MACHINE LEARNING

WEEK 3: OPTIMIZATION

LECTURE 2: HAMILTONIAN, BACKPROPAGATION

Lagrange Multipliers

Minimization is often complicated by the addition of constraints. The simplest kind of constraint is an equality constraint, for example, $G(x) = 0$. Formally this addition is stated as

$$\min_x \tilde{F}(x) \text{ subject to } G(x) = 0$$

The method of Lagrange reduces the constrained problem to a new, unconstrained minimization problem with additional variables. The additional variables are known as Lagrange multipliers. To handle this problem, append $G(x)$ to the function $\tilde{F}(x)$ using a Lagrange multiplier λ:

$$F(x, \lambda) = \tilde{F}(x) + \lambda G(x)$$

The Lagrange multiplier is an extra scalar variable, so the number of degrees of freedom of the problem has increased, but the plus side is that now simple, unconstrained minimization techniques can be applied to the composite function. The problem becomes

$$\min_{x,\lambda} F(x, \lambda)$$

Dynamic Programming: Discretization step

First step: make all variables discrete:

The dynamics are now expressed by a difference equation,

$$\boldsymbol{x}(k+1) = f[\boldsymbol{x}(k), \boldsymbol{u}(k)]$$

The initial condition is:

$$\boldsymbol{x}(0) = \boldsymbol{x}_0$$

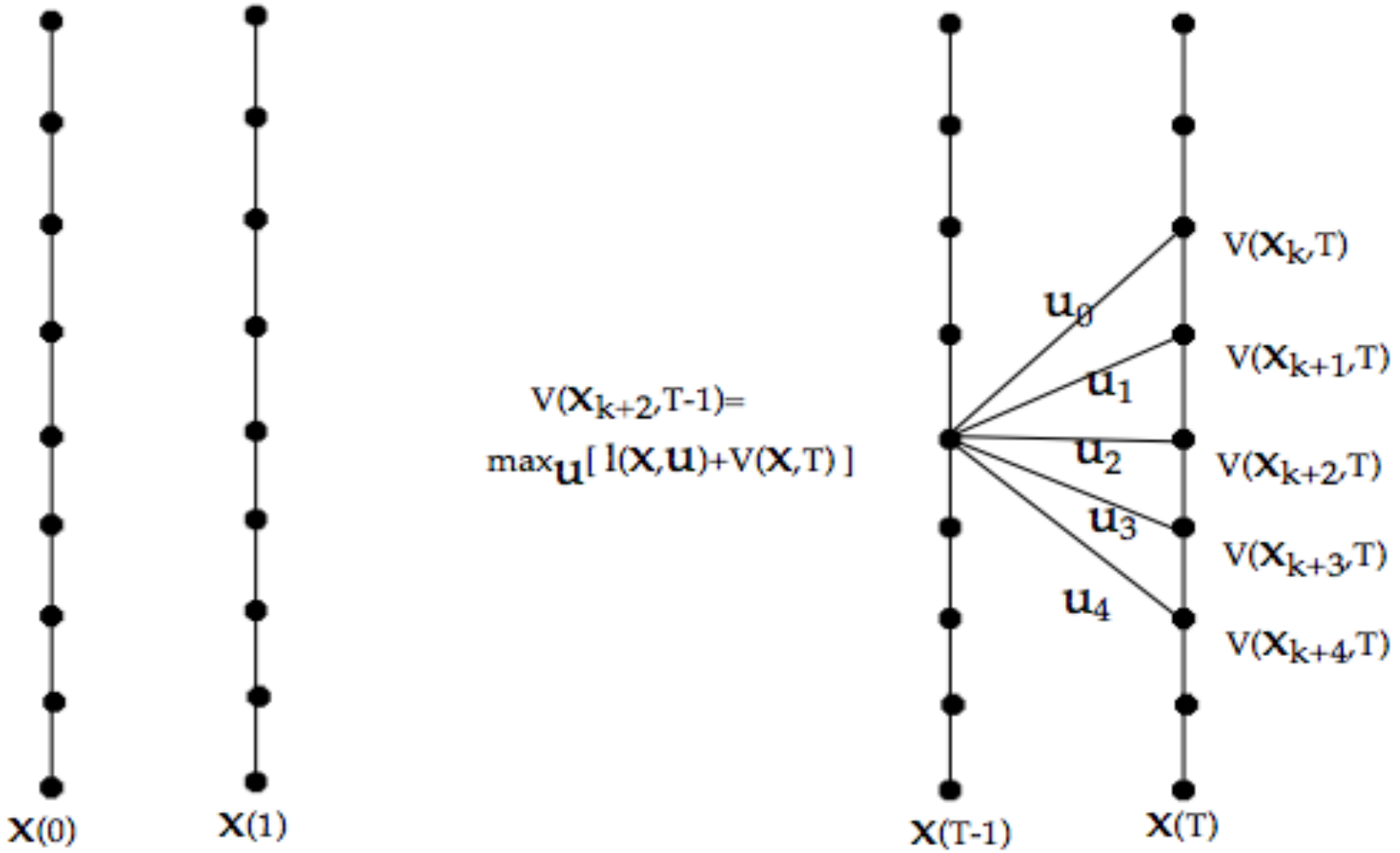
The allowable control is also discrete:

$$\boldsymbol{u}(k) \in U, k = 0, \ldots, N$$

Finally, the integral in the objective function is replaced by a sum:

$$J = \psi[\boldsymbol{x}(T)] + \sum_{0}^{N-1} \ell[\boldsymbol{u}(k), \boldsymbol{x}(k)]$$

DP direct solution: Illustration



Euler-Lagrange Method

$$\max_{\boldsymbol{u}} J \text{ subject to the constraint } \dot{\boldsymbol{x}} = \boldsymbol{F}(\boldsymbol{x}, \boldsymbol{u})$$

The strategy will be to assume that $\boldsymbol{u}$ maximizes J and then use this assumption to derive other conditions for a maximum. These arguments depend on making a perturbation in $\boldsymbol{u}$ and seeing what happens. Since $\boldsymbol{u}$ affects $\boldsymbol{x}$, the calculations become a little involved, but the argument is just a matter of careful bookkeeping. The main trick is to add additional terms to J that sum to zero. Let's start by appending the dynamic equation to J as before, but this time using continuous Lagrange multipliers $\boldsymbol{\lambda}(t)$:

$$\bar{J} = J - \int_0^T \boldsymbol{\lambda}^T [\dot{\boldsymbol{x}} - \boldsymbol{F}(\boldsymbol{x}, \boldsymbol{u})] dt$$

