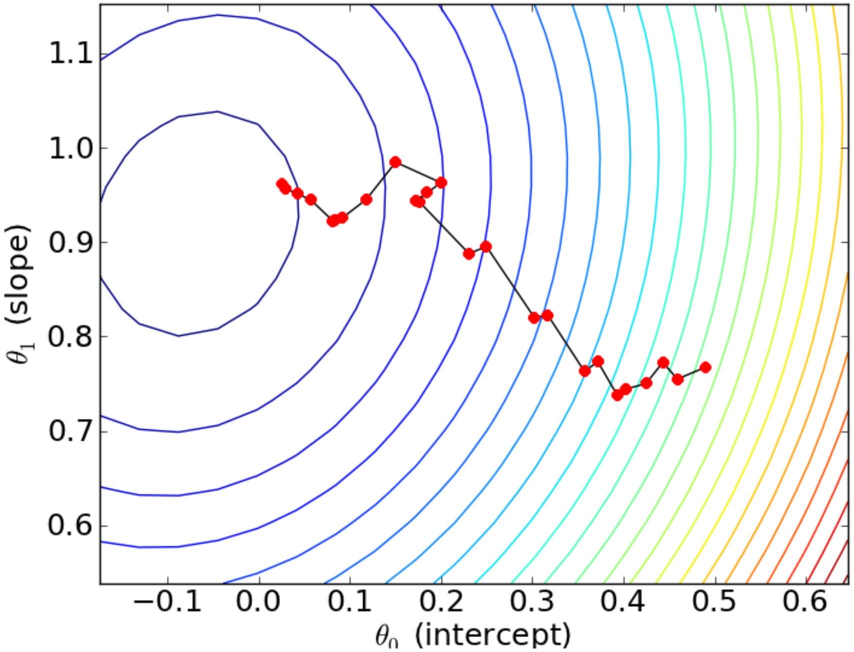
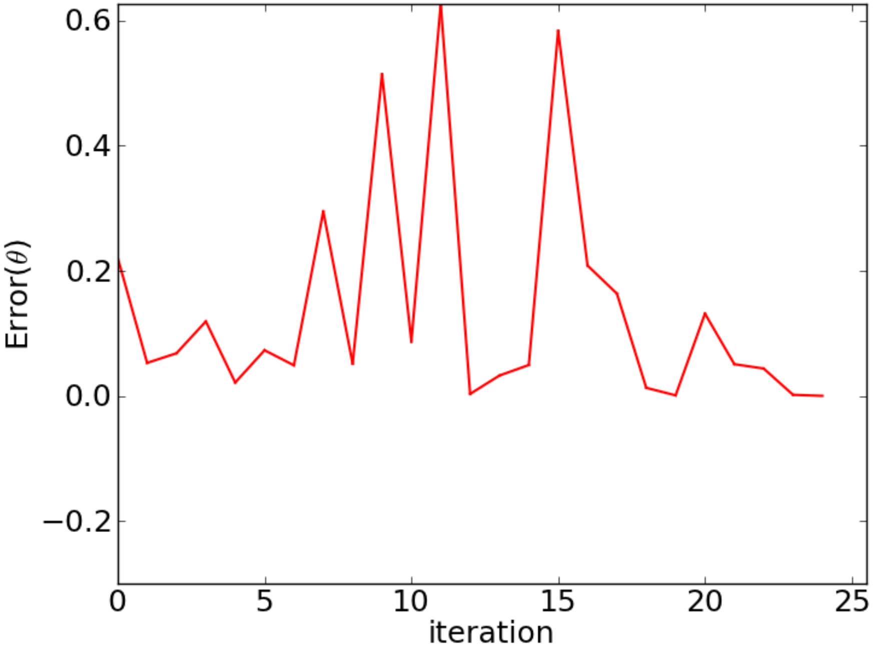
Back to Gradient Descent

