Machine Intelligence 1

1.4 Additional Topics

Prof. Dr. Klaus Obermayer

Fachgebiet Neuronale Informationsverarbeitung (NI)

WS 2017/2018

1.4.1 Stochastic Approximation and Online Learning

Online learning

$$\Delta \mathrm{w}_{ij}^{v'v} \quad = \quad -\eta \frac{\partial E_{[\underline{\mathbf{w}}]}^{T}}{\partial \mathrm{w}_{ij}^{v'v}} \quad = \quad -\eta \frac{1}{p} \sum_{\alpha=1}^{p} \frac{\partial e_{[\underline{\mathbf{w}}]}^{(\alpha)}}{\partial \mathrm{w}_{ij}^{v'v}}$$

Online learning

$$\Delta \mathrm{w}_{ij}^{v'v} = -\eta \frac{\partial E^T_{[\underline{\mathbf{w}}]}}{\partial \mathrm{w}_{ij}^{v'v}} = -\eta \frac{1}{p} \sum_{\alpha=1}^{p} \frac{\partial e^{(\alpha)}_{[\underline{\mathbf{w}}]}}{\partial \mathrm{w}_{ij}^{v'v}}$$

while *convergence criterion not met* **do**

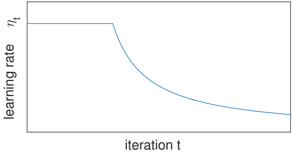
select the a data point: $\left(\underline{\mathbf{x}}^{(\alpha)}, y_T^{(\alpha)}\right)$

change weights according to: $\left(\Delta \mathrm{w}_{ij}^{v'v}\right)^{(t+1)} = -\eta_t \frac{\partial e^{(\alpha)}_{[\underline{\mathbf{w}}^{(t)}]}}{\partial \mathrm{w}_{ij}^{v'v}}$

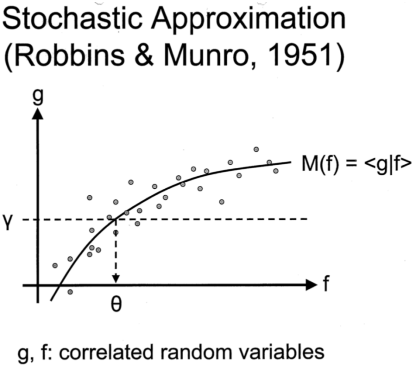
update learning rate η_t

end

- Adaptive learning rate
  - first constant $\eta_t = \eta_0$
  - then decaying $\eta_t = \frac{\eta_0}{t}$

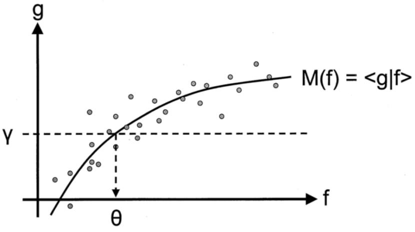


Convergence in online learning (1)



Convergence in online learning (1)

Stochastic Approximation (Robbins & Munro, 1951)



g, f: correlated random variables

Application to online learning:

- $f \rightarrow$ model parameters $\underline{\mathbf{w}}$
- $M(f) \rightarrow$ batch-gradients $\frac{\partial E^T}{\partial \underline{\mathbf{w}}}\Big|_{[\underline{\mathbf{w}}]}$
- $g \rightarrow$ individual gradients $\frac{\partial e^\alpha}{\partial \underline{\mathbf{w}}}\Big|_{[\underline{\mathbf{w}}]}$
- $\theta \rightarrow$ optimal parameters $\underline{\mathbf{w}}^*$, i.e.

$$\frac{\partial E^T}{\partial \underline{\mathbf{w}}}\Big|_{[\underline{\mathbf{w}}^*]} = M(\theta) \stackrel{!}{=} \gamma = 0$$

Convergence in online learning (2)

Let $g_t, f_t \in \mathbb{R},\ g_t|f_t,$ be correlated random variables and let the initial f_1 be an arbitrary real number. If

- g_t are bounded
- $M(f)$ is monotonously increasing
- $M(\theta) = \gamma$ and $\frac{\partial M(\theta)}{\partial f} > 0$
- and for the learning rates η_t holds

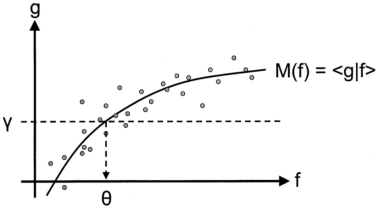
$$\sum_{t=1}^{\infty} \eta_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty ,$$

then the sequence

$$f_{t+1} = f_t + \eta_t (\gamma - g_t) , \qquad \eta_t > 0 ,$$

converges to θ in the sense $\lim_{t \to \infty} \langle (f_t - \theta)^2 \rangle = 0 .$

Stochastic Approximation (Robbins & Munro, 1951)



g, f: correlated random variables

Convergence in online learning (2)

Let $\frac{\partial e^\alpha}{\partial \underline{\mathbf{w}}}\Big|_{[\underline{\mathbf{w}}_t]}, \underline{\mathbf{w}}_t \in \mathbb{R}^{|C|}$ be correlated random variables and let the initial $\underline{\mathbf{w}}_1$ be an arbitrary real vector. If

- $\frac{\partial e^\alpha}{\partial \underline{\mathbf{w}}}\Big|_{[\underline{\mathbf{w}}_t]}$ are bounded
- E^T is a convex function
- $\frac{\partial E^T}{\partial \underline{\mathbf{w}}}\Big|_{[\underline{\mathbf{w}}^*]} = 0$ and $\frac{\partial^2 E^T[\underline{\mathbf{w}}^*]}{\partial^2 \underline{\mathbf{w}}} > 0$ (min. $\underline{\mathbf{w}}^*$)
- and for the learning rates η_t holds

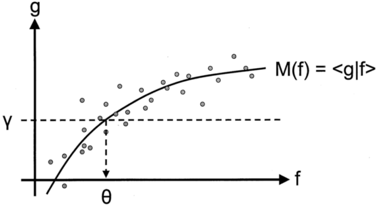
$$\sum_{t=1}^{\infty} \eta_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty ,$$

then the sequence

$$\underline{\mathbf{w}}_{t+1} = \underline{\mathbf{w}}_t - \eta_t \frac{\partial e^\alpha}{\partial \underline{\mathbf{w}}}\Big|_{[\underline{\mathbf{w}}_t]}, \qquad \eta_t > 0 ,$$

converges to $\underline{\mathbf{w}}^*$ in the sense $\lim_{t \to \infty} \langle (\underline{\mathbf{w}}_t - \underline{\mathbf{w}}^*)^2 \rangle = 0$.

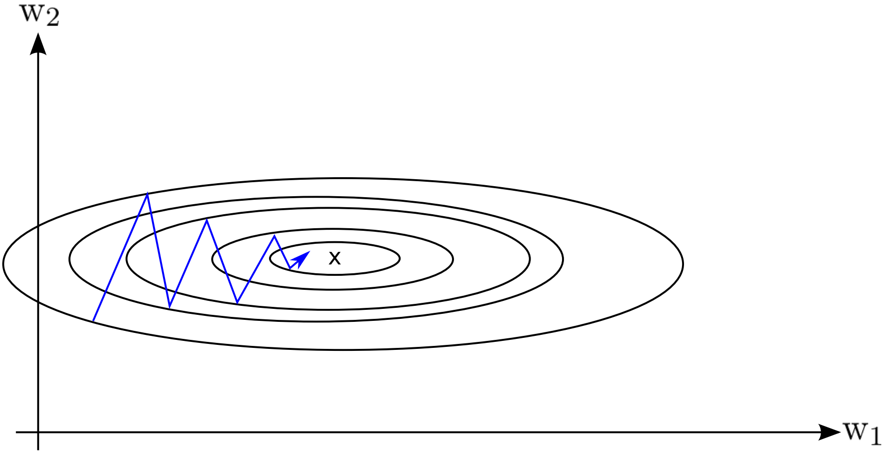
Stochastic Approximation (Robbins & Munro, 1951)



g, f: correlated random variables

1.4.2 Improving Gradient-Descent Optimization

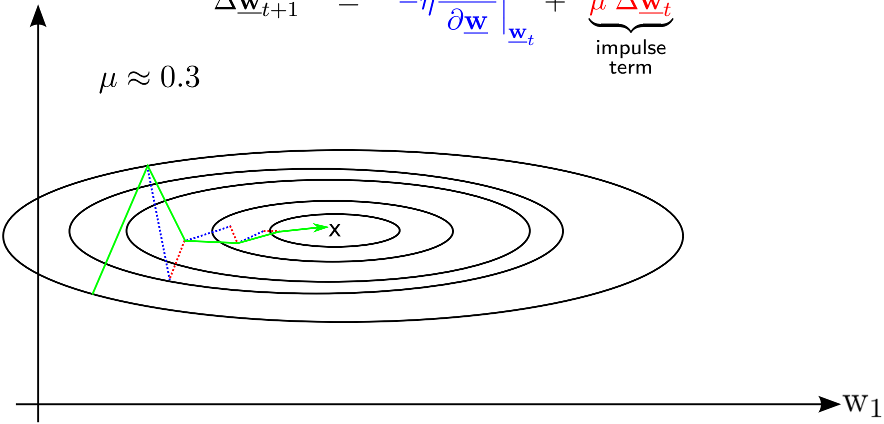
Impulse terms



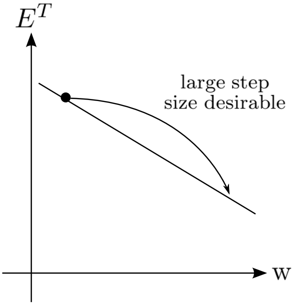
Impulse terms

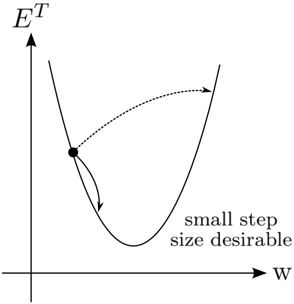
$$\Delta \underline{\mathbf{w}}_{t+1} = -\eta \frac{\partial E^T}{\partial \underline{\mathbf{w}}}\bigg|_{\underline{\mathbf{w}}_t} + \underbrace{\mu \, \Delta \underline{\mathbf{w}}_t}_{\text{impulse term}}$$