Hamiltonian

Anticipating what is about to happen, we define the *Hamiltonian* $H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u})$ as

$$H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u}) \equiv \boldsymbol{\lambda}^T[\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{u})] + \ell(\boldsymbol{x}, \boldsymbol{u})$$

so that the expression for $\bar{J}$ becomes

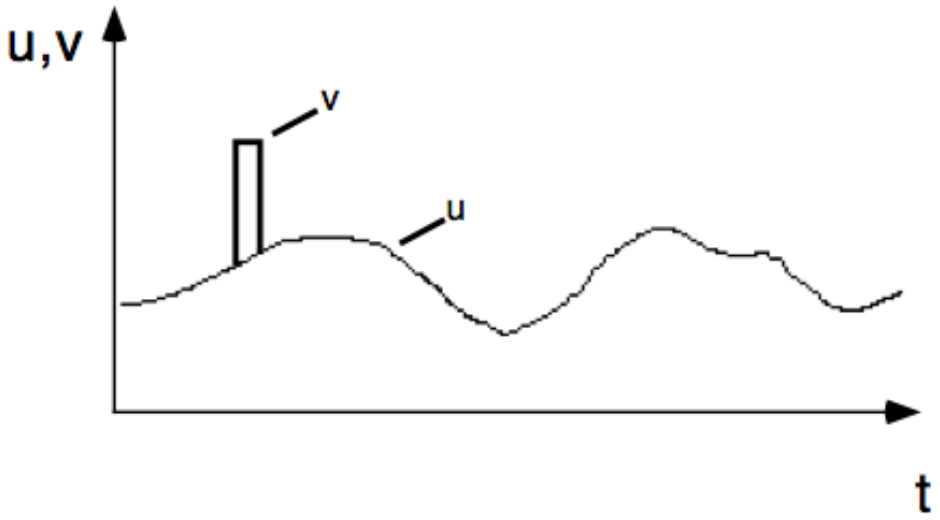
$$\bar{J} = \psi[\boldsymbol{x}(T)] + \int_0^T [H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u}) - \boldsymbol{\lambda}^T \dot{\boldsymbol{x}}]dt$$

δJ

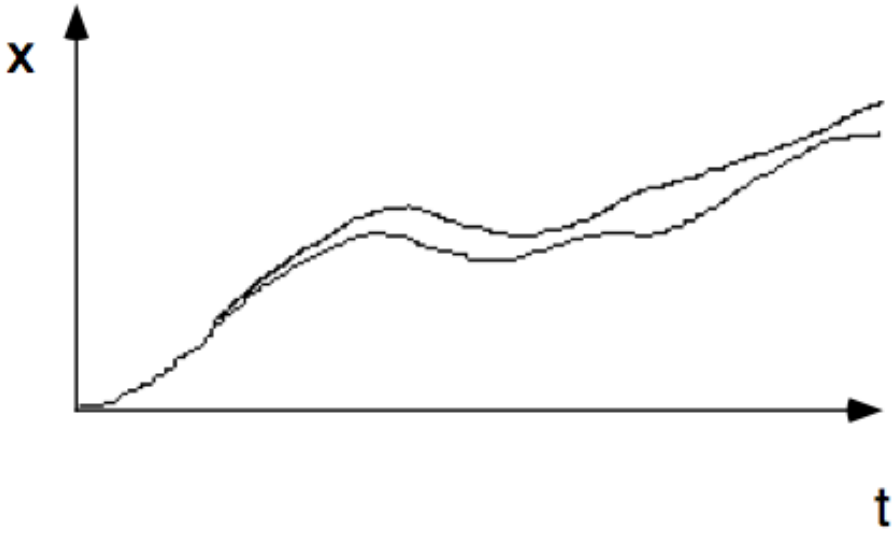
Now let's examine the effects of a small change in $\boldsymbol{u}$, as shown in Figure 6 on $\bar{J}$, just keeping track of the change $\delta\bar{J}$:

$$\delta\bar{J} = \psi[\boldsymbol{x}(T)+\delta\boldsymbol{x}(T)]-\psi[\boldsymbol{x}(T)]+\int_0^T [H(\boldsymbol{\lambda},\boldsymbol{x}+\delta\boldsymbol{x},\boldsymbol{v})-H(\boldsymbol{\lambda},\boldsymbol{x},\boldsymbol{u})-\boldsymbol{\lambda}^T\delta\dot{\boldsymbol{x}}]dt$$

V is like u but w
A small perturbation …



…this causes x to
change a by a small amount



Integration by parts

Using the expression for integration by parts for $\int \boldsymbol{\lambda}^T \delta\dot{\boldsymbol{x}} dt$:

$$\int_0^T \boldsymbol{\lambda}^T \delta\dot{\boldsymbol{x}}\, dt = \boldsymbol{\lambda}^T(T)\delta\boldsymbol{x}(T) - \boldsymbol{\lambda}^T(0)\delta\boldsymbol{x}(0) - \int_0^T \dot{\boldsymbol{\lambda}}^T \delta\boldsymbol{x}\, dt$$

Now substitute this into the expression for $\delta\bar{J}$,

$$\begin{aligned}\delta\bar{J} \;=\; & \psi[\boldsymbol{x}(T) + \delta\boldsymbol{x}(T)] - \psi[\boldsymbol{x}(T)] - \boldsymbol{\lambda}(T)^T\delta\boldsymbol{x}(T) \\ & + \int_0^T [H(\boldsymbol{\lambda}, \boldsymbol{x} + \delta\boldsymbol{x}, \boldsymbol{v}) - H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u}) + \dot{\boldsymbol{\lambda}}^T \delta\boldsymbol{x}]dt\end{aligned}$$

Variational analysis

Now concentrate just on the first two terms in the integral:

$$\int_0^T [H(\boldsymbol{\lambda}, \boldsymbol{x} + \delta\boldsymbol{x}, \boldsymbol{v}) - H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u})]dt$$

First add and subtract $H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{v})$:

$$= \int_0^T [H(\boldsymbol{\lambda}, \boldsymbol{x} + \delta\boldsymbol{x}, \boldsymbol{v}) - H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{v}) + H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{v}) - H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u})]dt$$

Next expand the first term inside the integral in a Taylor series and neglect terms above first order,

$$\cong \int_0^T (H_{\boldsymbol{x}}(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{v})^T \delta\boldsymbol{x} + H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{v}) - H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u}))dt$$

where $H_{\boldsymbol{x}}$ is the partial $\frac{\partial H}{\partial \boldsymbol{x}}$, which is

$$\begin{pmatrix} H_{x_1} \\ \vdots \\ H_{x_n} \end{pmatrix}$$

Variational Analysis Cont.

