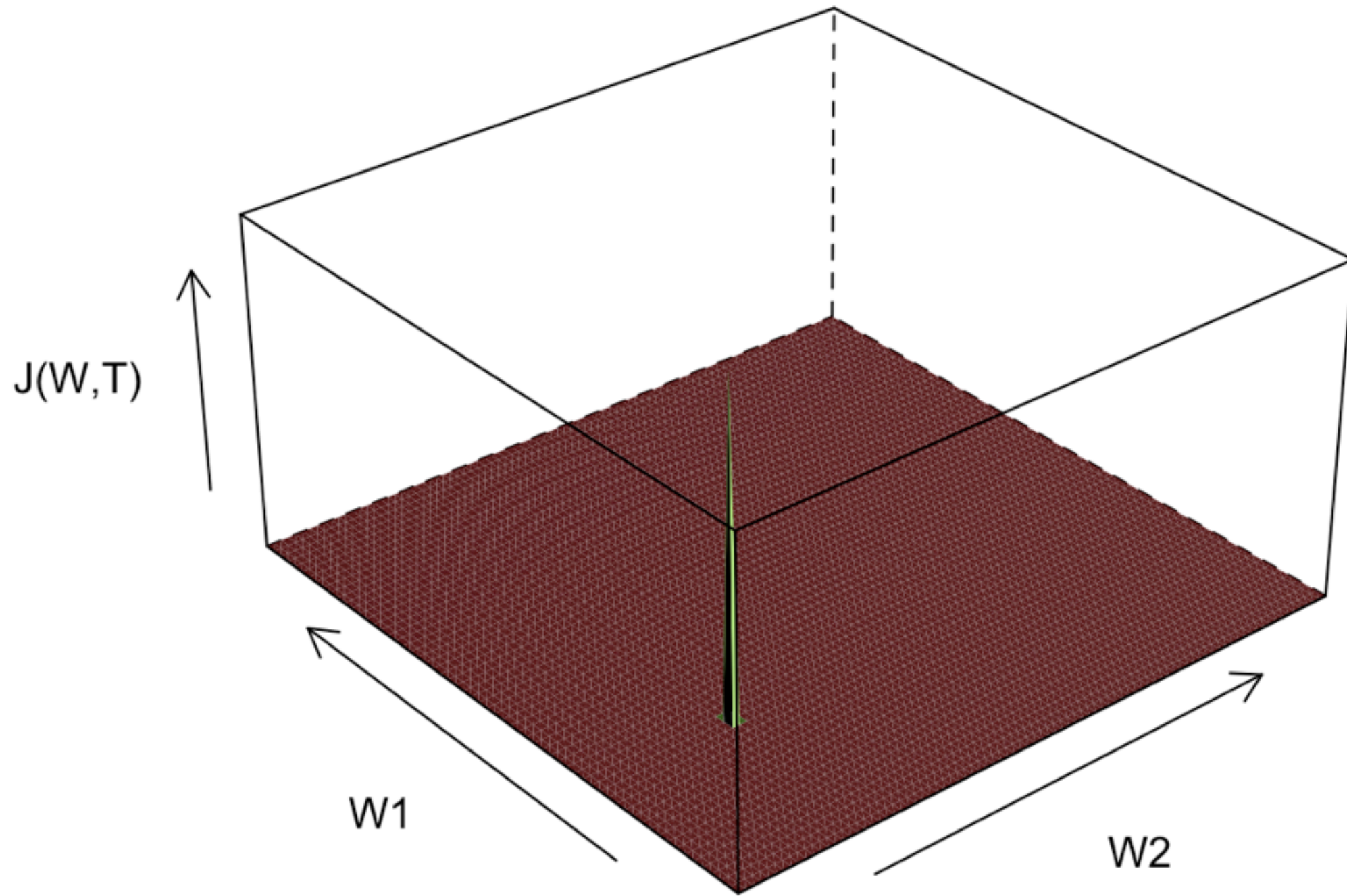