$\mu \approx 0.3$



Adaptive step size



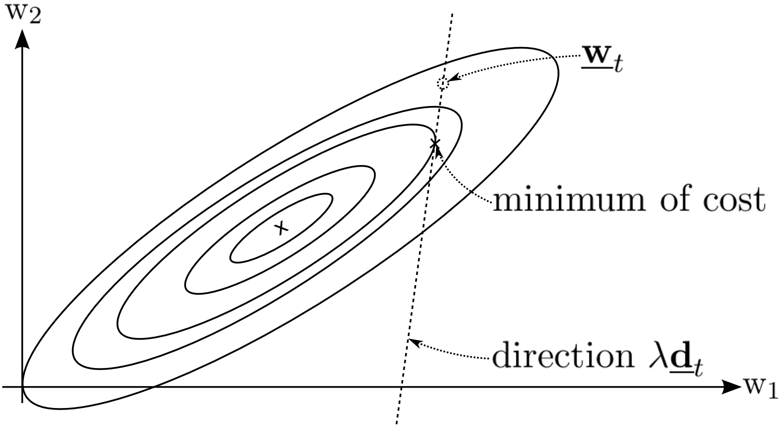


$$\eta_{t+1} = \begin{cases} \rho \eta_t, & \text{if } \Delta E^T < 0, \quad \text{increase step size, if } E^T \downarrow \\ \delta \eta_t, & \text{if } \Delta E^T > 0, \quad \text{decrease step size, if } E^T \uparrow \end{cases}$$

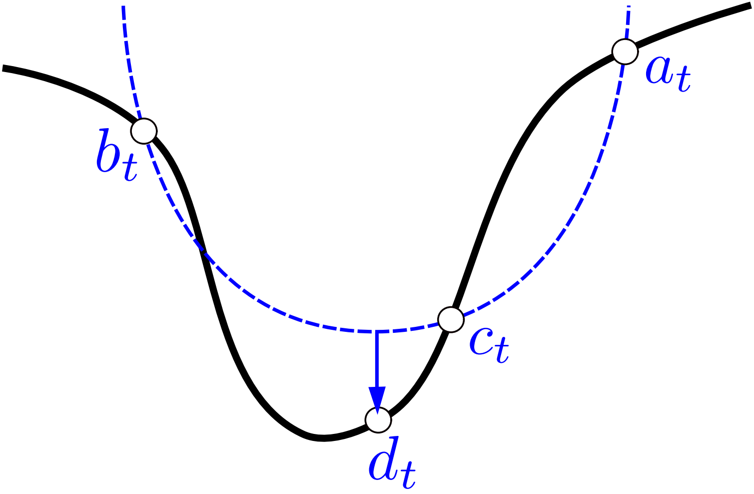
typical values: $\rho = 1.1, \delta = 0.5$

1.4.3 The Conjugate Gradient Method

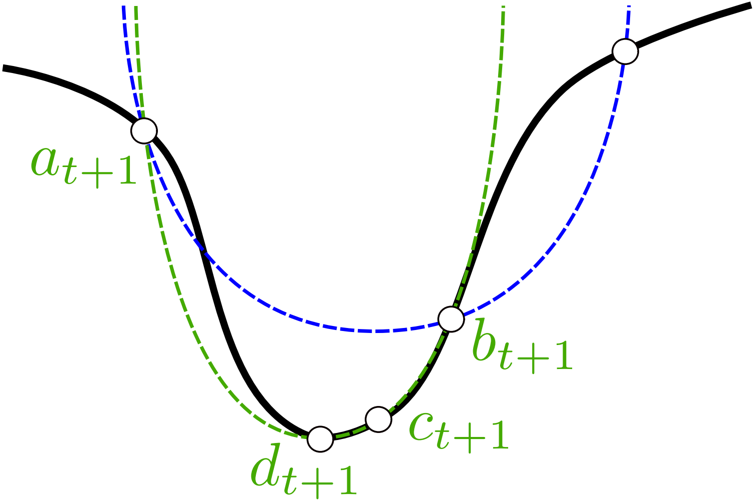
Optimal step size



Parabolic interpolation



Parabolic interpolation



Line search

Line search using successive parabolic interpolation

Initialization: a_0, b_0, c_0 (on $\lambda \underline{\mathbf{d}}_t$); $\quad E^T_{(a_0)}, \quad E^T_{(b_0)} > E^T_{(c_0)}$

while *stopping criterion not fulfilled* **do**

Fit a parabola through the three points a_t, b_t, c_t

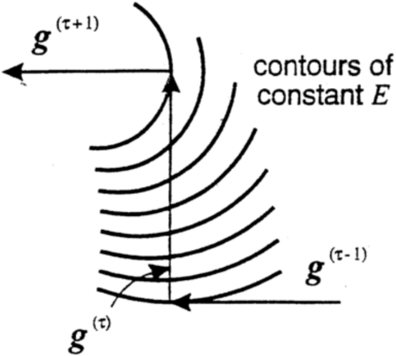
Calculate location d_t of its minimum

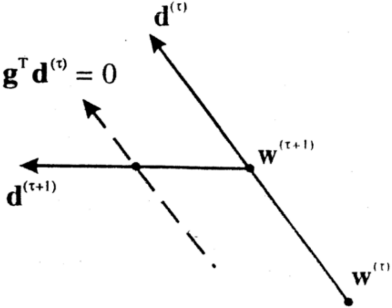
Set $c_{t+1} = d_t, \quad b_{t+1} = c_t, \quad a_{t+1} = \begin{cases} a_t, & E^T_{(a_t)} < E^T_{(b_t)} \\ b_t, & \text{else} \end{cases}$

end

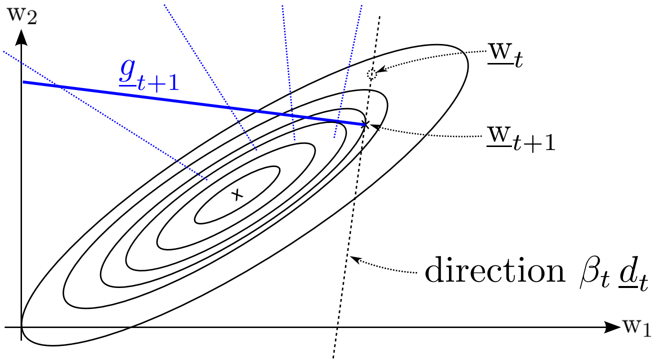
For details and implementation see e.g. Numerical Recipes, 2nd edition, Chapter 10.2.

The conjugate direction

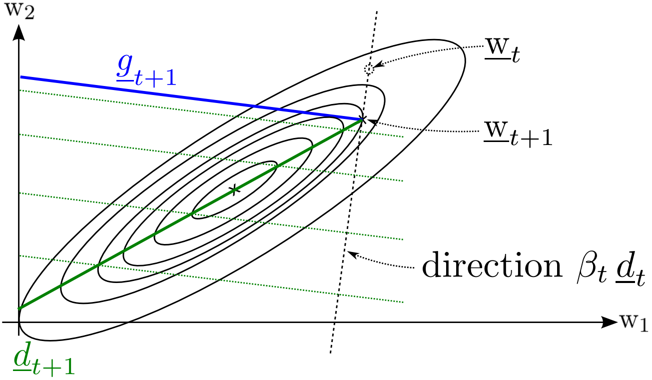




Parabolic cost function



Parabolic cost function



$$\underline{\mathbf{d}}_t^\top \, \underline{\mathbf{H}} \, \underline{\mathbf{d}}_{t+1} \stackrel{!}{=} 0 \,, \qquad H_{ij} := \frac{\partial^2 E^T}{\partial w_i \partial w_j} \quad \text{(Hessian Matrix)}$$

Adaptive momentum: the Polak-Ribiere rule

$$\underline{\mathbf{g}}_{t+1} := \frac{\partial E^T}{\partial \underline{\mathbf{w}}}\bigg|_{\underline{\mathbf{w}}_{t+1}} \qquad \text{(gradient at } \underline{\mathbf{w}}_{t+1}\text{)}$$

$$\underline{\mathbf{d}}_{t+1} = -\underline{\mathbf{g}}_{t+1} + \beta_t \underline{\mathbf{d}}_t \qquad \text{(conjugate direction)}$$

$$\beta_t = \underbrace{\frac{\underline{\mathbf{g}}_{t+1}^T \left(\underline{\mathbf{g}}_{t+1} - \underline{\mathbf{g}}_t\right)}{\underline{\mathbf{g}}_t^T \underline{\mathbf{g}}_t}}_{\text{Polak-Ribiere rule}} \qquad \text{("smart momentum")}$$