Now add and subtract $H_{\boldsymbol{x}}(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u})^T \delta \boldsymbol{x}$:

$$= \int_0^T \{H_{\boldsymbol{x}}(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u})^T \delta \boldsymbol{x} + [H_{\boldsymbol{x}}(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{v}) - H_{\boldsymbol{x}}(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u})]^T \delta \boldsymbol{x} + H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{v}) - H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u})\} dt$$

The term $[H_{\boldsymbol{x}}(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{v}) - H_{\boldsymbol{x}}(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u})]^T \delta \boldsymbol{x}$ can be neglected because it is the product of two small terms, $[H_{\boldsymbol{x}}(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{v}) - H_{\boldsymbol{x}}(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u})]$ and $\delta \boldsymbol{x}$, and thus is a second-order term. Thus

$$\cong \int_0^T (H_{\boldsymbol{x}}(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u}) \delta \boldsymbol{x} + H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{v}) - H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u})) dt$$

Adjoint Equation

Finally, substitute this expression back into the original equation for δJ, yielding

$$\begin{aligned}\delta \bar{J} \cong & \{\psi_{\boldsymbol{x}}[\boldsymbol{x}(T)] - \boldsymbol{\lambda}^T(T)\}\delta\boldsymbol{x}(T) \\ & + \int_0^T [H_{\boldsymbol{x}}(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u}) + \dot{\boldsymbol{\lambda}}^T]\delta\boldsymbol{x}\, dt \\ & + \int_0^T [H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{v}) - H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u})]dt\end{aligned}$$

Since we have the freedom to pick $\boldsymbol{\lambda}$, just to make matters simpler, pick it so that the first integral vanishes:

$$-\dot{\boldsymbol{\lambda}}^T = H_{\boldsymbol{x}}(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u})$$

$$\boldsymbol{\lambda}^T(T) = \psi_{\boldsymbol{x}}[\boldsymbol{x}(T)]$$

Condition for a maximum

Now all $\delta \bar{J}$ has left is

$$\delta \bar{J} = \int_0^T [H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{v}) - H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u})] dt$$

From this equation it follows that the optimal control $\boldsymbol{u}^*$ must be such that

$$H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u}^*) \geq H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u}), \ \boldsymbol{u} \in U \tag{3}$$

To see this point, suppose that it were not true, that is, that for some interval of time there was a $\boldsymbol{v} \in U$ such that

$$H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{v}) \geq H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u}^*)$$

This assumption would mean that you could adjust the integral so that the perturbation $\delta \bar{J}$ is positive, contradicting the original assumption that $\bar{J}$ is maximimized by $\boldsymbol{u}^*$. Therefore Equation 3 must hold.

Summary

In addition to the dynamic equations

$$\dot{\boldsymbol{x}} = f(\boldsymbol{x}, \boldsymbol{u})$$

and associated initial condition

$$\boldsymbol{x}(0) = \boldsymbol{x}_0$$

the Lagrange multipliers also must obey a constraint equation

$$-\dot{\boldsymbol{\lambda}}^T = H_{\boldsymbol{x}}$$

that has a *final condition*

$$\boldsymbol{\lambda}^T(T) = \psi_{\boldsymbol{x}}[\boldsymbol{x}(T)]$$

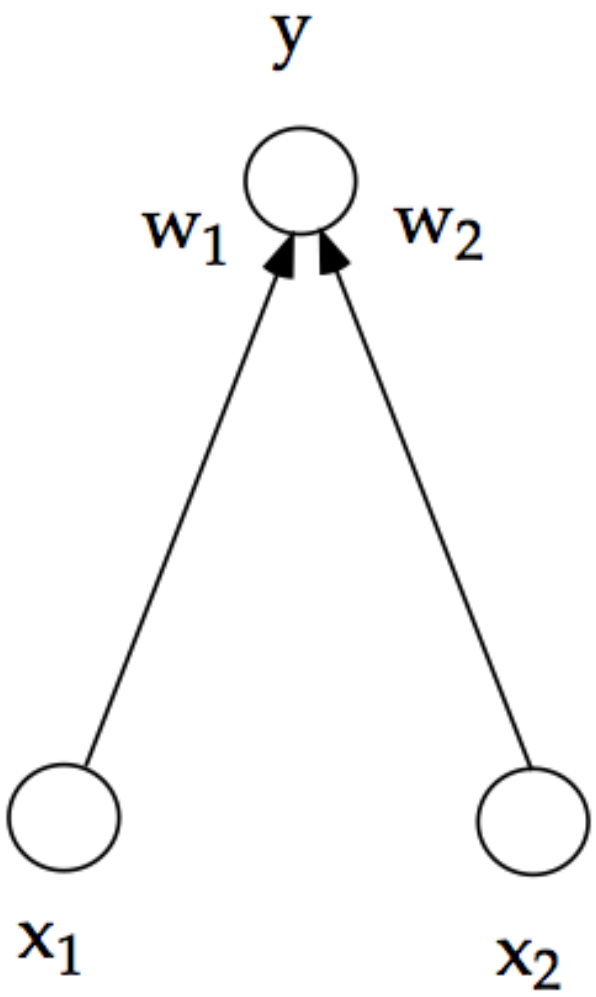
The equation for $\boldsymbol{\lambda}$ is known as the *adjoint equation*. In addition, for all t, the optimal control u is such that

$$H[\boldsymbol{\lambda}(t), \boldsymbol{x}(t), \boldsymbol{v}] \leq H[\boldsymbol{\lambda}(t), \boldsymbol{x}(t), \boldsymbol{u}(t)]$$

where H is the Hamiltonian

$$H = \boldsymbol{\lambda}^T f(\boldsymbol{x}, \boldsymbol{u}) + \ell(\boldsymbol{x}, \boldsymbol{u})$$

A Neural Network



$$y = \sum_k w_k x_k$$

Error Function

"(What you wanted – what you got)^2 "

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_p (y^p - y)^2$$

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_p \left(y^p - \sum_k w_k x_k\right)^2$$

