

Lecture 2

182/282A Deep Learning

Optimization

- ① Gradient Descent
- ② Implicit Regularization
- ③ Stochastic Gradient Descent.

Admin

- ① Discussion Room expansion
- ② Quiz dates posted today
- ③ HWO due today.

Least Squares + Gradient Descent

$$X \in \mathbb{R}^{n \times d}, \vec{y} \in \mathbb{R}^n, \vec{w} \in \mathbb{R}^d$$

$$L(w) = \|X\vec{w} - \vec{y}\|_2^2$$

$$\nabla L(\vec{w}) = 2X^T(X\vec{w} - \vec{y})$$

Gradient Descent:

$$\vec{w}_{t+1} = \underbrace{(I - 2\eta X^T X)}_{\text{Eigenvectors determine convergence}} \vec{w}_t + 2\eta X^T \vec{y}$$

Eigenvectors determine convergence.

$$|1 - 2\eta \lambda_{\max}| < 1 \quad \Rightarrow \quad -1 < 1 - 2\eta \lambda_{\max} < 1$$

$$\Rightarrow \eta < \frac{1}{\lambda_{\max}}$$

Convergence condition. $\eta < \frac{1}{\lambda_{\min}}$

$$\lambda_{\max}(X^T X) = \sigma_{\max}(X)^2$$

$$\vec{w}_{t+1} = (I - 2\eta X^T X) \vec{w}_t + \underbrace{2\eta X^T \vec{y}}_{\text{Updates only in the span of the training data.}}$$

Updates only in the span of the training data.

$R(X^T)$

$\Rightarrow$ Cannot move in $N(X)$

Fundamental thm.

$$N(X) \oplus R(X^T) = \mathbb{R}^n$$

$\Rightarrow$ Initialization matters.

Ridge Regression

$$L(\vec{w}) = \|X\vec{w} - \vec{y}\|_2^2 + \lambda \|\vec{w}\|_2^2$$

$$\vec{w}_* = \underbrace{(X^T X + \lambda I)^{-1}}_{\text{Shift improves condition number of } X^T X} X^T \vec{y}$$

Shift improves condition number of $X^T X$.

$\rightarrow$ Helps with overfitting.

$$\vec{w}_* = (X^T X + \lambda I)^{-1} X^T \vec{y} = X^T (X X^T + \lambda I)^{-1} \vec{y}$$

$$X = U \Sigma V^T$$

$$\vec{w}_* = (V \Sigma^T \cancel{U^T U} \Sigma V^T + \lambda \underbrace{I}_{V V^T})^{-1} V \Sigma^T U^T \vec{y}$$

$$= V (\Sigma^T \Sigma + \lambda I)^{-1} \cancel{V^T V} \Sigma^T U^T \vec{y}$$

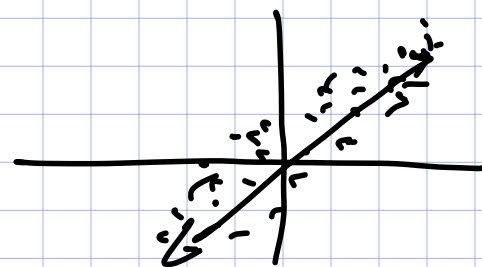
$$= V (\underbrace{\Sigma^T \Sigma + \lambda I}_{\text{diagonal}})^{-1} \Sigma^T U^T \vec{y}$$

$$= V \begin{bmatrix} \sigma_1^2 + \lambda & & 0 \\ & \sigma_2^2 + \lambda & \dots \\ 0 & \dots & \lambda \end{bmatrix}^{-1} \Sigma^T U^T \vec{y}$$

$$\vec{w}_* = \sum_{i=1}^n \vec{u}_i \left[\frac{\sigma_i}{\sigma_i^2 + \lambda} \right] \vec{u}_i^T \cdot \vec{y}$$

$$\frac{\sigma_i}{\sigma_i^2 + \lambda} \text{ if } \lambda \gg \sigma_i \rightarrow 0 \text{ as } \lambda \rightarrow \infty$$

$$\frac{\sigma_i}{\sigma_i^2 + \lambda} \rightarrow \frac{1}{\sigma_i} \text{ if } \sigma_i \gg \lambda$$



GD update:
No regularization

$$\vec{w}_{t+1} = \vec{w}_t - 2\eta \underbrace{x^T (x\vec{w}_t - \vec{y})}_{\text{residual}}$$

always
remains in
the span of
the data.