Conjugate gradient descent algorithm

$$\mathrm{w}_{ij}(t+1) = \mathrm{w}_{ij}(t) + \eta_t \, \underline{\mathbf{d}}_{t-1}$$

Initialization: $\underline{\mathbf{w}}, \quad \underline{\mathbf{d}} = -\underline{\mathbf{g}}$

while *stopping criterion not fulfilled* **do**

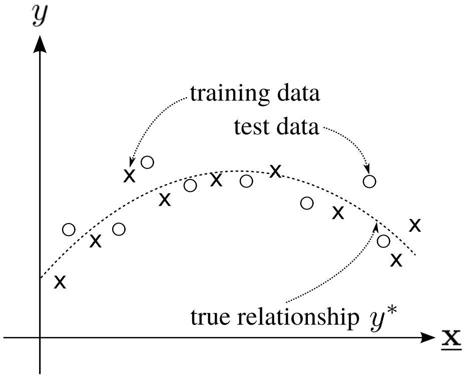
  Minimize E^T along $\underline{\mathbf{d}}$ using line-search $\rightarrow$ new $\underline{\mathbf{w}}$

  Calculate the new conjugate direction $\rightarrow$ new $\underline{\mathbf{d}}$

end

1.4.4 Overfitting and Underfitting

Overfitting and underfitting



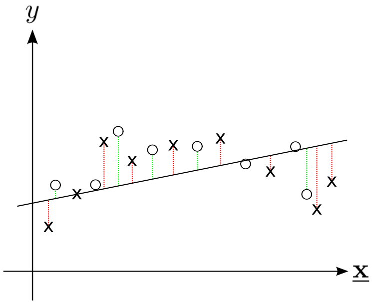
data generation:

e.g. $y_T = y^*_{(\underline{\mathbf{x}})} + \underbrace{\eta}_{\text{noise, zero mean}}$

Goal

Find a **good** model for $y^*_{(\cdot)}$ (explanation, prediction)

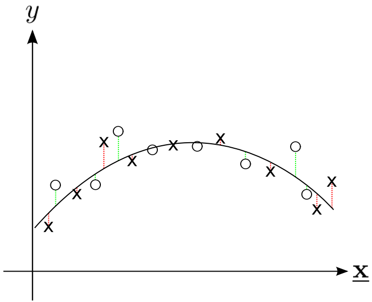
Underfitting



Diagnostics

$$\left.\begin{array}{l} E^T \text{ large} \\ E^G \text{ large} \end{array}\right\} E^T \approx E^G$$

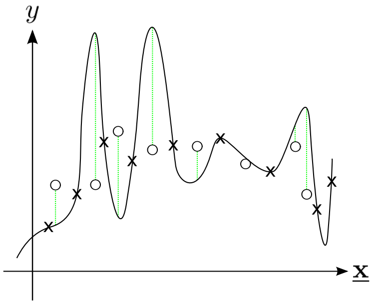
Well fitted model



Diagnostics

$$\left.\begin{array}{l} E^T \text{ small} \\ E^G \text{ small} \end{array}\right\} E^T \approx E^G$$

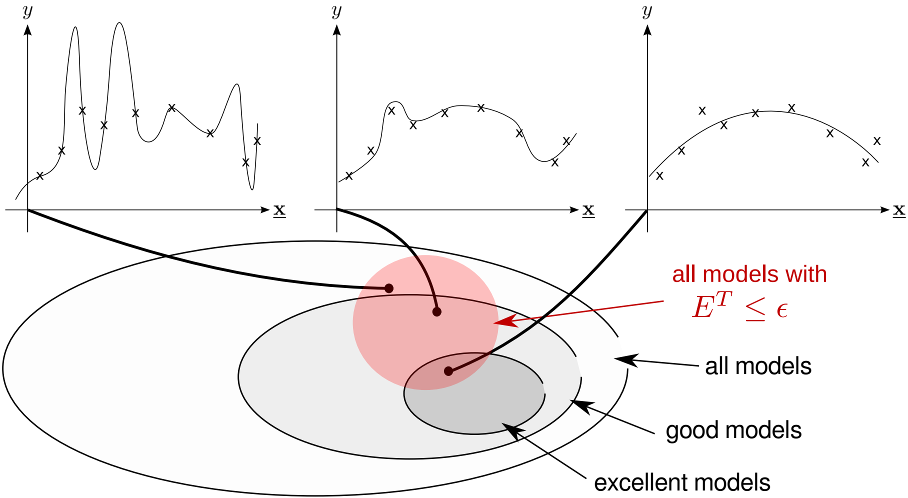
Overfitting



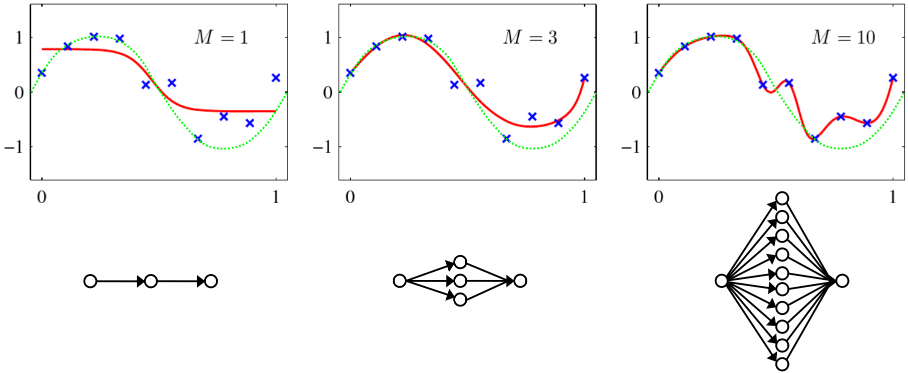
Diagnostics

$$\left.\begin{array}{l} E^T \text{ small} \\ E^G \text{ large} \end{array}\right\} E^T \ll E^G$$

Overfitting and model selection



Example with MLPs



MLP with M sigmoid hidden neurons $f_j^1(x) = \tanh(x)$ and linear output neuron $f_1^2(x) = x$, from Bishop (2006)

1.4.5 Bias and Variance

Bias and variance

Example scenario

Observations: $$y_T \quad = \quad \underbrace{y^*_{(\underline{\mathbf{x}})}}_{\substack{\text{true}\\ \text{relationship}}} + \underbrace{\eta}_{\text{noise}}, \qquad \underline{\mathbf{x}} \in \mathbb{R}^N, \;\; y_T \in \mathbb{R}, \;\; \eta \sim \mathcal{N}(0, \sigma_\eta)$$

- $\underline{\mathbf{x}}, y_T$ are random variables
- Many iid. datasets of equal length from $P_{(y_T, \underline{\mathbf{x}})} = P_{(y_T | \underline{\mathbf{x}})} P_{(\underline{\mathbf{x}})}$

Bias and variance

Example scenario

Observations: $$y_T = \underbrace{y^*_{(\underline{\mathbf{x}})}}_{\text{true relationship}} + \underbrace{\eta}_{\text{noise}}, \quad \underline{\mathbf{x}} \in \mathbb{R}^N, \; y_T \in \mathbb{R}, \; \eta \sim \mathcal{N}(0, \sigma_\eta)$$

- $\underline{\mathbf{x}}, y_T$ are random variables
- Many iid. datasets of equal length from $P_{(y_T, \underline{\mathbf{x}})} = P_{(y_T | \underline{\mathbf{x}})} P_{(\underline{\mathbf{x}})}$

- fitting one MLP to every dataset

$\rightarrow$ $\underline{\mathbf{w}}$ (model parameters) are random variables

$\rightarrow$ $y_{(\underline{\mathbf{x}}; \underline{\mathbf{w}})}$ (predicted values) are random variables

Bias and variance

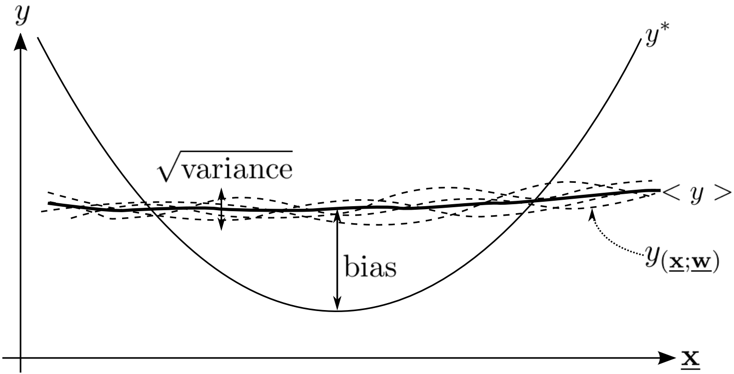
Example: squared error cost

$$\left\langle \left(\underbrace{y_{(\underline{\mathbf{x}};\underline{\mathbf{w}})}}_{y} - \underbrace{y^*_{(\underline{\mathbf{x}})}}_{y^*} \right)^2 \right\rangle_{\text{all datasets}}$$