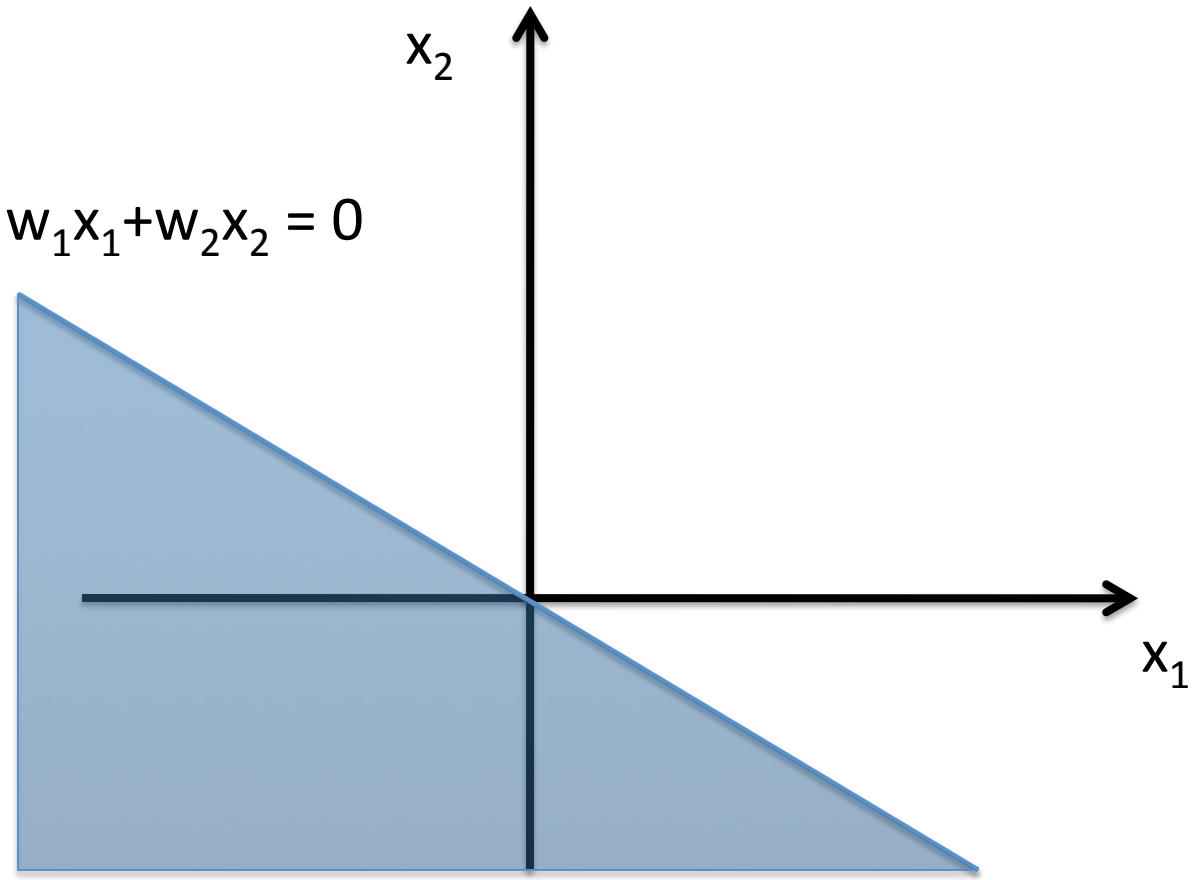
Gradient Minimization

$$w_k^{new} = w_k^{old} - \alpha \frac{\partial E(\boldsymbol{w})}{\partial w_k}$$

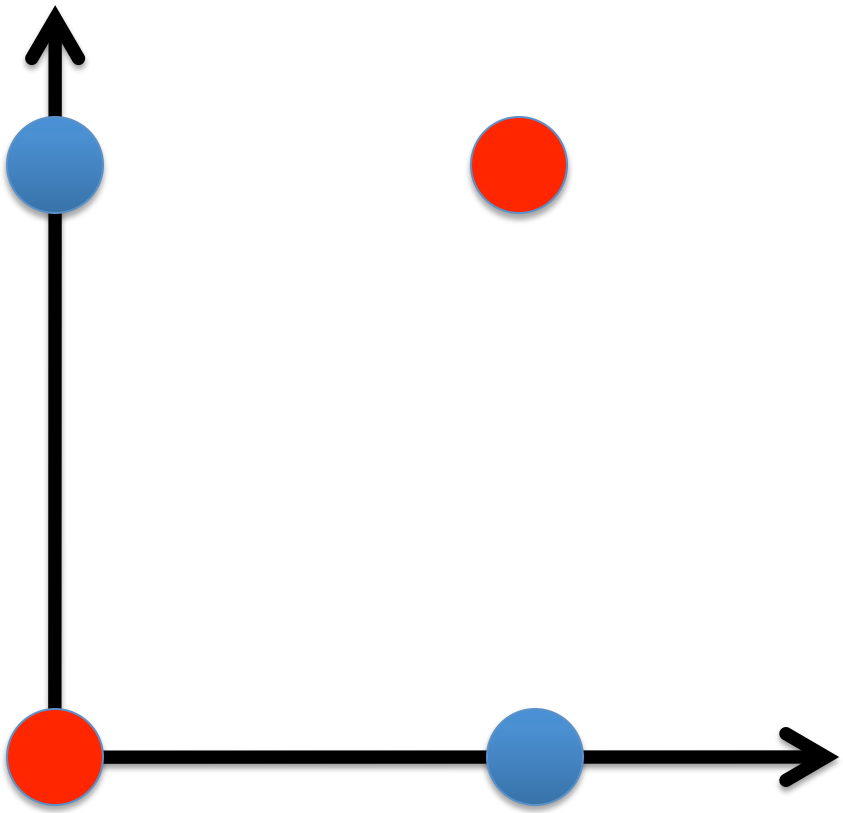
$$\frac{\partial E(\boldsymbol{w})}{\partial w_k} = -\sum_p (y^p - y) x_k$$

Widrow Hoff learning rule

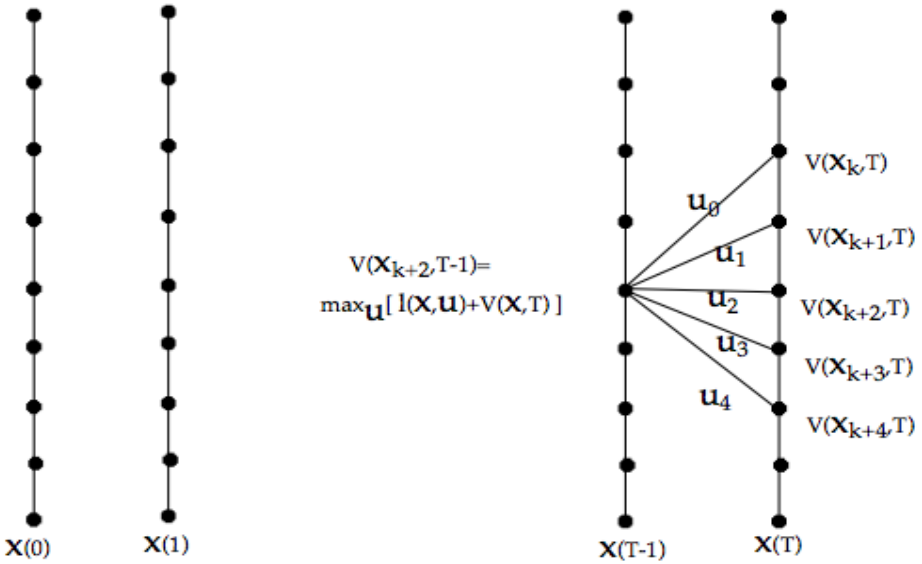
Network is linear!

