Step Size Matters in Deep Learning

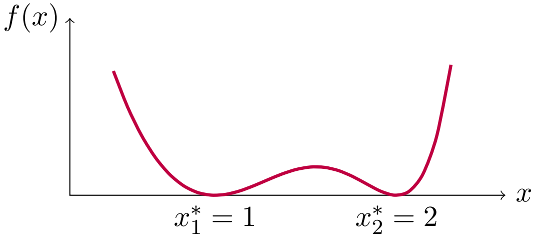
Kamil Nar   Shankar Sastry

Neural Information Processing Systems
December 4, 2018

Gradient Descent: Effect of Step Size

Example

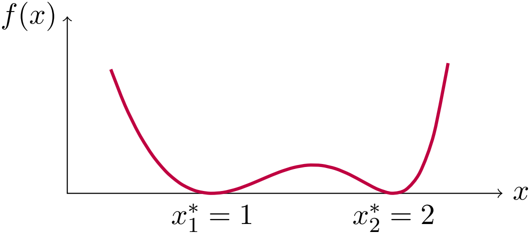
$$\min_{x\in\mathbb{R}}\ (x^2+1)(x-1)^2(x-2)^2$$



Gradient Descent: Effect of Step Size

Example

$$\min_{x\in\mathbb{R}} \ (x^2+1)(x-1)^2(x-2)^2$$



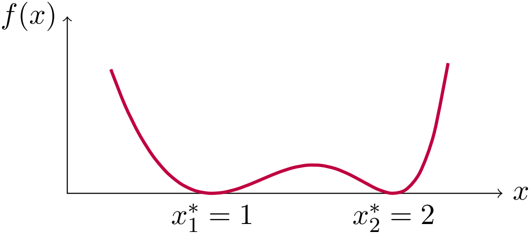
From random initialization

- converges to x_1^* only if $\delta \leq 0.5$
- converges to x_2^* only if $\delta \leq 0.2$

Gradient Descent: Effect of Step Size

Example

$$\min_{x\in\mathbb{R}}\ (x^2+1)(x-1)^2(x-2)^2$$



From random initialization

- converges to x_1^* only if $\delta \leq 0.5$
- converges to x_2^* only if $\delta \leq 0.2$

If the algorithm converges with $\delta = 0.3$, the solution is x_1^*.

Deep Linear Networks

$$x \;\mapsto\; W_L W_{L-1} \cdots W_2 W_1 x$$

Deep Linear Networks

$$x \mapsto W_L W_{L-1} \cdots W_2 W_1 x$$

- Cost function has infinitely many local minimum
- Different dynamic characteristics at different optima

Lyapunov Stability of Gradient Descent

Deep Linear Networks

Proposition

- $\lambda \in \mathbb{R}$ and $\lambda \neq 0$
- λ is estimated as multiplication of scalar parameters $\{w_i\}$

$$\min_{\{w_i\}} \; \frac{1}{2} \left(w_L \dots w_2 w_1 - \lambda \right)^2 .$$

Lyapunov Stability of Gradient Descent

Deep Linear Networks

Proposition

- $\lambda \in \mathbb{R}$ and $\lambda \neq 0$
- λ is estimated as multiplication of scalar parameters $\{w_i\}$

$$\min_{\{w_i\}} \frac{1}{2} \left(w_L \dots w_2 w_1 - \lambda \right)^2 .$$

For convergence to $\{w_i^*\}$ with $w_L^* \dots w_2^* w_1^* = \lambda$, step size must satisfy

$$\delta \leq \frac{2}{\sum_{i=1}^{L} \left(\frac{\lambda}{w_i^*} \right)^2} .$$

Lyapunov Stability of Gradient Descent

Deep Linear Networks

- δ needs to be very small for equilibria with disproportionate $\{w_i^*\}$
- For each δ, the algorithm can converge only to a subset of optima