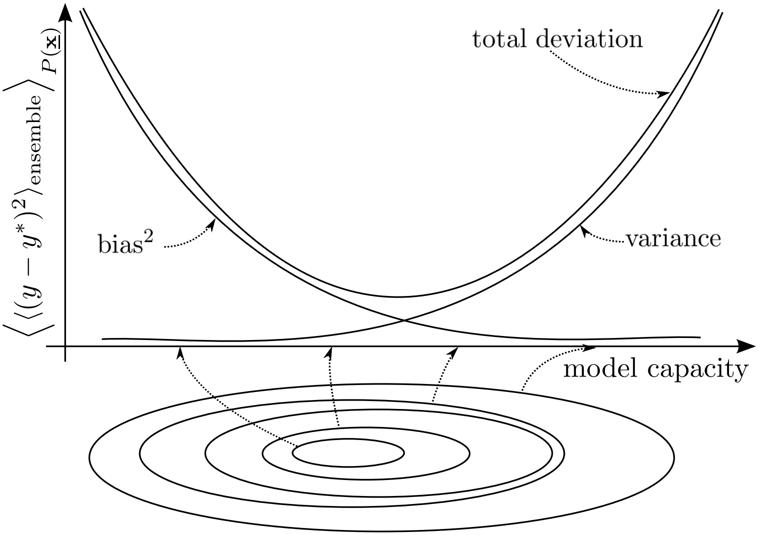
with

$$\begin{aligned} \left\langle (y - y^*)^2 \right\rangle &= \left\langle (y - \langle y \rangle + \langle y \rangle - y^*)^2 \right\rangle \\ &= \underbrace{(y^* - \langle y \rangle)^2}_{\text{bias}^2} + \underbrace{\left\langle (y - \langle y \rangle)^2 \right\rangle}_{\text{variance}} + 2 \underbrace{\langle y - \langle y \rangle \rangle}_{=0} (\langle y \rangle - y^*) \end{aligned}$$

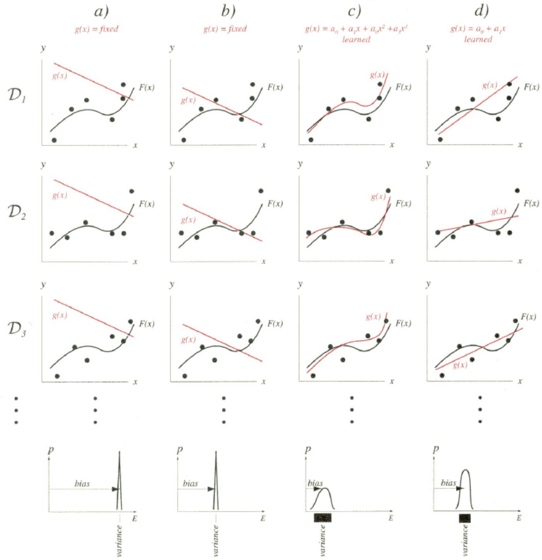
Bias and variance



Bias and variance



Regression example



from Duda & Hart 467

Bias, variance and generalization error

Relation to generalization performance (for previous example)

$$E^G = \left\langle e^G \right\rangle_{P_{(\underline{\mathbf{x}})}}, \text{ where } e^G = \left\langle \left(y_{(\underline{\mathbf{x}};\underline{\mathbf{w}})} - y_T \right)^2 \right\rangle_{P_{(y_T|\underline{\mathbf{x}})}}$$

Bias, variance and generalization error

Relation to generalization performance (for previous example)

$$E^G = \left\langle e^G \right\rangle_{P_{(\underline{\mathbf{x}})}}, \text{ where } e^G = \left\langle \left(y_{(\underline{\mathbf{x}};\underline{\mathbf{w}})} - y_T \right)^2 \right\rangle_{P_{(y_T|\underline{\mathbf{x}})}}$$

$$\begin{aligned} e^G &= \left\langle (y - y_T)^2 \right\rangle \\ &= \left\langle (y - y^* + y^* - y_T)^2 \right\rangle \\ &= (y - y^*)^2 + \left\langle (y^* - y_T)^2 \right\rangle + 2(y - y^*) \underbrace{\left\langle (y^* - y_T) \right\rangle}_{=0} \\ &= (y - y^*)^2 + \sigma_\eta^2 \end{aligned}$$

$$\left\langle e^G \right\rangle_{\text{ensemble}} = \left\langle \underbrace{(y - y^*)^2}_{\text{bias}^2 + \text{variance}} \right\rangle_{\text{ensemble}} + \sigma_\eta^2$$

$\Rightarrow$ bias-variance trade-off applies to wide range of inductive learning problems

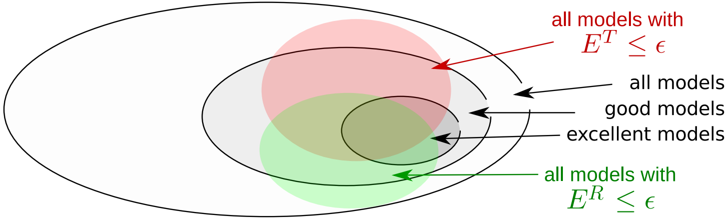
1.4.6 Regularization

Regularization

Risk function

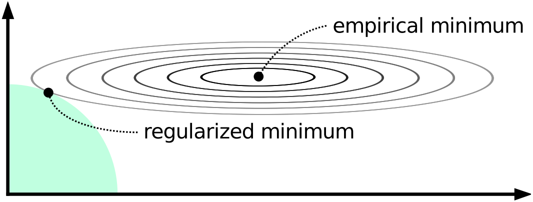
$$R_{[\underline{\mathbf{w}}]} = \underbrace{E^T_{[\underline{\mathbf{w}}]}}_{\text{training error}} + \underbrace{\lambda E^R_{[\underline{\mathbf{w}}]}}_{\text{regularization term}} \overset{!}{=} \min$$

- E^R : prior knowledge of solution
- λ : regularization parameter



Regularization example: weight decay

$$E^R_{[\mathbf{w}]} = \frac{1}{2} \sum_{i,j,v,v'} \left(\mathrm{w}_{ij}^{vv'}\right)^2$$

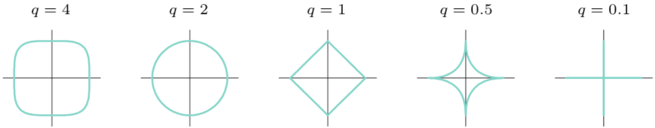


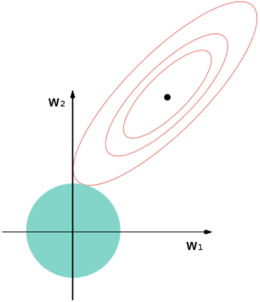
Minimization of R through gradient descent

$$\Delta \mathrm{w}_{ij}^{vv'} \sim -\frac{\partial R}{\partial \mathrm{w}_{ij}^{vv'}} = -\underbrace{\frac{\partial E^T}{\partial \mathrm{w}_{ij}^{vv'}}}_{\text{e.g. via backprop}} - \underbrace{\lambda \mathrm{w}_{ij}^{vv'}}_{\text{decay term}}$$

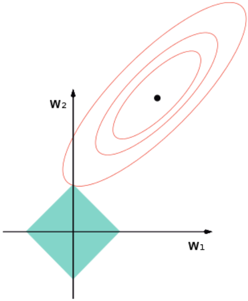
Other forms of regularization

- general form of regularization: $E^R = \sum_{ijv'v} |w_{ij}^{v'v}|^q$





L_2 *weight decay* with $q = 2$



L_1 *lasso* with $q = 1$

Regularization example: symmetries

Odd vs. even function

$$E^R_{[\mathrm{w}]} = \frac{1}{2p} \sum_{\alpha=1}^{p} \left(y_{(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} \pm y_{(-\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} \right)^2$$

Regularization example: symmetries

Odd vs. even function

$$E^R_{[\mathrm{w}]} = \frac{1}{2p} \sum_{\alpha=1}^{p} \left(y_{(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} \pm y_{(-\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} \right)^2$$

Invariance under translation:

$$E^R_{[\mathrm{w}]} = \frac{1}{2p} \sum_{\alpha=1}^{p} \left(y_{(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} - y_{(\underline{\mathbf{x}}^{(\alpha)}-\underline{\mathbf{t}};\underline{\mathbf{w}})} \right)^2$$

Regularization example: symmetries

Odd vs. even function

$$E^R_{[\mathrm{w}]} = \frac{1}{2p} \sum_{\alpha=1}^{p} \Big(y_{(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} \pm y_{(-\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} \Big)^2$$

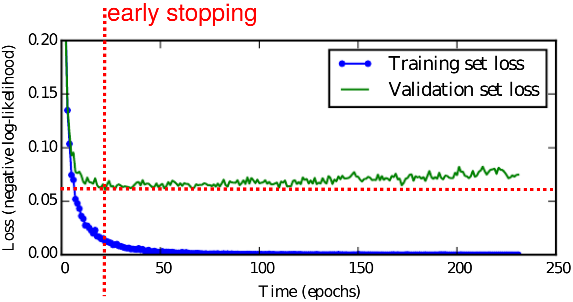
Invariance under translation:

$$E^R_{[\mathrm{w}]} = \frac{1}{2p} \sum_{\alpha=1}^{p} \Big(y_{(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} - y_{(\underline{\mathbf{x}}^{(\alpha)}-\underline{\mathbf{t}};\underline{\mathbf{w}})} \Big)^2$$