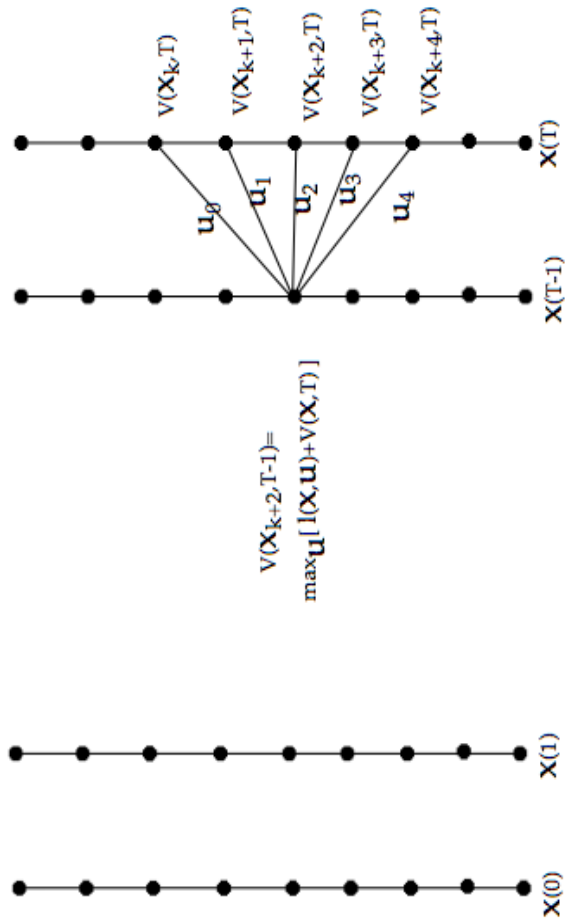


Linear networks cant solve XOR



Nonlinear networks: idea #1 from DP





Concept #2 From Euler-Lagrange

$$\bar{J} = J - \int_0^T \boldsymbol{\lambda}^T [\dot{\boldsymbol{x}} - \boldsymbol{F}(\boldsymbol{x}, \boldsymbol{u})] dt$$

$$-\dot{\boldsymbol{\lambda}}^T = H_{\boldsymbol{x}}$$

that has a *final condition*

$$\boldsymbol{\lambda}^T(T) = \psi_{\boldsymbol{x}}[\boldsymbol{x}(T)]$$

The equation for $\boldsymbol{\lambda}$ is known as the *adjoint equation*. In addition, for all t, the optimal control u is such that

$$H[\boldsymbol{\lambda}(t), \boldsymbol{x}(t), \boldsymbol{v}] \leq H[\boldsymbol{\lambda}(t), \boldsymbol{x}(t), \boldsymbol{u}(t)]$$

where H is the Hamiltonian

$$H = \boldsymbol{\lambda}^T f(\boldsymbol{x}, \boldsymbol{u}) + \ell(\boldsymbol{x}, \boldsymbol{u})$$

And now for the Backprop Algorithm …

$$E = \frac{1}{2}\|\boldsymbol{x}^K - \boldsymbol{d}\|^2 \tag{1}$$

where K is an index denoting the last layer in the network and d is the desired output.

$$E = \frac{1}{2}\|\boldsymbol{x}^K - \boldsymbol{d}\|^2 \tag{1}$$

where K is an index denoting the last layer in the network and d is the desired output.

The equation for updating the states is given by

$$\boldsymbol{x}^{k+1} = g(W^k \boldsymbol{x}^k) \tag{2}$$

$$E = \frac{1}{2}\|\boldsymbol{x}^K - \boldsymbol{d}\|^2 \tag{1}$$

where K is an index denoting the last layer in the network and d is the desired output.

The equation for updating the states is given by

$$\boldsymbol{x}^{k+1} = g(W^k \boldsymbol{x}^k) \tag{2}$$

where W^k is a matrix used to store the weights between layer k and layer $k+1$,

$$W^k = \begin{bmatrix} w_{11}^k & w_{12}^k & \cdots \\ w_{21}^k & \ddots & \\ \vdots & & w_{nn}^k \end{bmatrix} = \begin{pmatrix} \boldsymbol{w}_1^k \\ \boldsymbol{w}_2^k \\ \vdots \\ \boldsymbol{w}_n^k \end{pmatrix}$$

and the special understanding we shall have is that the function g applied to a vector is just that function applied to its elements; that is,

$$g(W\boldsymbol{x}) = \begin{pmatrix} g(\boldsymbol{w}_1 \cdot \boldsymbol{x}) \\ g(\boldsymbol{w}_2 \cdot \boldsymbol{x}) \\ \vdots \end{pmatrix}$$

Adjoint Equation

$$H = \frac{1}{2}\|\boldsymbol{x}^K - \boldsymbol{d}\|^2 + \sum_{k=0}^{k=K-1} (\boldsymbol{\lambda}^{k+1})^T[-\boldsymbol{x}^{k+1} + g(W^k \boldsymbol{x}^k)]$$

Adjoint Equation

$$H = \frac{1}{2}\|\boldsymbol{x}^K - \boldsymbol{d}\|^2 + \sum_{k=0}^{k=K-1} (\boldsymbol{\lambda}^{k+1})^T[-\boldsymbol{x}^{k+1} + g(W^k\boldsymbol{x}^k)]$$

Differentiating with respect to $\boldsymbol{x}^k$ provides the adjoint system of equations,

$$H_{\boldsymbol{x}^k} = \mathbf{0} = -\boldsymbol{\lambda}^k + [W^{kT}\Lambda^{k+1}g'(W^k\boldsymbol{x}^k)] \text{ for } k = 0, \ldots, K-1 \qquad (3)$$

where

$$\Lambda^{k+1} = \begin{bmatrix} \lambda_1^{k+1} & 0 & \cdots \\ 0 & \lambda_2^{k+1} & \\ \vdots & & \lambda_n^{k+1} \end{bmatrix}$$

and with the final condition given by

$$H_{\boldsymbol{x}^K} = \mathbf{0} = \boldsymbol{x}^K - \boldsymbol{d} - \boldsymbol{\lambda}^K \text{ for } k = K$$