Implicit Regularization

$$V^T \vec{w}_{t+1} = V^T \vec{w}_t - 2\eta V^T (X \Sigma^T U^T) (X \vec{w}_t - \vec{y})$$

$$\vec{\tilde{w}}_{t+1} = \vec{\tilde{w}}_t - 2\eta \Sigma^T (U^T \cancel{U} \Sigma V^T \vec{w}_t - U^T \vec{y})$$

$$= \vec{\tilde{w}}_t - 2\eta \Sigma^T (\Sigma \vec{\tilde{w}}_t - U^T \vec{y})$$

$$\tilde{w}_{t+1,i} = \tilde{w}_{t,i} - 2\eta \cdot [\sigma_i^2 \tilde{w}_{t,i} - \sigma_i \tilde{y}_i]$$

$$= \tilde{w}_{t,i} - 2\eta \cdot \sigma_i [\sigma_i \tilde{w}_{t,i} - \tilde{y}_i]$$

"Early Stopping"

Implicit Regularization.

Ridge $\rightarrow \lambda$ is large $\rightarrow$

GD with early stopping $\rightarrow$

Stochastic Gradient Descent

- GD is great for convex functions.
- Neural network fitting requires optimization over non-convex landscapes
- GD is also expensive.

SGD

- Adds some noise to gradient direction. In expectation, still descends.

$$\min_{\theta} f(\theta)$$

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{1}{B} \sum_{i=1}^B \nabla f_i(\theta_t)$$

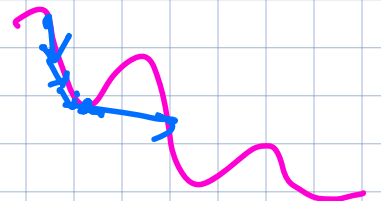
B: batch size
"mini-batch"

$$\text{where } f(\theta) = \frac{1}{n} \sum_{i=1}^n f_i(\theta)$$

Gradient descent is
"full-batch"

$$E[\nabla f_i(\theta)] = \frac{1}{n} \cdot \sum_{i=1}^n \nabla f_i(\theta) = \nabla f(\theta)$$

draw i uniformly iid



- Many standard loss functions have this form.

e.g. squared loss functions:

$$\begin{aligned} \operatorname{argmin}_{\vec{x}} \|\mathbf{A}\vec{x} - \vec{y}\|_2^2 &= \operatorname{argmin}_{\vec{x}} \frac{1}{m} \|\mathbf{A}\vec{x} - \vec{y}\|_2^2 = \\ &= \operatorname{argmin}_{\vec{x}} \frac{1}{m} \sum_{i=1}^m (\vec{q}_i^T \vec{x} - y_i)^2. \end{aligned}$$

loss on individual data point.

- Though noisy, improves fit to subset of data each step.
- Computationally less expensive
- Noise helps us (can help us) escape local minima and saddle points.

Question: Does it converge?

Optimization : Must use a learning-rate schedule with decreasing step size for SGD

In practice: can see it converge with constant ~~to~~ step size.