Lyapunov Stability of Gradient Descent

Deep Linear Networks

- δ needs to be very small for equilibria with disproportionate $\{w_i^*\}$
- For each δ, the algorithm can converge only to a subset of optima
- No finite Lipschitz constant for the gradient on the whole parameter space

Deep Linear Networks

Theorem

- $\{x_i\}_{i\in[N]}$ satisfies $\frac{1}{N}\sum_{i=1}^{N} x_i x_i^\top = I$
- R is estimated as multiplication of $\{W_j\}$ by

$$\min_{\{W_j\}} \frac{1}{2N}\sum_{i=1}^{N} \|Rx_i - W_L W_{L-1}\cdots W_2 W_1 x_i\|_2^2$$

Deep Linear Networks

Theorem

- $\{x_i\}_{i\in[N]}$ satisfies $\frac{1}{N}\sum_{i=1}^{N} x_i x_i^\top = I$
- R is estimated as multiplication of $\{W_j\}$ by

$$\min_{\{W_j\}} \frac{1}{2N}\sum_{i=1}^{N} \|Rx_i - W_L W_{L-1}\cdots W_2 W_1 x_i\|_2^2$$

Assume the gradient descent algorithm with random initialization has converged to $\hat{R}$. Then,

$$\rho(\hat{R}) \leq \left(\frac{2}{L\delta}\right)^{L/(2L-2)} \qquad \text{almost surely.}$$

Deep Linear Networks

Theorem

- $\{x_i\}_{i\in[N]}$ satisfies $\frac{1}{N}\sum_{i=1}^{N} x_i x_i^\top = I$
- R is estimated as multiplication of $\{W_j\}$ by

$$\min_{\{W_j\}} \frac{1}{2N}\sum_{i=1}^{N} \|Rx_i - W_L W_{L-1}\cdots W_2 W_1 x_i\|_2^2$$

Assume the gradient descent algorithm with random initialization has converged to $\hat{R}$. Then,

$$\rho(\hat{R}) \leq \left(\frac{2}{L\delta}\right)^{L/(2L-2)} \qquad \text{almost surely.}$$

- Step size bounds the Lipschitz constant of the estimated function

Deep Linear Networks

Theorem

- $\{x_i\}_{i\in[N]}$ satisfies $\frac{1}{N}\sum_{i=1}^{N} x_i x_i^\top = I$
- R is estimated as multiplication of $\{W_j\}$ by

$$\min_{\{W_j\}} \frac{1}{2N}\sum_{i=1}^{N} \|Rx_i - W_L W_{L-1}\cdots W_2 W_1 x_i\|_2^2$$

Assume the gradient descent algorithm with random initialization has converged to $\hat{R}$. Then,

$$\rho(\hat{R}) \leq \left(\frac{2}{L\delta}\right)^{L/(2L-2)} \quad \text{almost surely.}$$

- Step size bounds the Lipschitz constant of the estimated function
- Contrary to ordinary-least-squares

Deep Linear Networks

Symmetric PSD matrices:

- The bound is tight with identity initialization
- Identity initialization allows convergence with the largest step size

Nonlinear Networks

Poster #8

Two-layer ReLU network:

$$x \;\mapsto\; W(Vx - b)_+$$

Nonlinear Networks

Poster #8

Two-layer ReLU network:

$$x \ \mapsto \ W(Vx - b)_+$$

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be estimated by

$$\min_{W,V} \ \frac{1}{2} \sum\nolimits_{i=1}^{N} \|W(Vx_i - b)_+ - f(x_i)\|_2^2.$$

Nonlinear Networks

Poster #8

Two-layer ReLU network:

$$x \;\mapsto\; W(Vx - b)_+$$

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be estimated by

$$\min_{W,V} \; \frac{1}{2} \sum_{i=1}^{N} \|W(Vx_i - b)_+ - f(x_i)\|_2^2.$$

If the algorithm converges, then the estimate $\hat{f}(x_i)$ satisfies

$$\max_{i \in [N]} \; \|x_i\| \|\hat{f}(x_i)\| \leq \frac{1}{\delta}$$

almost surely.