$[W^{kT}\Lambda^{k+1}g'(W^k\boldsymbol{x}^k)]$ = ?????

$$[W^{kT}\Lambda^{k+1}g'(W^k\boldsymbol{x}^k)]$$

$$(\boldsymbol{\lambda}^{k+1})^T g(W^k\boldsymbol{x}^k) = \lambda_1^{k+1} g(w_{11}x_1^k + w_{12}x_2^k) + \lambda_2^{k+1} g(w_{21}x_1^k + w_{22}x_2^k)$$

$[W^{kT}\Lambda^{k+1}g'(W^k\boldsymbol{x}^k)]$

$$(\boldsymbol{\lambda}^{k+1})^T g(W^k\boldsymbol{x}^k) = \lambda_1^{k+1} g(w_{11}x_1^k + w_{12}x_2^k) + \lambda_2^{k+1} g(w_{21}x_1^k + w_{22}x_2^k)$$

Taking the partial derivative with respect to x_1^k,

$$\lambda_1^{k+1} g'(w_{11}x_1^k + w_{12}x_2^k)w_{11} + \lambda_2^{k+1} g'(w_{21}x_1^k + w_{22}x_2^k)w_{21}$$

$[W^{kT}\Lambda^{k+1}g'(W^k\boldsymbol{x}^k)]$

$$(\boldsymbol{\lambda}^{k+1})^T g(W^k\boldsymbol{x}^k) = \lambda_1^{k+1} g(w_{11}x_1^k + w_{12}x_2^k) + \lambda_2^{k+1} g(w_{21}x_1^k + w_{22}x_2^k)$$

Taking the partial derivative with respect to x_1^k,

$$\lambda_1^{k+1} g'(w_{11}x_1^k + w_{12}x_2^k)w_{11} + \lambda_2^{k+1} g'(w_{21}x_1^k + w_{22}x_2^k)w_{21}$$

Similarly the partial derivative with respect to x_2^k,

$$\lambda_1^{k+1} g'(w_{11}x_1^k + w_{12}x_2^k)w_{12} + \lambda_2^{k+1} g'(w_{21}x_1^k + w_{22}x_2^k)w_{22}$$

$[W^{kT}\Lambda^{k+1}g'(W^k\boldsymbol{x}^k)]$

$$(\boldsymbol{\lambda}^{k+1})^T g(W^k\boldsymbol{x}^k) = \lambda_1^{k+1} g(w_{11}x_1^k + w_{12}x_2^k) + \lambda_2^{k+1} g(w_{21}x_1^k + w_{22}x_2^k)$$

Taking the partial derivative with respect to x_1^k,

$$\lambda_1^{k+1} g'(w_{11}x_1^k + w_{12}x_2^k)w_{11} + \lambda_2^{k+1} g'(w_{21}x_1^k + w_{22}x_2^k)w_{21}$$

Similarly the partial derivative with respect to x_2^k,

$$\lambda_1^{k+1} g'(w_{11}x_1^k + w_{12}x_2^k)w_{12} + \lambda_2^{k+1} g'(w_{21}x_1^k + w_{22}x_2^k)w_{22}$$

Reassembling these terms into a vector results in

$$= W^T \begin{pmatrix} \lambda_1^{k+1} g'(\boldsymbol{w}_1^k \cdot \boldsymbol{x}^k) \\ \lambda_2^{k+1} g'(\boldsymbol{w}_2^k \cdot \boldsymbol{x}^k) \\ \vdots \end{pmatrix} = W^{kT}\Lambda^{k+1} g'(W^k\boldsymbol{x}^k)$$

Backpropagation Algorithm

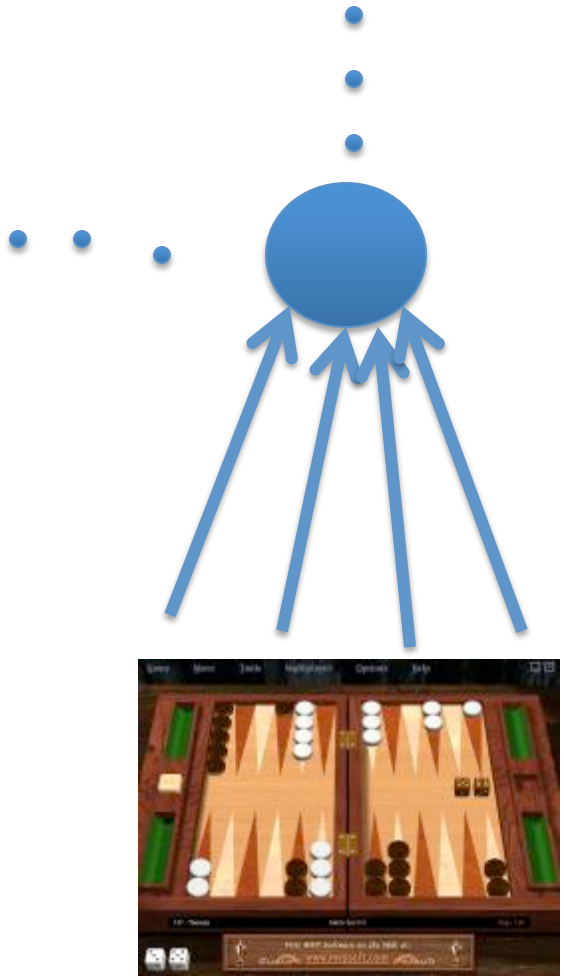
$$\boldsymbol{\lambda}^K = \boldsymbol{x}^K - \boldsymbol{d} \quad \text{for } k = K$$