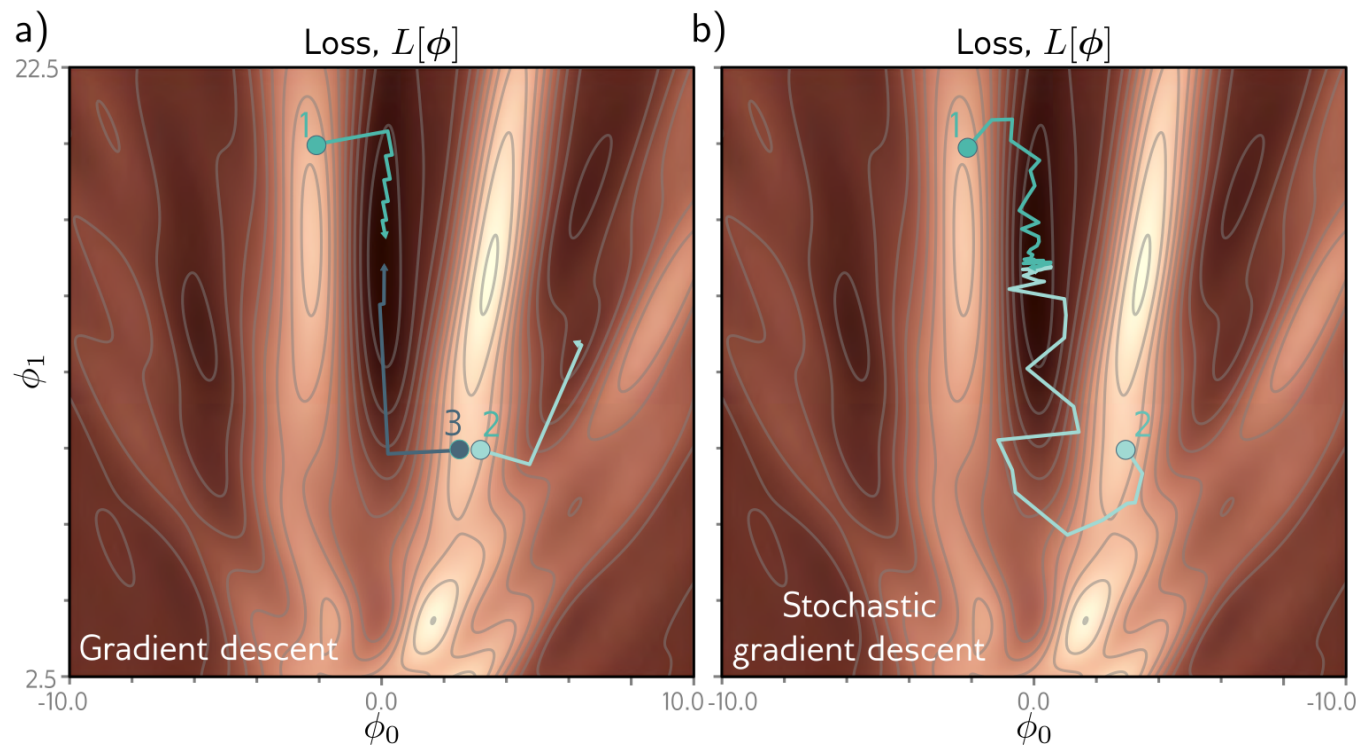


Figure 6.5 Gradient descent vs. stochastic gradient descent. a) Gradient descent with line search. As long as the gradient descent algorithm is initialized in the right “valley” of the loss function (e.g., points 1 and 3), the parameter estimate will move steadily toward the global minimum. However, if it is initialized outside this valley (e.g., point 2), it will descend toward one of the local minima. b) Stochastic gradient descent adds noise to the optimization process, so it is possible to escape from the wrong valley (e.g., point 2) and still reach the global minimum.

Prince 2025.

SGD convergence

$$X\vec{w} = \vec{y}, \quad X \in \mathbb{R}^{n \times d}, \quad d > n$$

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}$$

X : full rank.

$$\text{rank}(X) = n.$$

overparameterized regime.

Min-norm solution: $\vec{w}_* = X^T (XX^T)^{-1} \vec{y} \in \mathcal{R}(X^T)$

SGD. Batch size 1

Sample I_t is drawn at iteration t .

$$\vec{w}_0 = 0$$

$$\vec{w}_{t+1} = \vec{w}_t - 2\eta (\vec{x}_{I_t}^T \vec{w}_t - y_{I_t}) \cdot \vec{x}_{I_t} \quad \eta > 0$$

Let $\vec{q}_t = \vec{w}_t - \vec{w}_*$ Error / distance to optimal.

How does this evolve?

$$\begin{aligned} \vec{q}_{t+1} &= \vec{w}_t - 2\eta (\vec{x}_{I_t}^T \vec{w}_t - y_{I_t}) \cdot \vec{x}_{I_t} - \vec{w}_* \\ &= \vec{q}_t - 2\eta (\vec{x}_{I_t}^T (\vec{q}_t + \vec{w}_*) - y_{I_t}) \vec{x}_{I_t} \end{aligned}$$

$$= \vec{q}_t - 2\eta (\vec{x}_{I_t}^\top \vec{q}_t) \vec{x}_{I_t}.$$

$$[\vec{x}_{I_t}^\top \vec{w}_* = y_{I_t}]$$

$$\vec{q}_{t+1} = (\mathbf{I} - 2\eta \vec{x}_{I_t} \vec{x}_{I_t}^\top) \vec{q}_t$$

① Note $\vec{q}_t \in \mathcal{R}(X^\top) \cdot \forall t$ why?

$$\textcircled{2} \quad \|\mathbf{X}\vec{q}\|^2 \geq \sigma_{\min}^2 \|\vec{q}\|^2.$$

Why is $\vec{q} \in R(\mathbf{X}^T)$ important?