Monotony:

$$E^R_{[\mathrm{w}]} = \frac{1}{n_p} \sum_{\mathrm{x}^{(\alpha)} > \mathrm{x}^{(\beta)}} \begin{cases} \Big(y_{(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} - y_{(\underline{\mathbf{x}}^{(\beta)};\underline{\mathbf{w}})} \Big)^2 & \text{, if } y_{(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} < y_{(\underline{\mathbf{x}}^{(\beta)};\underline{\mathbf{w}})} \\ 0 & \text{, else} \end{cases}$$

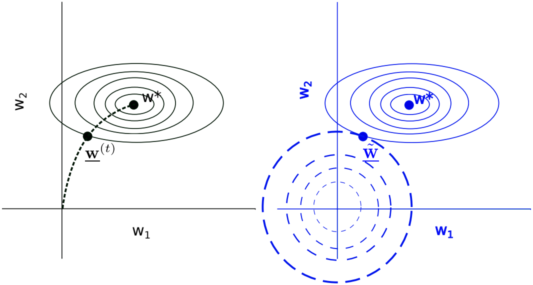
Early stopping



- estimate generalization error with a validation set during training
- stop training when the validation error rises

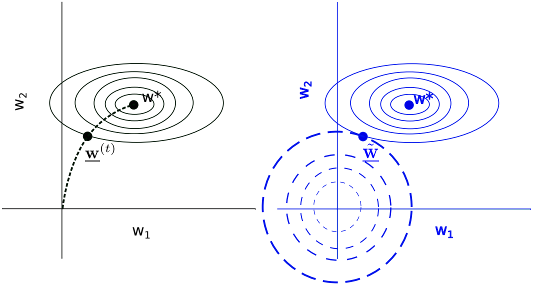
(trained on MNIST, Goodfellow et al., 2016)

Early stopping as regularization



- Is there a relationship between (unregularized) early stopping at time t and converged *weight-decay* regularization with constant α?

Early stopping as regularization



- Is there a relationship between (unregularized) early stopping at time t and converged *weight-decay* regularization with constant α?

- second order Taylor approximation around the minimum $\underline{\mathbf{w}}^*$

$$E^T_{(\underline{\mathbf{w}})} \quad \approx \quad E^T_{(\underline{\mathbf{w}}^*)} + \frac{1}{2}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*)^\top \underline{\mathbf{H}}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*) + \frac{\alpha}{2}\, \underline{\mathbf{w}}^\top \underline{\mathbf{w}}$$

$$t \approx \frac{1}{\eta\, \alpha}\,, \qquad\qquad \alpha \approx \frac{1}{\eta\, t}$$

derivation here

Choice of regularization parameter

Testset-Method

1. perform model selection for different values of λ (on training data)

train	test	validation

Choice of regularization parameter

Testset-Method

1. perform model selection for different values of λ (on training data)
2. select value of λ, which provides best prediction results (on test data)

train	test	validation

Choice of regularization parameter

Testset-Method

1. perform model selection for different values of λ (on training data)
2. select value of λ, which provides best prediction results (on test data)
3. estimate generalization performance of selected λ (on validation data)

train	test	validation

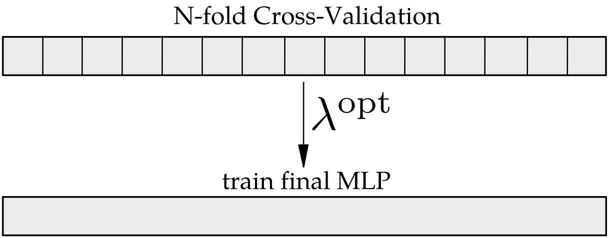
Choice of regularization parameter

n-fold cross-validation

for $\lambda = \lambda_1$ *TO* λ_n *DO* **do**
  perform n-fold cross-validation on data with regularization λ
end

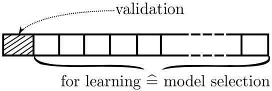
pick optimal λ^{opt} with minimum $\widehat{E}^G$

final model: train network on all data with λ^{opt}



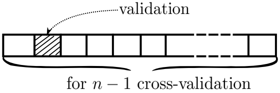
Validation

①



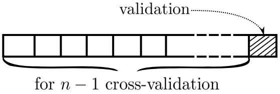
- do n - 1 cross-validation for all values of λ
- train with best λ
- validate with learned model

②



- do n - 1 cross-validation for all values of λ
- train with best λ
- validate with learned model

ⓝ



- do n - 1 cross-validation for all values of λ
- train with best λ
- validate with learned model

- $\widehat{E}^G \quad \widehat{=} \quad$ average over all n validation errors

- Never use test data for model selection (including hyper-parameter search).
- Always embed the whole selection procedure (including hyper-parameter search) within an n-fold cross-validation run.

1.4.7 Classification Problems (Multi-Class)

Inductive learning

data representation

$\downarrow$

model class

$\downarrow$

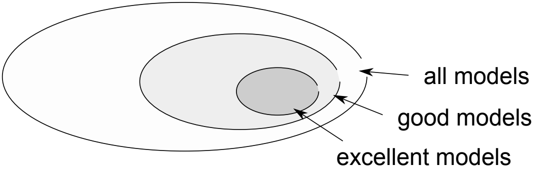
performance measure

$\downarrow$

optimization

$\downarrow$

validation



Data representation

Prediction of class labels

observations: $\left\{ \left(\underline{\mathbf{x}}^{(\alpha)}, y_T^{(\alpha)} \right) \right\}, \quad \alpha \in \{1, \ldots, p\}$

c classes $C_k, \quad k \in \{1, \ldots, c\} \quad \Rightarrow \quad y_T^{(\alpha)} \in \{C_1, \ldots, C_c\}$

Data representation

Prediction of class labels

observations: $\left\{\left(\underline{\mathbf{x}}^{(\alpha)}, y_T^{(\alpha)}\right)\right\}, \quad \alpha \in \{1, \ldots, p\}$

c classes $C_k, \quad k \in \{1, \ldots, c\} \quad \Rightarrow \quad y_T^{(\alpha)} \in \{C_1, \ldots, C_c\}$

Prediction of class probabilities

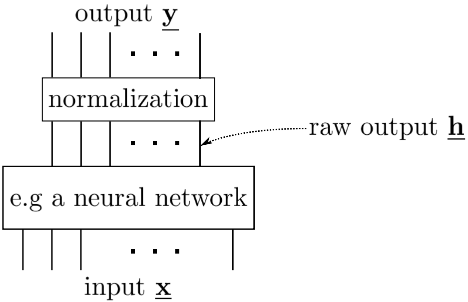
1-out-of-c-code

$$y_{Tk}^{(\alpha)} = \begin{cases} 0, & \underline{\mathbf{x}}^{(\alpha)} \notin C_k \\ 1, & \underline{\mathbf{x}}^{(\alpha)} \in C_k \end{cases}$$

$\Rightarrow$ binary vector $\underline{\mathbf{y}}_T^{(\alpha)}$, one non-zero element

Limiting case of probabilities: true labels are known.

Model class



probabilistic interpretation of network output:

$$y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} := P_{(C_k|\underline{\mathbf{x}};\underline{\mathbf{w}})}$$

$$0 \leq \quad y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} \quad \leq 1$$

$$\sum_{k=1}^{c} y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} \quad = 1$$

Softmax normalization

$$y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} = \frac{\exp(h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})})}{\sum_l \exp(h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})})}$$

Performance measure

True probability: $P_{(C_k|\underline{\mathbf{x}})}$ $\leftrightarrow$ Model prediction: $P_{(C_k|\underline{\mathbf{x}};\underline{\mathbf{w}})}$

Kullback-Leibler-Divergence D_{KL}

$$D_{KL}\left(P_{(C,\underline{\mathbf{x}})} \middle\| P_{(C,\underline{\mathbf{x}}|\underline{\mathbf{w}})}\right) = \sum_{k=1}^{c} \int d\underline{\mathbf{x}} \, \overbrace{P_{(\underline{\mathbf{x}})} P_{(C_k|\underline{\mathbf{x}})}}^{P_{(C,\underline{\mathbf{x}})}} \ln \left(\frac{\overbrace{P_{(C_k|\underline{\mathbf{x}})} \cdot P_{(\underline{\mathbf{x}})}}^{P_{(C,\underline{\mathbf{x}})}}}{\underbrace{P_{(C_k|\underline{\mathbf{x}};\underline{\mathbf{w}})} \cdot P_{(\underline{\mathbf{x}})}}_{P_{(C,\underline{\mathbf{x}}|\underline{\mathbf{w}})}}} \right)$$

- distance measure between probability distributions and densities
- non-negative: $D_{KL} = 0$ iff $P_{(C,\underline{\mathbf{x}})} \equiv P_{(C,\underline{\mathbf{x}}|\underline{\mathbf{w}})}$ (distributions are equal)
- asymmetric: $D_{KL}(p\|q)$ does not generally equal $D_{KL}(q\|p)$

Performance measure

True probability: $P_{(C_k|\underline{\mathbf{x}})}$ $\leftrightarrow$ Model prediction: $P_{(C_k|\underline{\mathbf{x}};\underline{\mathbf{w}})}$