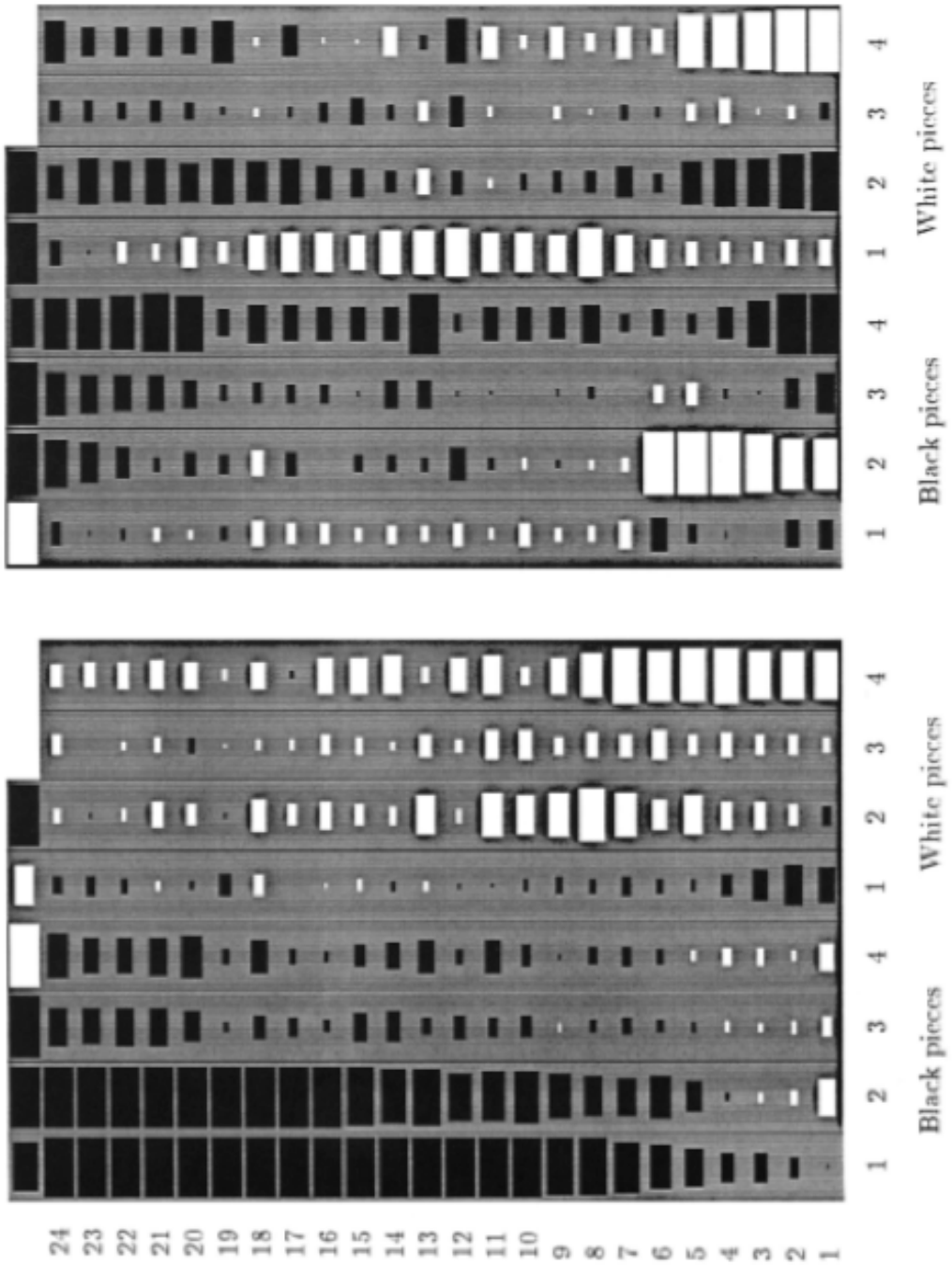
$$\boldsymbol{\lambda}^k = -W^T \Lambda^{k+1} g'(W\boldsymbol{x}^k) \quad \text{for } k = 0, \ldots, K-1$$

$$\frac{\partial H}{\partial w_{ij}^k} = \lambda_i^{k+1} x_j^k g'(\boldsymbol{w}_i^k \cdot \boldsymbol{x}^k)$$

