\begin{bmatrix} 2.3 & 6.1 \\ 1.1 & -0.2 \end{bmatrix} > 0}$$

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

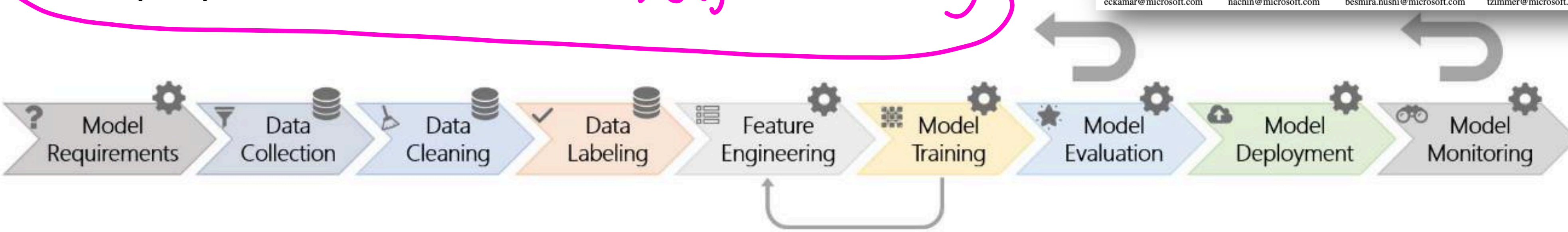
```python
def abs(self) -> List2D:
    data = [
        [abs(self.data[i][j]) for j in range(self.ncol)]
        for i in range(self.nrow)
    ]
    return List2D(*self.shape, data)

def __abs__(self) -> List2D:
    return self.abs()
```

General Neural Network Process

1. Prepare your dataset
   - Normalize inputs
   - Create a proxy (smaller) dataset for debugging and testing workflows
   - Split the dataset into training, validation, and evaluation
2. Design a neural network (architecture)
3. Set initial hyperparameters (learning rate, number of epochs, etc.)
4. Train model parameters (includes validation)
5. Evaluate the trained model
6. Deploy the trained model longest lasting


Software Engineering for Machine Learning: A Case Study

Saleema Amershi, Microsoft Research, Redmond, WA USA, samershi@microsoft.com
Andrew Begel, Microsoft Research, Redmond, WA USA, andrew.begel@microsoft.com
Christian Bird, Microsoft Research, Redmond, WA USA, cbird@microsoft.com
Robert DeLine, Microsoft Research, Redmond, WA USA, rdeline@microsoft.com
Harald Gall, University of Zurich, Zurich, Switzerland, gall@ifi.uzh.ch
Ece Kamar, Microsoft Research, Redmond, WA USA, eckamar@microsoft.com
Nachiappan Nagappan, Microsoft Research, Redmond, WA USA, nachin@microsoft.com
Besmira Nushi, Microsoft Research, Redmond, WA USA, besmira.nushi@microsoft.com
Thomas Zimmermann, Microsoft Research, Redmond, WA USA, tzimmer@microsoft.com

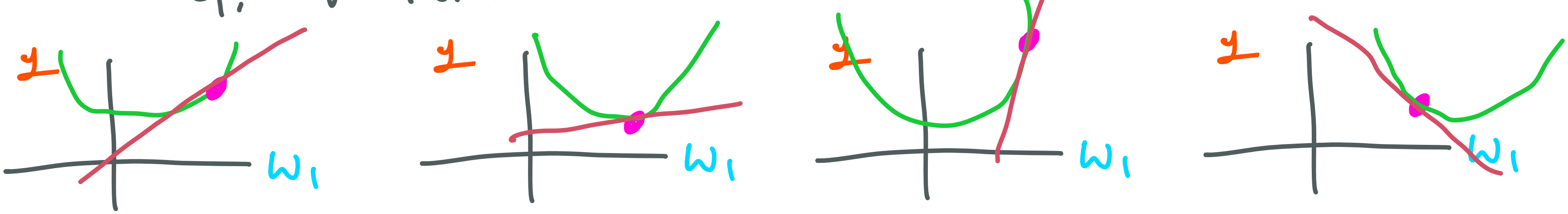


(Batch) Gradient Descent

total_epochs = 16

for epoch in range(...):

1. compute gradients w.r.t. all examples
2. average the gradients
3. update parameters
4. validate the results