Lemma: Let $\rho = \max_{1 \leq i \leq n} \|\vec{x}_i\|^2$ be max row norm of X

Let $\alpha = 1 - \frac{4\eta(1-\eta\rho^2)\sigma_{\min}^2}{n}$. Let $0 < \eta < \frac{1}{\rho^2}$.

Then: $\mathbb{E}[\|\vec{q}_{t+1}\|^2 \mid \vec{q}_t] \leq \alpha \|\vec{q}_t\|^2$

Proof:

$$\begin{aligned} \textcircled{1} \quad \|\vec{q}_{t+1}\|^2 &= \|\vec{q}_t - 2\eta(\vec{x}_{I_t}^\top \vec{q}_t)\vec{x}_{I_t}\|^2 \\ &= \|\vec{q}_t\|^2 + 4\eta^2 \|\vec{x}_{I_t}\|^2 (\vec{x}_{I_t}^\top \vec{q}_t)^2 - 4\eta (\vec{x}_{I_t}^\top \vec{q}_t)^2 \\ &= \|\vec{q}_t\|^2 - 4\eta(1 - \eta \|\vec{x}_{I_t}\|^2) (\vec{x}_{I_t}^\top \vec{q}_t)^2 \end{aligned}$$

② Want a lower bound on: $4\eta(1 - \eta \|\vec{x}_{I_t}\|^2) (\vec{x}_{I_t}^\top \vec{q}_t)^2$

$$\eta \|\vec{x}_{I_t}\|^2 \leq \eta \cdot \rho^2 < 1 \quad \leftarrow \text{Key assumption.}$$

$$\Rightarrow 1 - \eta \|\vec{x}_{I_t}\|^2 \geq 1 - \eta \rho^2 > 0$$

$$\Rightarrow \|\vec{q}_{t+1}\|^2 \leq \|\vec{q}_t\|^2 - 4\eta(1-\eta\sigma^2)(\vec{x}_{I_t}^\top \vec{q}_t)^2$$

③ Condition on $\vec{q}_t \rightarrow$

$$\begin{aligned} \mathbb{E}[\|\vec{q}_{t+1}\|^2 | \vec{q}_t] &\leq \|\vec{q}_t\|^2 - 4\eta(1-\eta\sigma^2) \mathbb{E}[(\vec{x}_{I_t}^\top \vec{q}_t)^2 | \vec{q}_t] \\ &= \|\vec{q}_t\|^2 - 4\eta(1-\eta\sigma^2) \frac{1}{n} \sum_{i=1}^n (\vec{x}_i^\top \vec{q}_t)^2 \\ &= \|\vec{q}_t\|^2 - 4\eta(1-\eta\sigma^2) \frac{1}{n} \|\mathbf{X} \vec{q}_t\|^2. \end{aligned}$$

④ But now $\|\mathbf{X} \vec{q}_t\|^2 \dots$ what do we know?

$$\|\mathbf{X} \vec{q}_t\|^2 \geq$$

$$\mathbb{E}[\|\vec{q}_{t+1}\|^2 | \vec{q}_t] \leq \alpha \|\vec{q}_t\|^2$$

Which gives us what we wanted!

Recall $\alpha = 1 - \frac{4\eta(1-\eta s^2)\sigma_{\min}^2}{n}$

$$\alpha \in [0, 1)$$

$\alpha < 1$: All subtracted quantities are positive.

$\alpha \geq 0$: $\eta(1-\eta s^2) \rightarrow$ maximized at $\eta = \frac{1}{2s^2}$. Neg. leading bef.

Max value: $\frac{1}{4s^2}$

Also: $\sigma_{\min}^2 = \lambda_{\min}(XX^T) \leq \vec{e}_i^T XX^T \vec{e}_i = \|\vec{x}_i\|^2 \leq s^2.$

$$\Rightarrow \frac{4\eta(1-\eta s^2) \cdot \sigma_{\min}^2}{n} \leq \frac{4 \cdot \frac{1}{4s^2} \cdot s^2}{n} = \frac{1}{n} \leq 1.$$