Backgammon

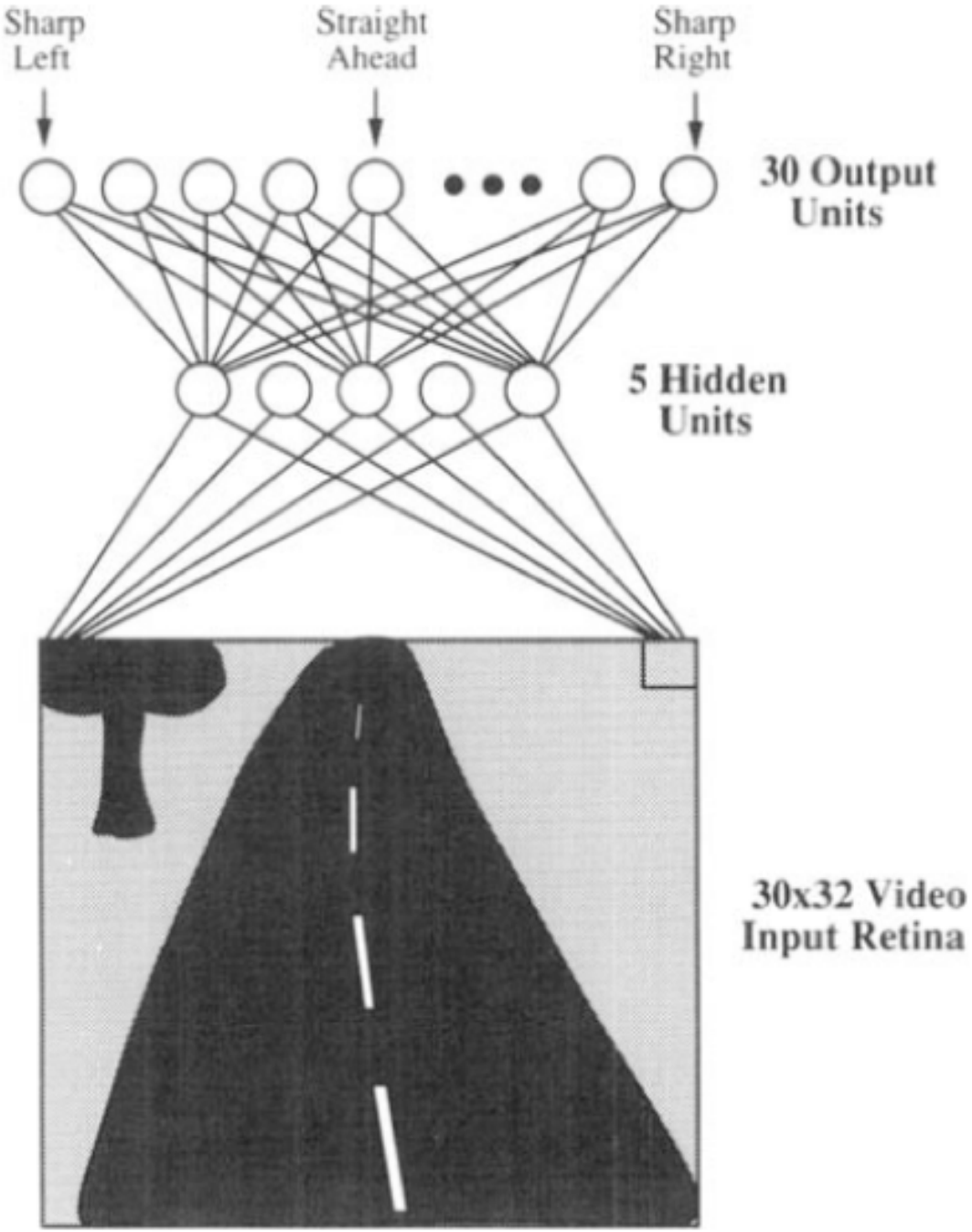
Backgammon



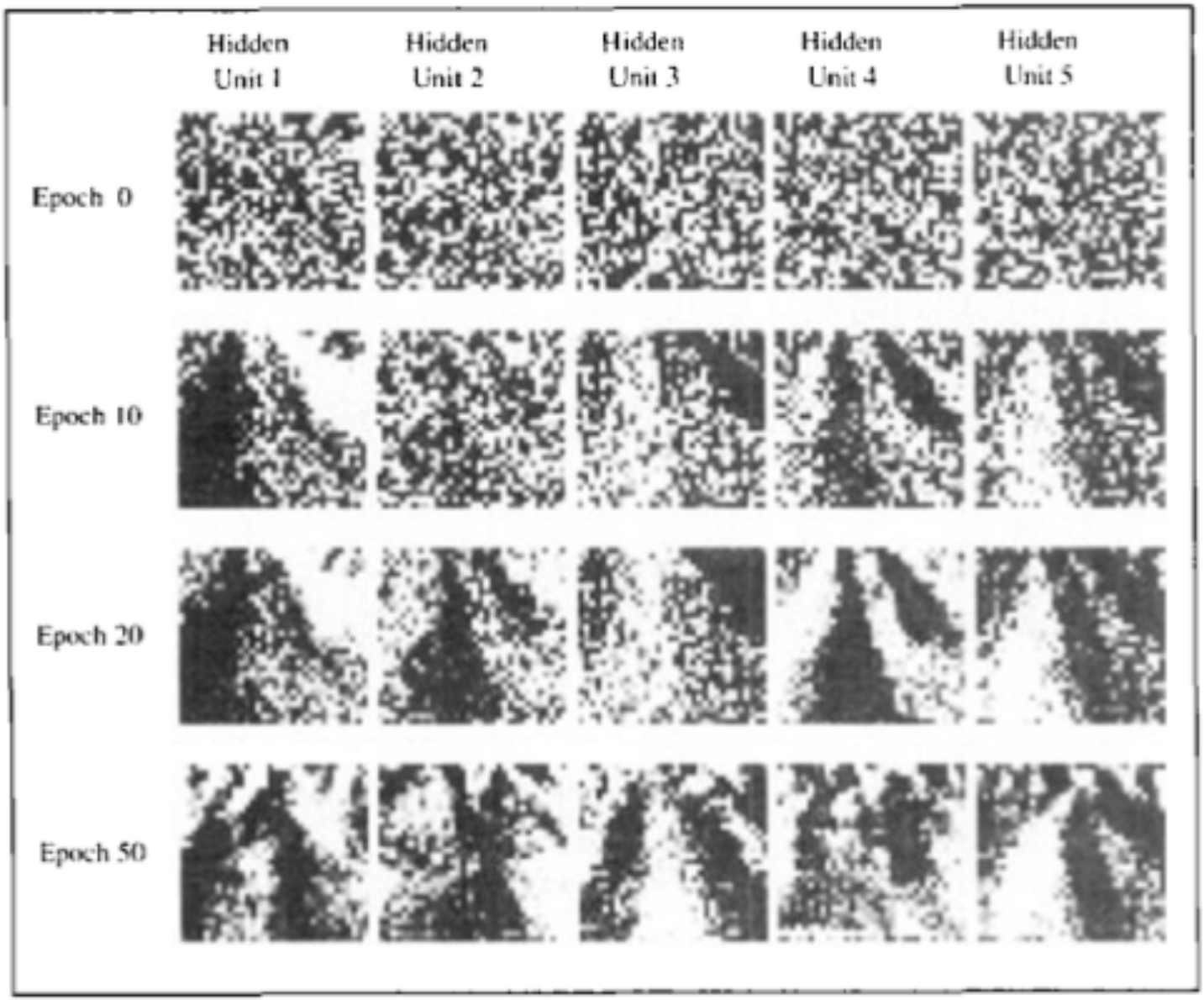

White pieces
Black pieces
White pieces
Black pieces
24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

Driving a Van

Network


Sharp Left
Straight Ahead
Sharp Right
30 Output Units
5 Hidden Units
30x32 Video Input Retina

Results


Hidden Unit 1
Hidden Unit 2
Hidden Unit 3
Hidden Unit 4
Hidden Unit 5
Epoch 0
Epoch 10
Epoch 20
Epoch 50

The Cart

The dynamic equation is

$$\ddot{x} = -\dot{x} + u(t)$$

with initial conditions

$$x(0) = 0$$
$$\dot{x}(0) = 0$$

The cost functional

$$J = x(T) - \frac{1}{2}\int_0^T u^2(t)dt$$

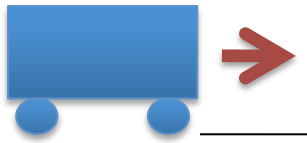
captures the desire to maximize the distance traveled in time T and at the same time penalize excessive accelerations.

Using the transformation of Section 5.2.1, the state variables x_1 and x_2 are defined by

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} x_2 \\ -x_2 + u \end{pmatrix}$$
$$x_1(0) = x_2(0) = 0$$
$$J = x_1(T) - \frac{1}{2}\int_0^T u^2 dt$$

The Hamiltonian is given by

$$H = \lambda_1 x_2 - \lambda_2 x_2 + \lambda_2 u - \frac{1}{2}u^2$$



Differentiating this equation allows the determination of the adjoint system as

$$-\dot{\lambda}_1 = \frac{\partial H}{\partial x_1} = 0$$

$$-\dot{\lambda}_2 = \frac{\partial H}{\partial x_2} = \lambda_1 - \lambda_2$$

Solution

and its final condition can be determined from

$$\psi = x_1(T)$$

$$\boldsymbol{\lambda}(T) = \psi_{\boldsymbol{x}}(T) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

The simple form of the adjoint equations allows their direct solution. For λ_1,

$$\lambda_1 = \text{ const } = 1$$

For λ_2,we could use Laplace transform methods, but they have not been discussed, so lets make the incredible lucky guess:

$$\lambda_2 = 1 - e^{t-T}$$

For a maximum differentiate H with respect to u,

$$\frac{\partial H}{\partial u} = 0 \Rightarrow \lambda_2 - u = 0$$

$$u = \lambda_2 = 1 - e^{t-T}$$