Stochastic Gradient Descent

for each epoch

shuffle data

for each example

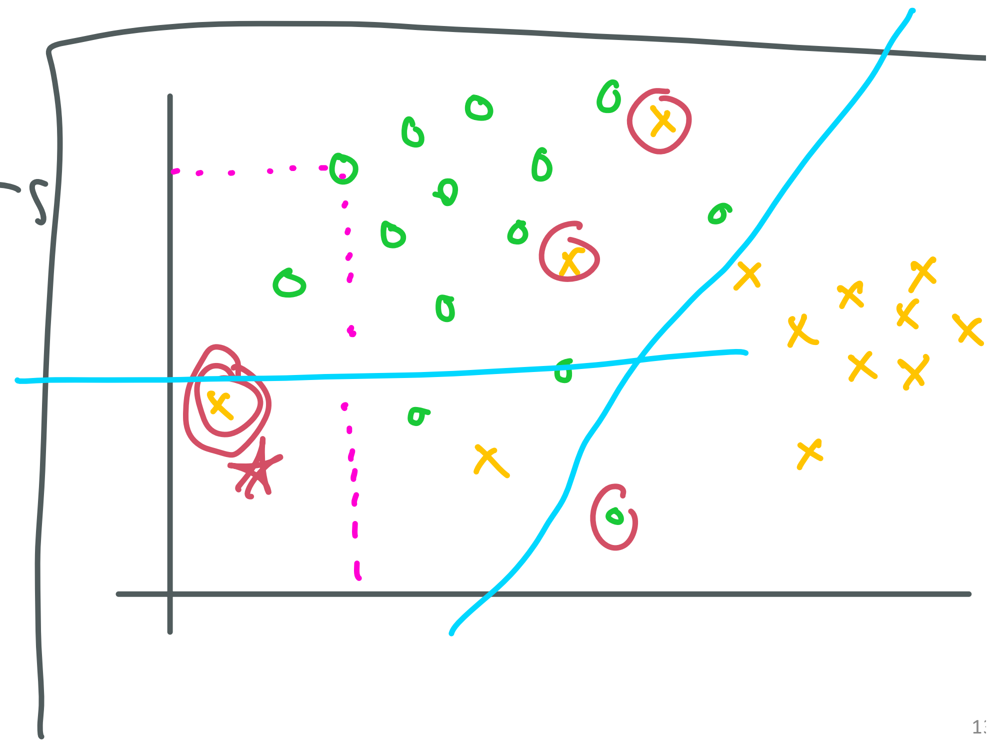
1. compute gradients

2. update parameters

4. validate

1 1 7 2

2 4 4



Minibatch Stochastic Gradient Descent

for each epoch

create shuffled batches

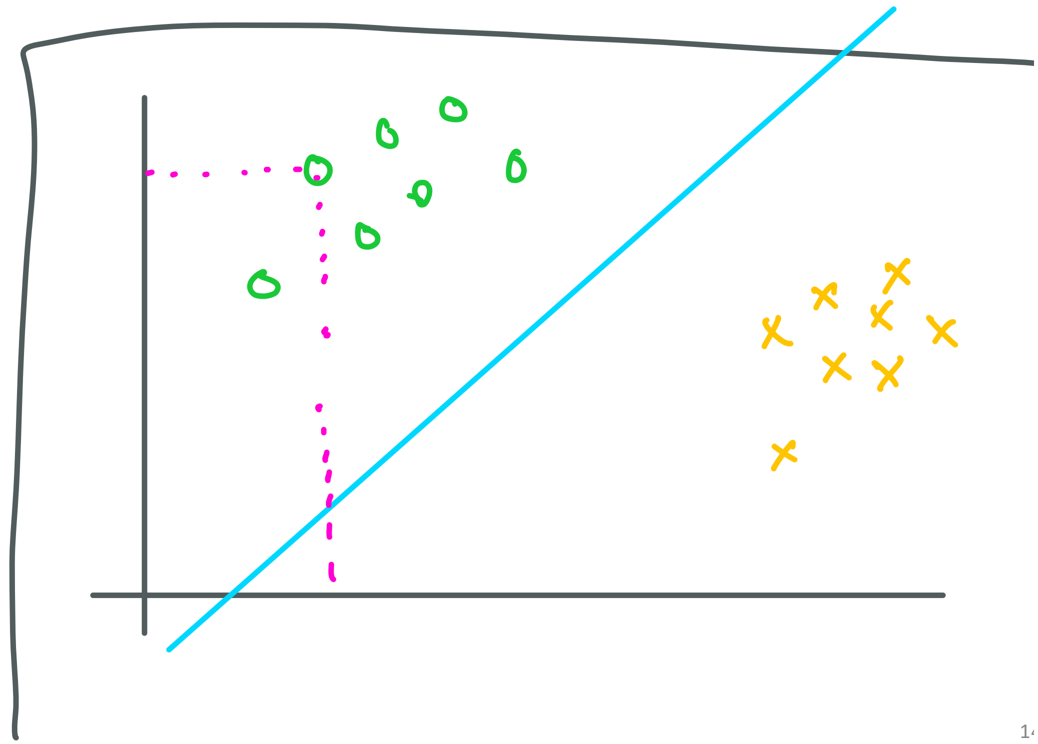
for each batch

1. compute grads
2. average grads
3. update params

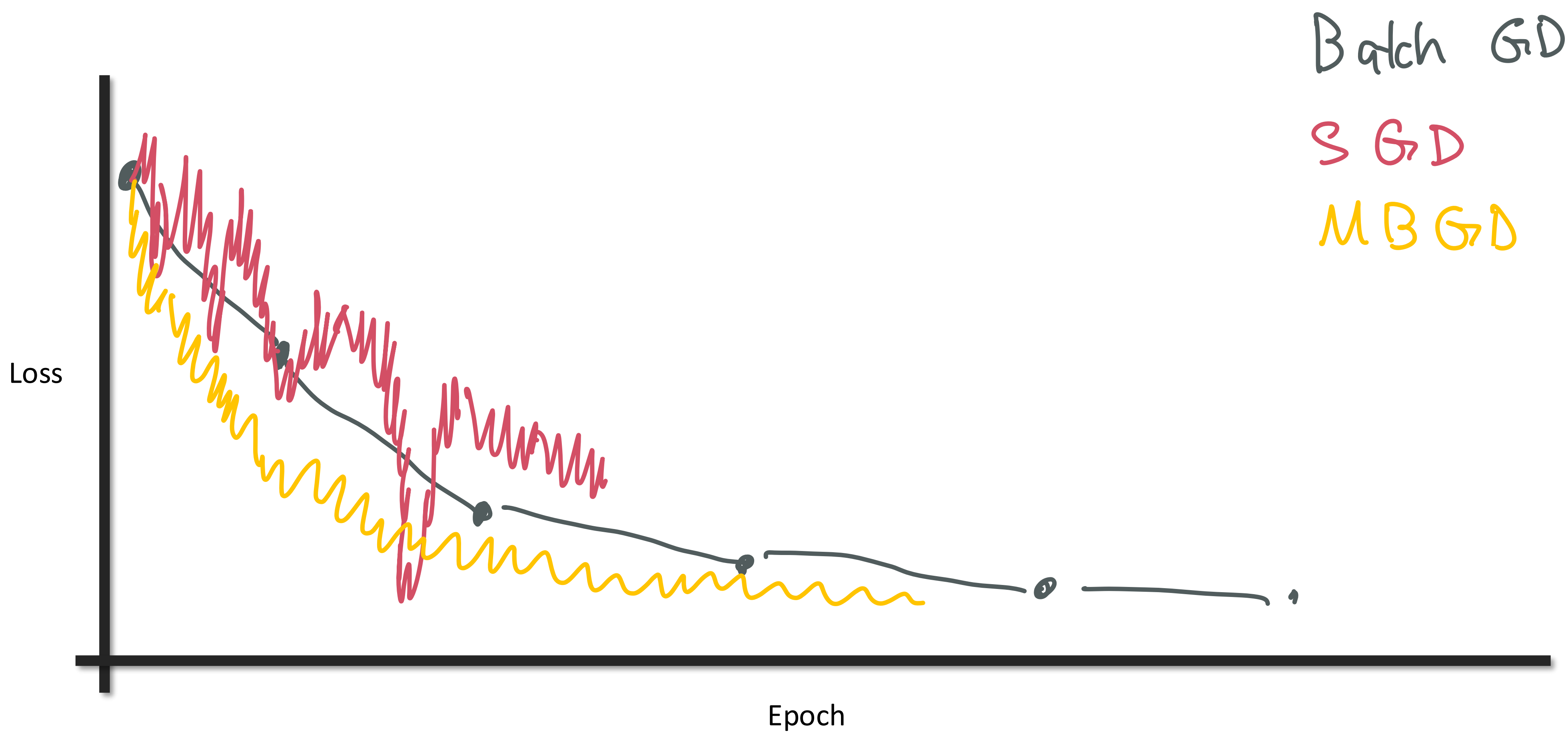
4. validate

batch_size

32, 64



Typical Behavior



Tradeoffs

- (Batch) gradient descent
  - Fewer updates means slower progress in terms of real time
  - Maximally leverages parallel computations (faster)

Tradeoffs

- (Batch) gradient descent
  - Fewer updates means slower progress in terms of real time
  - Maximally leverages parallel computations (faster)

- Stochastic gradient descent
  - Many updates can sometimes lead to faster convergence
  - Noisy data means that some updates are bad

Tradeoffs

- (Batch) gradient descent
  - Fewer updates means slower progress in terms of real time
  - Maximally leverages parallel computations (faster)

- Stochastic gradient descent
  - Many updates can sometimes lead to faster convergence
  - Noisy data means that some updates are bad

- Minibatch stochastic gradient descent
  - Noisy but less than SGD
  - Pretty fast but slower than batch gradient descent
  - Should always be your default choice
  - Add another hyperparameter called `batch_size`

Code Demo

https://github.com/anthonyjclark/cs152spring25/blob/main/Lectures/06-MinibatchSgd/SgdDemo.ipynb