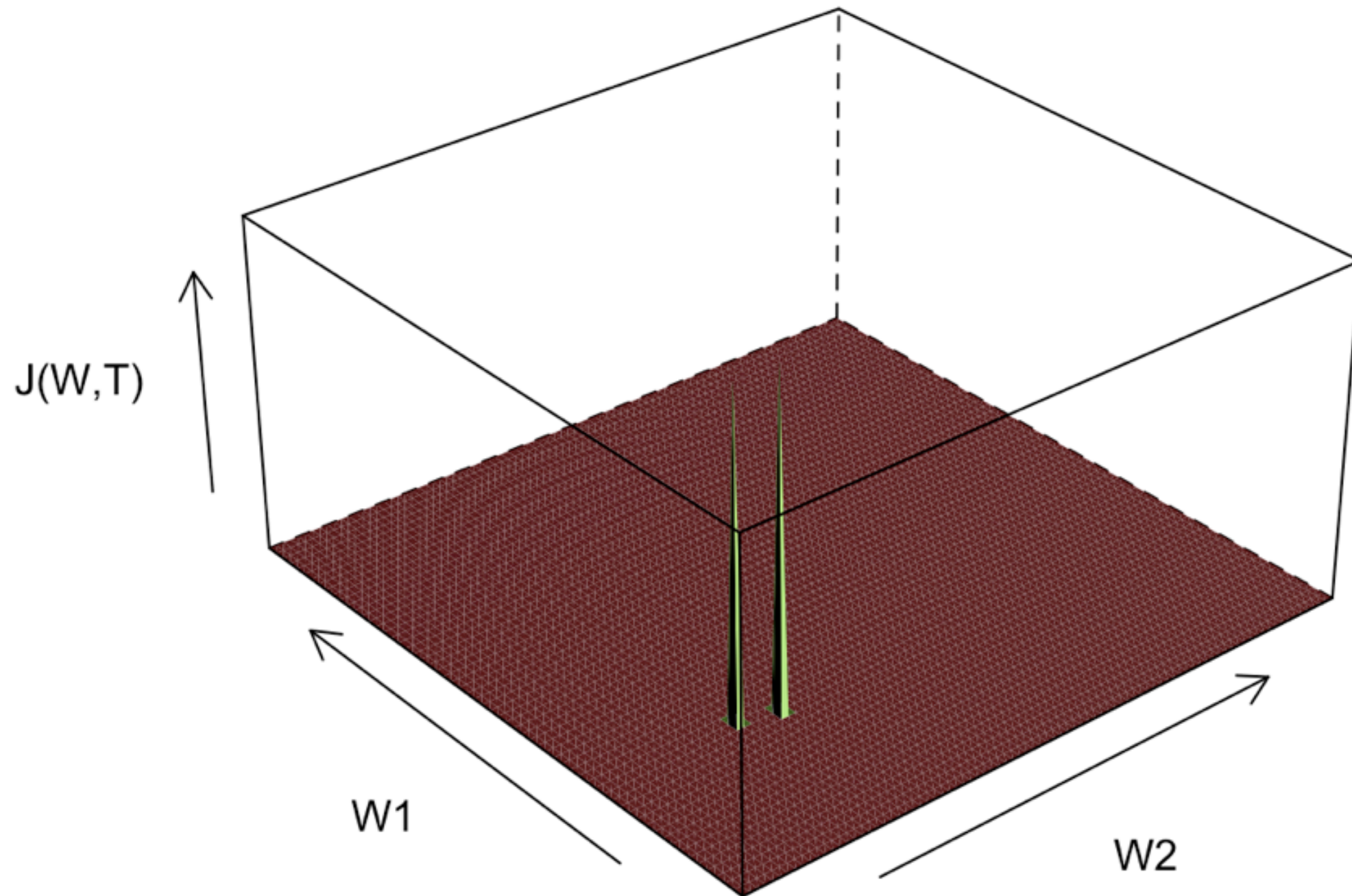
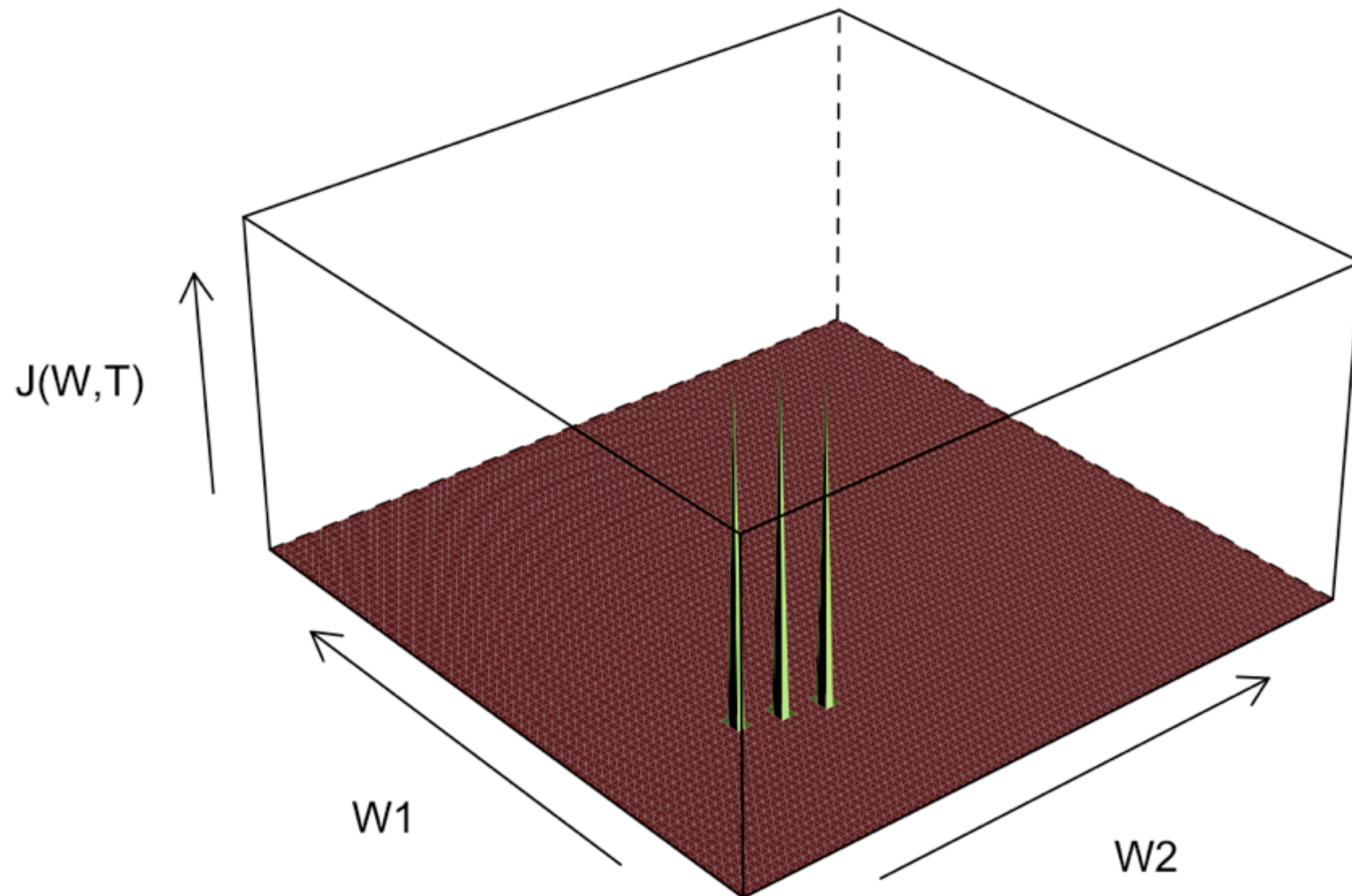