Kullback-Leibler-Divergence D_{KL}

$$D_{KL}\Big(P_{(C,\underline{\mathbf{x}})}\Big\|P_{(C,\underline{\mathbf{x}}|\underline{\mathbf{w}})}\Big) = \sum_{k=1}^{c}\int d\underline{\mathbf{x}}\; \overbrace{P_{(\underline{\mathbf{x}})}P_{(C_k|\underline{\mathbf{x}})}}^{P_{(C,\underline{\mathbf{x}})}} \ln\Bigg(\frac{\overbrace{P_{(C_k|\underline{\mathbf{x}})} \cdot P_{(\underline{\mathbf{x}})}}^{P_{(C,\underline{\mathbf{x}})}}}{\underbrace{P_{(C_k|\underline{\mathbf{x}};\underline{\mathbf{w}})} \cdot P_{(\underline{\mathbf{x}})}}_{P_{(C,\underline{\mathbf{x}}|\underline{\mathbf{w}})}}}\Bigg)$$

- distance measure between probability distributions and densities
- non-negative: $D_{KL} = 0$ iff $P_{(C,\underline{\mathbf{x}})} \equiv P_{(C,\underline{\mathbf{x}}|\underline{\mathbf{w}})}$ (distributions are equal)
- asymmetric: $D_{KL}(p\|q)$ does not generally equal $D_{KL}(q\|p)$

$$D_{KL} = -\sum_{k=1}^{c}\int d\underline{\mathbf{x}} P_{(\underline{\mathbf{x}})}P_{(C_k|\underline{\mathbf{x}})}\ln P_{(C_k|\underline{\mathbf{x}};\underline{\mathbf{w}})} + \underbrace{\sum_{k=1}^{c}\int d\underline{\mathbf{x}}\, P_{(\underline{\mathbf{x}})}P_{(C_k|\underline{\mathbf{x}})}\ln P_{(C_k|\underline{\mathbf{x}})}}_{\text{independent of model parameters}}$$

Cross entropy

$$E^G \quad \equiv \quad -\sum_{k=1}^{c} \int d\underline{\mathbf{x}} \, \underbrace{P_{(\underline{\mathbf{x}})} P_{(C_k|\underline{\mathbf{x}})}}_{\text{unknown!}} \ln P_{(C_k|\underline{\mathbf{x}};\underline{\mathbf{w}})}$$

Cross entropy

$$E^G \quad \equiv \quad - \sum_{k=1}^{c} \int d\underline{\mathbf{x}} \, \underbrace{P_{(\underline{\mathbf{x}})} P_{(C_k|\underline{\mathbf{x}})}}_{\text{unknown!}} \ln P_{(C_k|\underline{\mathbf{x}};\underline{\mathbf{w}})}$$

- multinomial distribution

$$P_{(\underline{\mathbf{y}}_T|\underline{\mathbf{x}},\underline{\mathbf{w}})} = \prod_{k=1}^{c} \left(y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}\right)^{y_{Tk}} ,$$

$$y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} \geq 0 \, , \quad \forall k \, , \quad \sum_{k=1}^{c} y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} = 1 \, , \quad \forall \underline{\mathbf{x}} \, , \forall \underline{\mathbf{w}}$$

Cross entropy

$$E^G \equiv -\sum_{k=1}^{c} \int d\underline{\mathbf{x}} \underbrace{P_{(\underline{\mathbf{x}})} P_{(C_k|\underline{\mathbf{x}})}}_{\text{unknown!}} \ln P_{(C_k|\underline{\mathbf{x}};\underline{\mathbf{w}})}$$

- multinomial distribution

$$P_{(\underline{\mathbf{y}}_T|\underline{\mathbf{x}},\underline{\mathbf{w}})} = \prod_{k=1}^{c} (y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})})^{y_{Tk}},$$

$$y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} \geq 0, \quad \forall k, \quad \sum_{k=1}^{c} y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} = 1, \quad \forall \underline{\mathbf{x}}, \forall \underline{\mathbf{w}}$$

- mathematical expectation $\rightarrow$ empirical average over training set:

$$E^T = \frac{1}{p} \sum_{\alpha=1}^{p} e^{(\alpha)} = -\frac{1}{p} \sum_{\alpha=1}^{p} \sum_{k=1}^{c} y_{Tk}^{(\alpha)} \ln \left(y_{k(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} \right)$$

Optimization via gradient descent (on-line)

$$\Delta \underline{\mathbf{w}} = -\eta \frac{\partial e^{\alpha}}{\partial \underline{\mathbf{w}}}$$

$$\frac{\partial e^{(\alpha)}}{\partial \underline{\mathbf{w}}} = -\frac{\partial}{\partial \underline{\mathbf{w}}} \sum_{k=1}^{c} y_{Tk}^{(\alpha)} \ln \left(y_{k(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} \right) \qquad \text{with} \qquad y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} = \frac{\exp h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\sum_l \exp h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})}}$$

$=$ see blackboard...

Optimization via gradient descent (on-line)

$$\Delta \underline{\mathbf{w}} = -\eta \frac{\partial e^{\alpha}}{\partial \underline{\mathbf{w}}}$$

$$\frac{\partial e^{(\alpha)}}{\partial \underline{\mathbf{w}}} = -\frac{\partial}{\partial \underline{\mathbf{w}}} \sum_{k=1}^{c} y_{Tk}^{(\alpha)} \ln \left(y_{k(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} \right) \qquad \text{with} \qquad y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} = \frac{\exp h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\sum_l \exp h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})}}$$

$$= \sum_{k=1}^{c} \left(y_{k(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} - y_{Tk}^{(\alpha)} \right) \underbrace{\frac{\partial h_{k(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}}}_{\leadsto \text{backprop}}$$

Validation

- test-set method or n-fold cross-validation
- BUT: using the cross-entropy based cost

$$\hat{E}^G = \frac{1}{q}\sum_{\beta=1}^{q} e^{(\beta)} = -\frac{1}{q}\sum_{\beta=1}^{q}\sum_{k=1}^{c} y_{Tk}^{(\beta)} \ln\left(y_{k(\underline{\mathbf{x}}^{(\beta)};\underline{\mathbf{w}})}\right)$$

Prediction of class labels

- Decision costs
  - $\$_{ij}$: cost for choosing C_i when $\underline{\mathbf{x}}$ is of class C_j

	patient is sick	patient is not sick
prediction: sick	buy medicine (\$20)	adverse effects (\$100)
prediction: not sick	sick leave (\$1500)	everything is fine (\$0)

- Choose C_k with minimal prediction costs,
  i.e. $k = \underset{i}{\operatorname{argmin}} \sum_{j=1}^{c} \$_{ij} \; y_j(\underline{\mathbf{x}}; \underline{\mathbf{w}})$

End of Section 1.4

the following slides contain

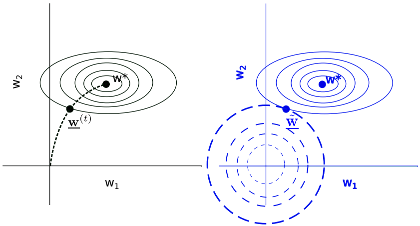
OPTIONAL MATERIAL

Early stopping as regularization

$$E^T_{(\underline{\mathbf{w}})} \quad \approx \quad E^T_{(\underline{\mathbf{w}}^*)} + \frac{1}{2}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*)^\top \underline{\mathbf{H}}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*)$$

$\underline{\mathbf{w}}^{(t)} - \underline{\mathbf{w}}^*$

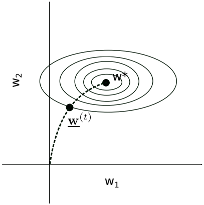
$=$

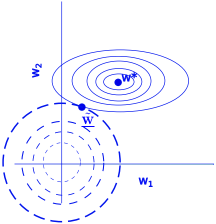


Early stopping as regularization

$$E^T_{(\underline{\mathbf{w}})} \quad \approx \quad E^T_{(\underline{\mathbf{w}}^*)} + \frac{1}{2}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*)^\top \underline{\mathbf{H}}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*)$$

$$\begin{aligned}
&\underline{\mathbf{w}}^{(t)} - \underline{\mathbf{w}}^* \\
&= \underline{\mathbf{w}}^{(t-1)} - \eta \frac{\partial E^T(\underline{\mathbf{w}}^{(t-1)})}{\partial \underline{\mathbf{w}}} \\
&= (\underline{\mathbf{I}} - \eta \underline{\mathbf{H}})(\underline{\mathbf{w}}^{(t-1)} - \underline{\mathbf{w}}^*) \\
&= -(\underline{\mathbf{I}} - \eta \underline{\mathbf{H}})^t \underline{\mathbf{w}}^*
\end{aligned}$$