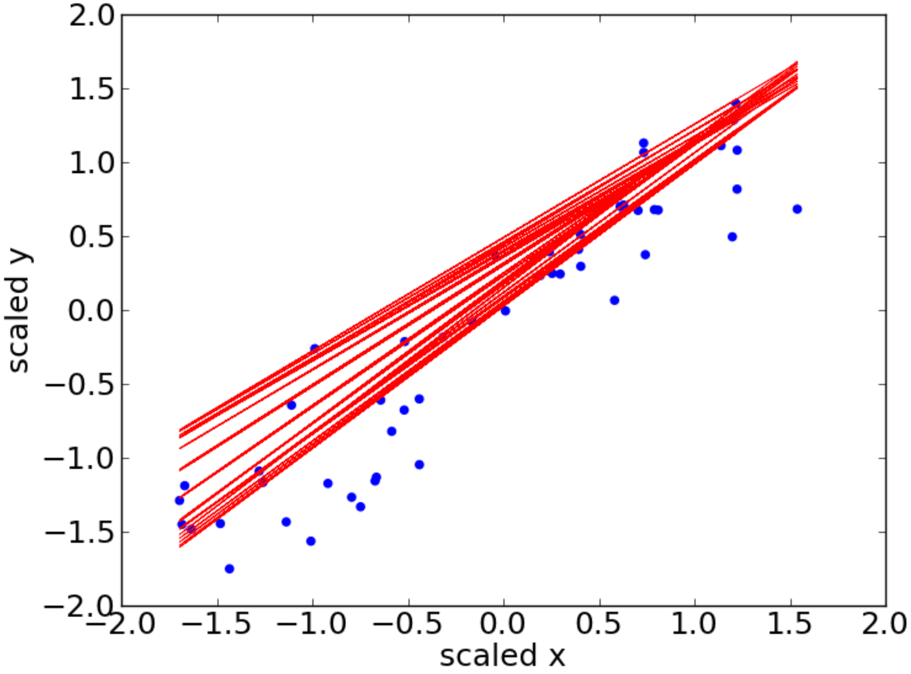
- We process the entire dataset on every iteration

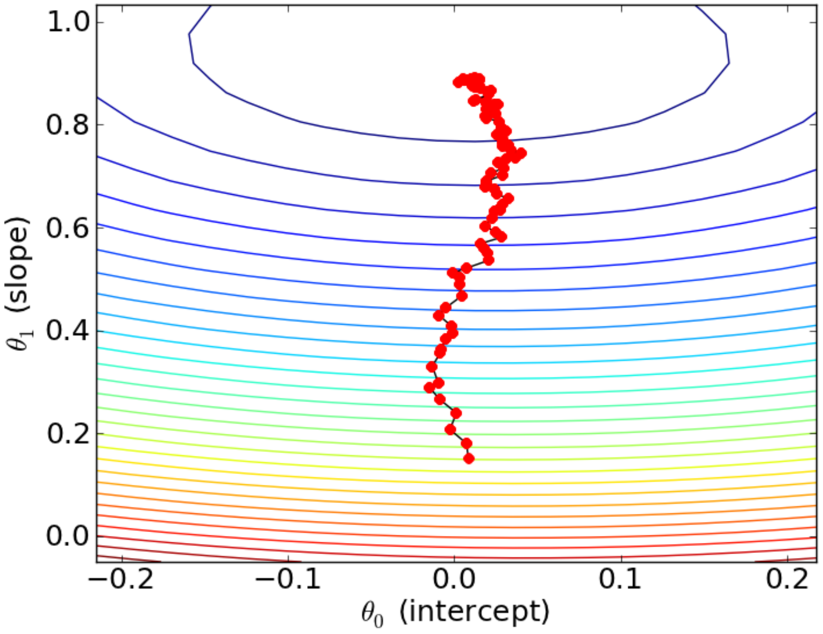
- Stochastic Gradient Descent
  - At each step, pick one random data point
  - Continue as if your entire dataset was just the one point
- Minibatch Gradient Descent
  - At each step, pick a small subset of data points
  - Continue as if your entire dataset was just this subset