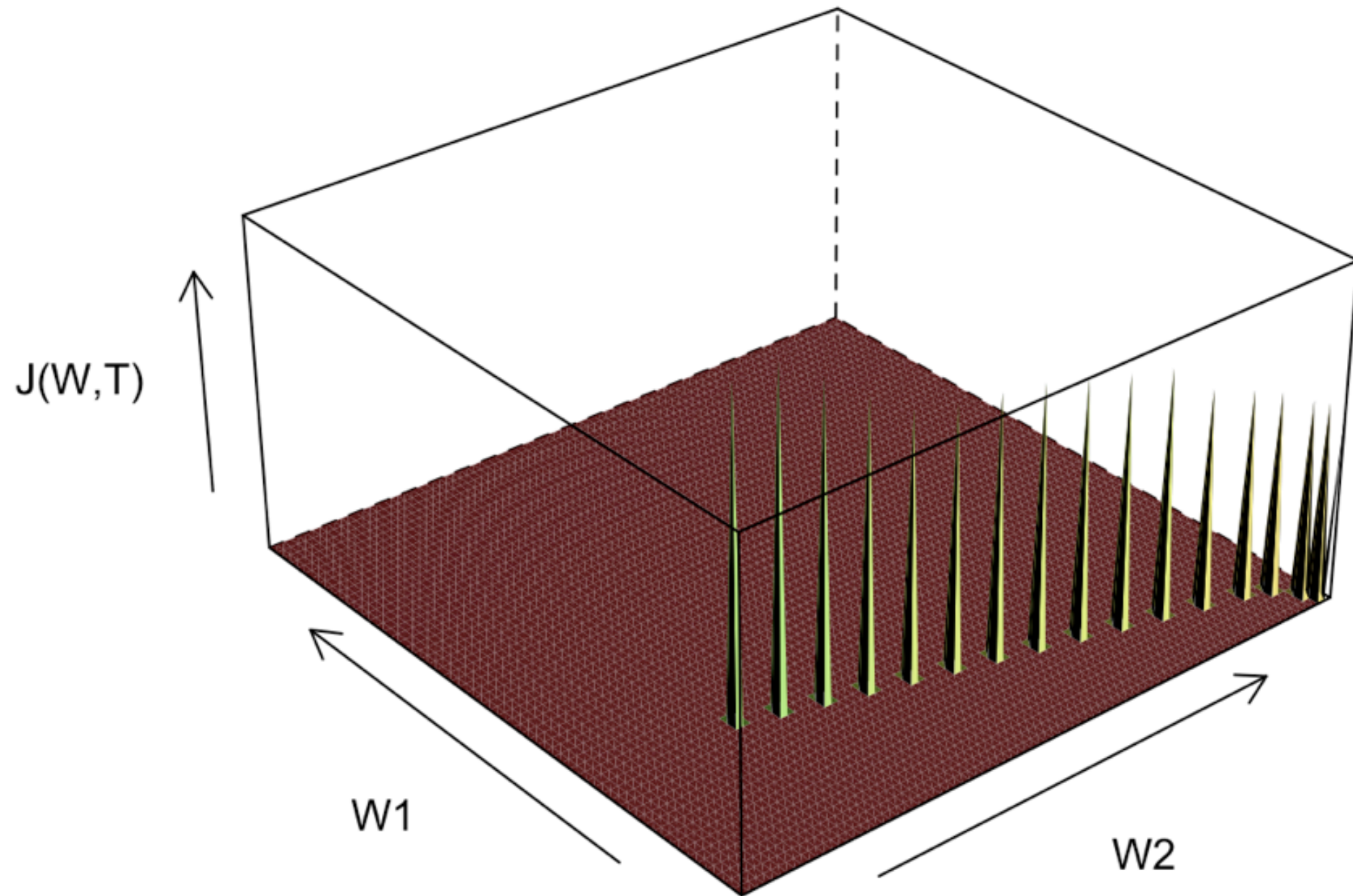