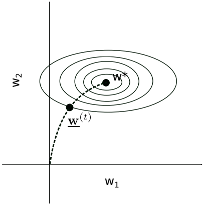


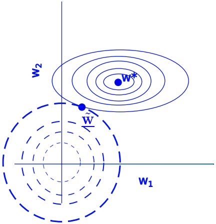


Early stopping as regularization

$$E^T_{(\underline{\mathbf{w}})} \quad \approx \quad E^T_{(\underline{\mathbf{w}}^*)} + \frac{1}{2}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*)^\top \underline{\mathbf{H}}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*) \; + \; \frac{\alpha}{2}\, \underline{\mathbf{w}}^\top \underline{\mathbf{w}}$$

$$\begin{aligned}
&\underline{\mathbf{w}}^{(t)} - \underline{\mathbf{w}}^* \\
&= \underline{\mathbf{w}}^{(t-1)} - \eta \frac{\partial E^T(\underline{\mathbf{w}}^{(t-1)})}{\partial \underline{\mathbf{w}}} \\
&= (\underline{\mathbf{I}} - \eta \underline{\mathbf{H}})(\underline{\mathbf{w}}^{(t-1)} - \underline{\mathbf{w}}^*) \\
&= -(\underline{\mathbf{I}} - \eta \underline{\mathbf{H}})^t \underline{\mathbf{w}}^*
\end{aligned}$$





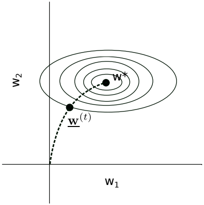
$$\frac{\partial E^T(\underline{\mathbf{w}}^{(t-1)})}{\partial \underline{\mathbf{w}}}$$

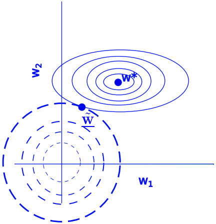
$$=$$

Early stopping as regularization

$$E^T_{(\underline{\mathbf{w}})} \quad \approx \quad E^T_{(\underline{\mathbf{w}}^*)} + \frac{1}{2}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*)^\top \underline{\mathbf{H}}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*) \; + \; \frac{\alpha}{2}\,\underline{\mathbf{w}}^\top \underline{\mathbf{w}}$$

$$\begin{aligned}
&\underline{\mathbf{w}}^{(t)} - \underline{\mathbf{w}}^* \\
&= \underline{\mathbf{w}}^{(t-1)} - \eta \frac{\partial E^T(\underline{\mathbf{w}}^{(t-1)})}{\partial \underline{\mathbf{w}}} \\
&= (\underline{\mathbf{I}} - \eta \underline{\mathbf{H}})(\underline{\mathbf{w}}^{(t-1)} - \underline{\mathbf{w}}^*) \\
&= -(\underline{\mathbf{I}} - \eta \underline{\mathbf{H}})^t \underline{\mathbf{w}}^*
\end{aligned}$$





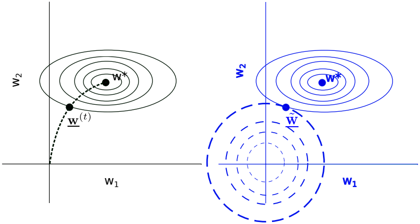
$$\frac{\partial E^T(\underline{\mathbf{w}}^{(t-1)})}{\partial \underline{\mathbf{w}}} = \underline{\mathbf{H}}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*) + \alpha\,\underline{\mathbf{w}} \overset{!}{=} 0$$

$$\Rightarrow \tilde{\underline{\mathbf{w}}} = (\underline{\mathbf{H}} + \alpha\,\underline{\mathbf{I}})^{-1}\underline{\mathbf{H}}\,\underline{\mathbf{w}}^*$$

Early stopping as regularization

$$E^T_{(\underline{\mathbf{w}})} \quad \approx \quad E^T_{(\underline{\mathbf{w}}^*)} + \frac{1}{2}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*)^\top \underline{\mathbf{H}}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*) \; + \; \frac{\alpha}{2}\, \underline{\mathbf{w}}^\top \underline{\mathbf{w}}$$

$$\begin{aligned} & \underline{\mathbf{w}}^{(t)} - \underline{\mathbf{w}}^* \\ &= \underline{\mathbf{w}}^{(t-1)} - \eta \frac{\partial E^T(\underline{\mathbf{w}}^{(t-1)})}{\partial \underline{\mathbf{w}}} \\ &= (\underline{\mathbf{I}} - \eta \underline{\mathbf{H}})(\underline{\mathbf{w}}^{(t-1)} - \underline{\mathbf{w}}^*) \\ &= -(\underline{\mathbf{I}} - \eta \underline{\mathbf{H}})^t \underline{\mathbf{w}}^* \end{aligned}$$



$$\begin{aligned} & \frac{\partial E^T(\underline{\mathbf{w}}^{(t-1)})}{\partial \underline{\mathbf{w}}} \\ &= \underline{\mathbf{H}}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*) + \alpha\, \underline{\mathbf{w}} \overset{!}{=} 0 \end{aligned}$$

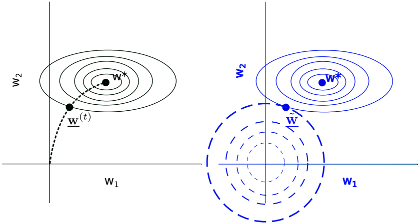
$$\Rightarrow \tilde{\underline{\mathbf{w}}} = (\underline{\mathbf{H}} + \alpha\, \underline{\mathbf{I}})^{-1} \underline{\mathbf{H}}\, \underline{\mathbf{w}}^*$$

$$\underline{\mathbf{w}}^{(t)} \overset{!}{=} \tilde{\underline{\mathbf{w}}} \quad \Rightarrow \quad \underline{\mathbf{I}} - (\underline{\mathbf{I}} - \eta \underline{\mathbf{H}})^t = (\underline{\mathbf{H}} + \alpha \underline{\mathbf{I}})^{-1} \underline{\mathbf{H}}$$

Early stopping as regularization

$$E^T_{(\underline{\mathbf{w}})} \approx E^T_{(\underline{\mathbf{w}}^*)} + \frac{1}{2}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*)^\top \underline{\mathbf{H}}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*) + \frac{\alpha}{2}\underline{\mathbf{w}}^\top \underline{\mathbf{w}}$$

$$\begin{aligned} & \underline{\mathbf{w}}^{(t)} - \underline{\mathbf{w}}^* \\ &= \underline{\mathbf{w}}^{(t-1)} - \eta \frac{\partial E^T(\underline{\mathbf{w}}^{(t-1)})}{\partial \underline{\mathbf{w}}} \\ &= (\underline{\mathbf{I}} - \eta \underline{\mathbf{H}})(\underline{\mathbf{w}}^{(t-1)} - \underline{\mathbf{w}}^*) \\ &= -(\underline{\mathbf{I}} - \eta \underline{\mathbf{H}})^t \underline{\mathbf{w}}^* \end{aligned}$$



$$\begin{aligned} & \frac{\partial E^T(\underline{\mathbf{w}}^{(t-1)})}{\partial \underline{\mathbf{w}}} \\ &= \underline{\mathbf{H}}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*) + \alpha\, \underline{\mathbf{w}} \stackrel{!}{=} 0 \\ \Rightarrow\ & \underline{\tilde{\mathbf{w}}} = (\underline{\mathbf{H}} + \alpha\, \underline{\mathbf{I}})^{-1} \underline{\mathbf{H}}\, \underline{\mathbf{w}}^* \end{aligned}$$

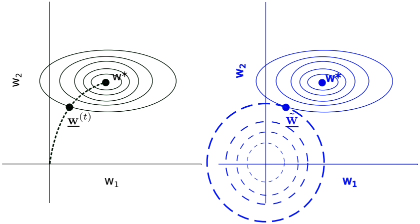
$$\underline{\mathbf{w}}^{(t)} \stackrel{!}{=} \underline{\tilde{\mathbf{w}}} \quad \Rightarrow \quad \underline{\mathbf{I}} - (\underline{\mathbf{I}} - \eta \underline{\mathbf{H}})^t = (\underline{\mathbf{H}} + \alpha \underline{\mathbf{I}})^{-1} \underline{\mathbf{H}}$$

$$\underline{\mathbf{H}} =: \underline{\mathbf{U}}\, \underbrace{\underline{\mathbf{\Lambda}}}_{\text{diagonal}}\, \underline{\mathbf{U}}^\top \quad \Rightarrow \quad \underline{\mathbf{U}}\Big(\underbrace{\underline{\mathbf{I}} - (\underline{\mathbf{I}} - \eta \underline{\mathbf{\Lambda}})^t}_{\text{diagonal}}\Big)\underline{\mathbf{U}}^\top = \underline{\mathbf{U}}\underbrace{(\underline{\mathbf{\Lambda}} + \alpha \underline{\mathbf{I}})^{-1} \underline{\mathbf{\Lambda}}}_{\text{diagonal}}\, \underline{\mathbf{U}}^\top$$

Early stopping as regularization

$$E^T_{(\underline{\mathbf{w}})} \approx E^T_{(\underline{\mathbf{w}}^*)} + \frac{1}{2}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*)^\top \underline{\mathbf{H}}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*) + \frac{\alpha}{2}\underline{\mathbf{w}}^\top \underline{\mathbf{w}}$$

$$\begin{aligned}
&\underline{\mathbf{w}}^{(t)} - \underline{\mathbf{w}}^* \\
&= \underline{\mathbf{w}}^{(t-1)} - \eta \frac{\partial E^T(\underline{\mathbf{w}}^{(t-1)})}{\partial \underline{\mathbf{w}}} \\
&= (\underline{\mathbf{I}} - \eta \underline{\mathbf{H}})(\underline{\mathbf{w}}^{(t-1)} - \underline{\mathbf{w}}^*) \\
&= -(\underline{\mathbf{I}} - \eta \underline{\mathbf{H}})^t \underline{\mathbf{w}}^*
\end{aligned}$$



$$\frac{\partial E^T(\underline{\mathbf{w}}^{(t-1)})}{\partial \underline{\mathbf{w}}} = \underline{\mathbf{H}}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*) + \alpha\, \underline{\mathbf{w}} \overset{!}{=} 0$$

$$\Rightarrow \tilde{\underline{\mathbf{w}}} = (\underline{\mathbf{H}} + \alpha\, \underline{\mathbf{I}})^{-1} \underline{\mathbf{H}}\, \underline{\mathbf{w}}^*$$