Example using Stochastic Gradient Descent

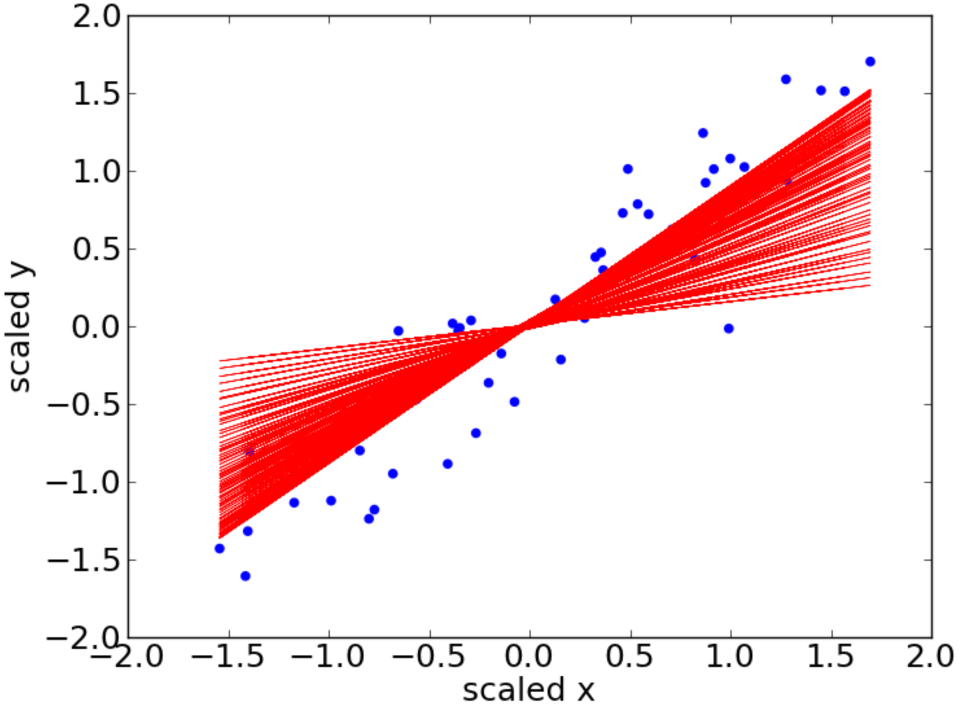


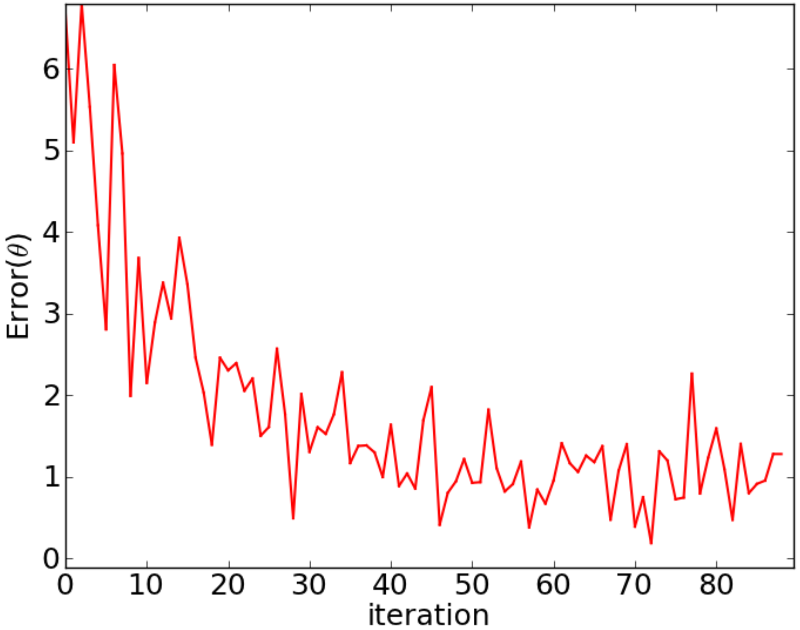






Example using Minibatches





Niu 2012

Parallelizing Stochastic Gradient Descent

- Stochastic Gradient Descent
  - At each step, pick a random data point
  - Continue as if your entire dataset was just the one point
- Parallel Stochastic Gradient Descent
  - In each of *k* threads, pick a random data point
  - Compute the gradient and update the weights
  - Weights will be “mixed”

HOGWILD!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent, Feng Niu, Benjamin Recht, Christopher Re, Stephen J. Wright, NIPS 2011 thread 0

thread 1

$$\theta_0^{(1)} \leftarrow \theta_0^{(0)} + (\theta_0^{(0)} + \theta_1^{(0)} x_3 - y_3)$$

$$\theta_0^{(2)} \leftarrow \theta_0^{(1)} + (\theta_0^{(1)} + \theta_1^{(0)} x_8 - y_8)$$

$$\theta_1^{(1)} \leftarrow \theta_1^{(0)} + (\theta_0^{(2)} + \theta_1^{(0)} x_3 - y_3) x_3$$

$$\theta_1^{(2)} \leftarrow \theta_1^{(1)} + (\theta_0^{(2)} + \theta_1^{(1)} x_8 - y_8) x_8$$

$$\theta_0^{(3)} \leftarrow \theta_0^{(2)} + (\theta_0^{(2)} + \theta_1^{(2)} x_5 - y_5)$$

$$\theta_0^{(4)} \leftarrow \theta_0^{(3)} + (\theta_0^{(3)} + \theta_1^{(2)} x_9 - y_9)$$

$$\theta_1^{(3)} \leftarrow \theta_1^{(2)} + (\theta_0^{(4)} + \theta_1^{(2)} x_5 - y_5) x_5$$

$$\theta_1^{(4)} \leftarrow \theta_1^{(3)} + (\theta_0^{(4)} + \theta_1^{(3)} x_9 - y_9) x_9$$