Gradient Descent Updater Strategies









Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \Delta_{\theta} J(\theta_t, X, Y)$$

Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \Delta_{\theta} J(\theta_t, X, Y)$$

Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \Delta_{\theta} J(\theta_t, X, Y)$$

Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \Delta_{\theta} J(\theta_t, X, Y)$$

Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \Delta_{\theta} J(\theta_t, X, Y)$$

Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \Delta_{\theta} J(\theta_t, X, Y)$$

Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \Delta_{\theta} J(\theta_t, X, Y)$$

Stochastic Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \Delta_{\theta} J(\theta_t, x^{(i)}, y^{(i)})$$

Stochastic Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \Delta_{\theta} J(\theta_t, x^{(i)}, y^{(i)})$$

(Mini) Batch Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \Delta_{\theta} J(\theta_t, x^{(i..i+n)}, y^{(i..i+n)})$$

(Mini) Batch Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \Delta_{\theta} J(\theta_t, x^{(i..i+n)}, y^{(i..i+n)})$$

Momentum

$$\nu_t = \gamma \nu_{t-1} - \eta \Delta_{\theta} J(\theta_t, X, Y)$$

$$\theta_{t+1} = \theta_t - \nu_t$$

Momentum

$$\nu_t = \gamma \nu_{t-1} - \eta \Delta_{\theta} J(\theta_t, X, Y)$$
$$\theta_{t+1} = \theta_t - \nu_t$$

Momentum

$$\nu_t = \gamma \nu_{t-1} - \eta \Delta_{\theta} J(\theta_t, X, Y)$$

$$\theta_{t+1} = \theta_t - \nu_t$$

Nesterov accelerated gradient

$$\begin{aligned}\nu_t &= \gamma \nu_{t-1} - \eta \Delta_{\theta} J(\theta_t - \gamma \nu_{t-1}, X, Y) \\ \theta &= \theta - \nu_t\end{aligned}$$

Nesterov accelerated gradient

$$\begin{aligned}\nu_t &= \gamma \nu_{t-1} - \eta \Delta_{\theta} J(\theta_t - \gamma \nu_{t-1}, X, Y) \\ \theta &= \theta - \nu_t\end{aligned}$$

Adagrad

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,i} + \epsilon}} \Delta_{\theta} J(\theta_{t,i}, X, Y)$$

Adagrad

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,i} + \epsilon}} \Delta_{\theta} J(\theta_{t,i}, X, Y)$$

Adagrad

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,i} + \epsilon}} \Delta_{\theta} J(\theta_{t,i}, X, Y)$$

Adagrad

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,i} + \epsilon}} \Delta_{\theta} J(\theta_{t,i}, X, Y)$$

Adagrad

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,i} + \epsilon}} \Delta_{\theta} J(\theta_{t,i}, X, Y)$$

Adagrad

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,i} + \epsilon}} \Delta_{\theta} J(\theta_{t,i}, X, Y)$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{G_t + \epsilon}} \Delta_{\theta} J(\theta_t, X, Y)$$

Adadelata

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[\Delta\theta^2]_t + \epsilon}} \Delta_{\theta} J(\theta_t, X, Y)$$

Adadelata

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[\Delta\theta^2]_t + \epsilon}} \Delta_{\theta} J(\theta_t, X, Y)$$

RMSProp

RMSProp

Adam

AdaMax

RMSProp

Adam

AdaMax

RMSProp

Adam

Nadam

Summary

Activation Functions