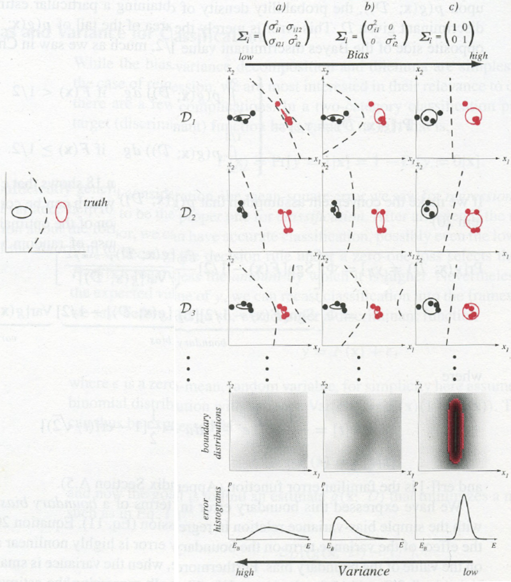
$$\underline{\mathbf{w}}^{(t)} \overset{!}{=} \tilde{\underline{\mathbf{w}}} \quad \Rightarrow \quad \underline{\mathbf{I}} - (\underline{\mathbf{I}} - \eta \underline{\mathbf{H}})^t = (\underline{\mathbf{H}} + \alpha \underline{\mathbf{I}})^{-1} \underline{\mathbf{H}}$$

$$\underline{\mathbf{H}} =: \underbrace{\underline{\mathbf{U}}\, \underline{\mathbf{\Lambda}}\, \underline{\mathbf{U}}^\top}_{\text{diagonal}} \quad \Rightarrow \quad \underline{\mathbf{U}}\Big(\underbrace{\underline{\mathbf{I}} - (\underline{\mathbf{I}} - \eta \underline{\mathbf{\Lambda}})^t}_{\text{diagonal}}\Big)\underline{\mathbf{U}}^\top = \underline{\mathbf{U}}\underbrace{(\underline{\mathbf{\Lambda}} + \alpha \underline{\mathbf{I}})^{-1} \underline{\mathbf{\Lambda}}}_{\text{diagonal}}\, \underline{\mathbf{U}}^\top$$

$$1 - (1 - \eta\lambda_i)^t = \frac{\lambda_i}{\lambda_i + \alpha} \quad \Rightarrow \quad t \approx \frac{1}{\eta(\lambda_i + \alpha)} \approx \frac{1}{\eta\, \alpha}, \quad \alpha \approx \frac{1}{\eta\, t}$$

using $\ln(1-x) \approx -x$ and assuming $\lambda_i \ll 1$

Classification example



- bias-variance trade-off also applies to classification ("boundary error")
- BUT – for classification: not additive & variance dominates

Gradient of softmax normalization

Network computation of probabilities

$$y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} = \frac{\exp(h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})})}{\sum_l \exp(h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})})} \qquad \text{(softmax function)}$$

Gradient of softmax normalization

Network computation of probabilities

$$y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} = \frac{\exp(h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})})}{\sum_l \exp(h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})})}$$

(softmax function)

$$\frac{\partial y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}} = \frac{\exp(h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}) \frac{\partial h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}} \sum_{l=1}^{c} \exp(h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})}) - \exp(h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}) \frac{\partial}{\partial \underline{\mathbf{w}}} \sum_{l=1}^{c} \exp(h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})})}{\left(\sum_{l=1}^{c} \exp(h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})}) \right)^2}$$

Gradient of softmax normalization

Network computation of probabilities

$$y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} = \frac{\exp\left(h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}\right)}{\sum_l \exp\left(h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})}\right)} \qquad \text{(softmax function)}$$

$$\begin{aligned}
\frac{\partial y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}} &= \frac{\exp(h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}) \frac{\partial h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}} \sum_{l=1}^{c} \exp(h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})}) - \exp(h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}) \frac{\partial}{\partial \underline{\mathbf{w}}} \sum_{l=1}^{c} \exp(h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})})}{\left(\sum_{l=1}^{c} \exp(h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})}) \right)^2} \\
&= y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} \frac{\partial h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}} - y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} \sum_{l=1}^{c} y_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})} \frac{\partial h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}}
\end{aligned}$$

Gradient of Cross Entropy

$$\frac{\partial e^{(\alpha)}}{\partial \underline{\mathbf{w}}} = -\frac{\partial}{\partial \underline{\mathbf{w}}} \sum_{k=1}^{c} y_{Tk}^{(\alpha)} \ln \left(y_{k(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} \right)$$

Gradient of Cross Entropy

$$\frac{\partial e^{(\alpha)}}{\partial \underline{\mathbf{w}}} \quad = \quad -\frac{\partial}{\partial \underline{\mathbf{w}}} \sum_{k=1}^{c} y_{Tk}^{(\alpha)} \ln \left(y_{k(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} \right) \qquad \text{with} \qquad y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} = \frac{\exp h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\sum_l \exp h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})}}$$

Gradient of Cross Entropy

$$\begin{aligned}
\frac{\partial e^{(\alpha)}}{\partial \underline{\mathbf{w}}} &= -\frac{\partial}{\partial \underline{\mathbf{w}}} \sum_{k=1}^{c} y_{Tk}^{(\alpha)} \ln\left(y_{k(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})}\right) \\
&= -\sum_{k=1}^{c} \frac{y_{Tk}}{y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}} \cdot \frac{\partial y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}}
\end{aligned}$$

with $$y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} = \frac{\exp h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\sum_l \exp h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})}}$$

see derivation on Slide 4

Gradient of Cross Entropy

$$\frac{\partial e^{(\alpha)}}{\partial \underline{\mathbf{w}}} = -\frac{\partial}{\partial \underline{\mathbf{w}}} \sum_{k=1}^{c} y_{Tk}^{(\alpha)} \ln \left(y_{k(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} \right) \qquad \text{with} \qquad y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} = \frac{\exp h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\sum_l \exp h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})}}$$

$$= -\sum_{k=1}^{c} \frac{y_{Tk}}{y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}} \cdot \frac{\partial y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}}$$

see derivation on Slide 4

$$= -\sum_{k=1}^{c} \frac{y_{Tk}}{y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}} \left(y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} \frac{\partial h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}} - y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} \sum_{l=1}^{c} y_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})} \frac{\partial h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}} \right)$$

Gradient of Cross Entropy

$$\frac{\partial e^{(\alpha)}}{\partial \underline{\mathbf{w}}} = -\frac{\partial}{\partial \underline{\mathbf{w}}} \sum_{k=1}^{c} y_{Tk}^{(\alpha)} \ln\left(y_{k(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})}\right) \qquad \text{with} \qquad y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} = \frac{\exp h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\sum_l \exp h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})}}$$

$$= -\sum_{k=1}^{c} \frac{y_{Tk}}{y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}} \cdot \frac{\partial y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}}$$

see derivation on Slide 4

$$= -\sum_{k=1}^{c} \frac{y_{Tk}}{y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}} \left(y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} \frac{\partial h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}} - y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} \sum_{l=1}^{c} y_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})} \frac{\partial h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}} \right)$$

$$= -\sum_{k=1}^{c} y_{Tk} \frac{\partial h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}} + \underbrace{\left(\sum_{k=1}^{c} y_{Tk} \right)}_{=1} \left(\sum_{l=1}^{c} y_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})} \frac{\partial h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}} \right)$$

Gradient of Cross Entropy

$$\frac{\partial e^{(\alpha)}}{\partial \underline{\mathbf{w}}} = -\frac{\partial}{\partial \underline{\mathbf{w}}} \sum_{k=1}^{c} y_{Tk}^{(\alpha)} \ln \left(y_{k(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} \right) \qquad \text{with} \qquad y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} = \frac{\exp h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\sum_l \exp h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})}}$$

$$= -\sum_{k=1}^{c} \frac{y_{Tk}}{y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}} \cdot \frac{\partial y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}}$$

see derivation on Slide 4

$$= -\sum_{k=1}^{c} \frac{y_{Tk}}{y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}} \left(y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} \frac{\partial h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}} - y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} \sum_{l=1}^{c} y_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})} \frac{\partial h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}} \right)$$

$$= -\sum_{k=1}^{c} y_{Tk} \frac{\partial h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}} + \underbrace{\left(\sum_{k=1}^{c} y_{Tk} \right)}_{=1} \left(\sum_{l=1}^{c} y_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})} \frac{\partial h_{l(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}} \right)$$

$$= \sum_{k=1}^{c} \left(y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})} - y_{Tk} \right) \underbrace{\frac{\partial h_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}}{\partial \underline{\mathbf{w}}}}_{\rightsquigarrow \text{backprop}}$$