Introduction to Machine Learning

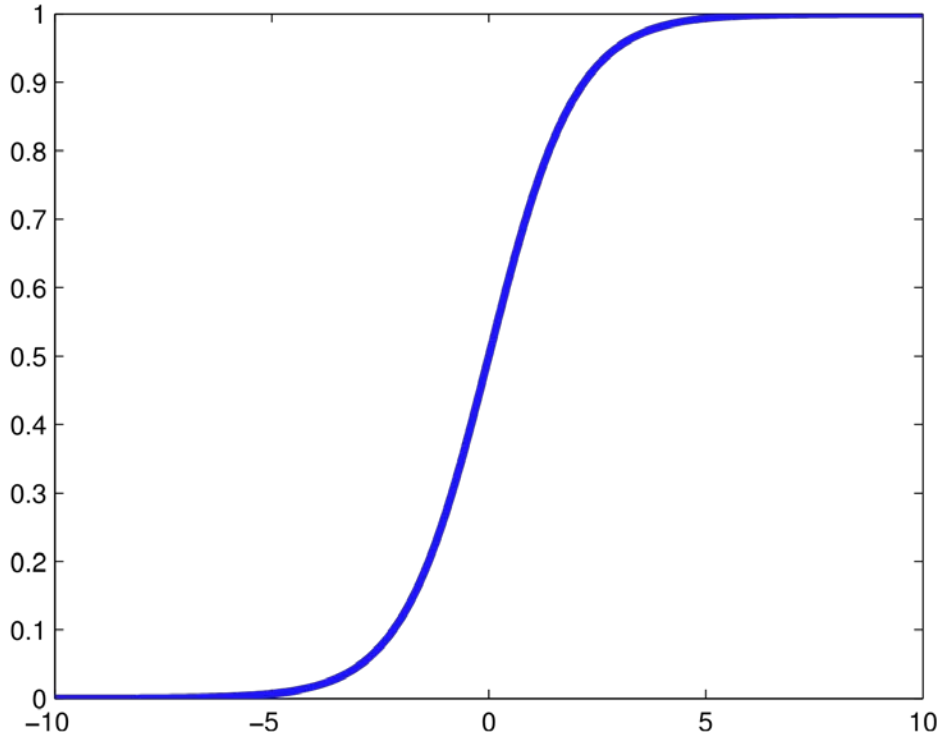
Brown University CSCI 1950-F, Spring 2011
Prof. Erik Sudderth

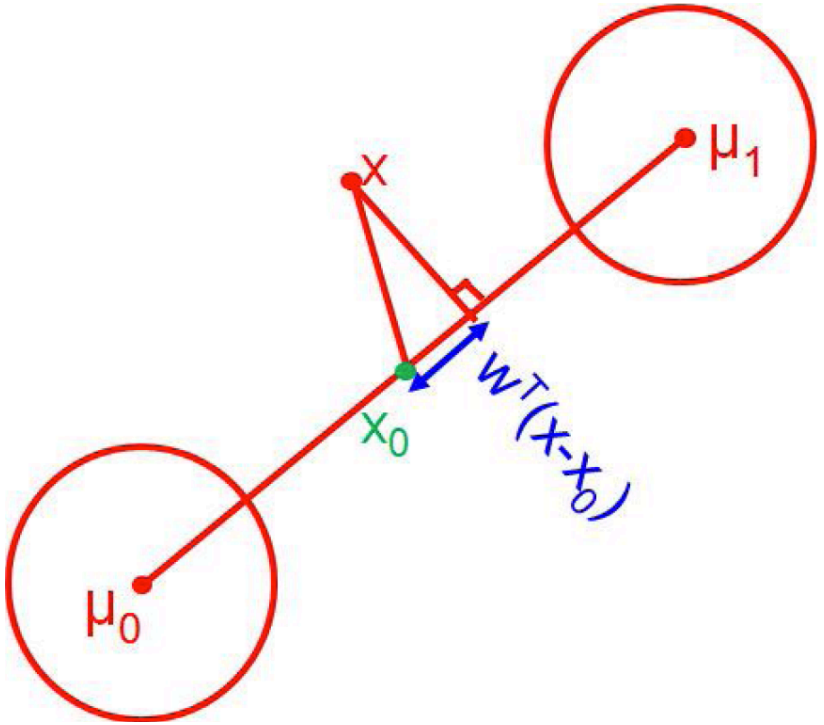
Lecture 10: Logistic Regression, Gradient-Based Optimization

Many figures courtesy Kevin Murphy's textbook,
Machine Learning: A Probabilistic Perspective

Binary Discriminant Analysis

Logistic Function





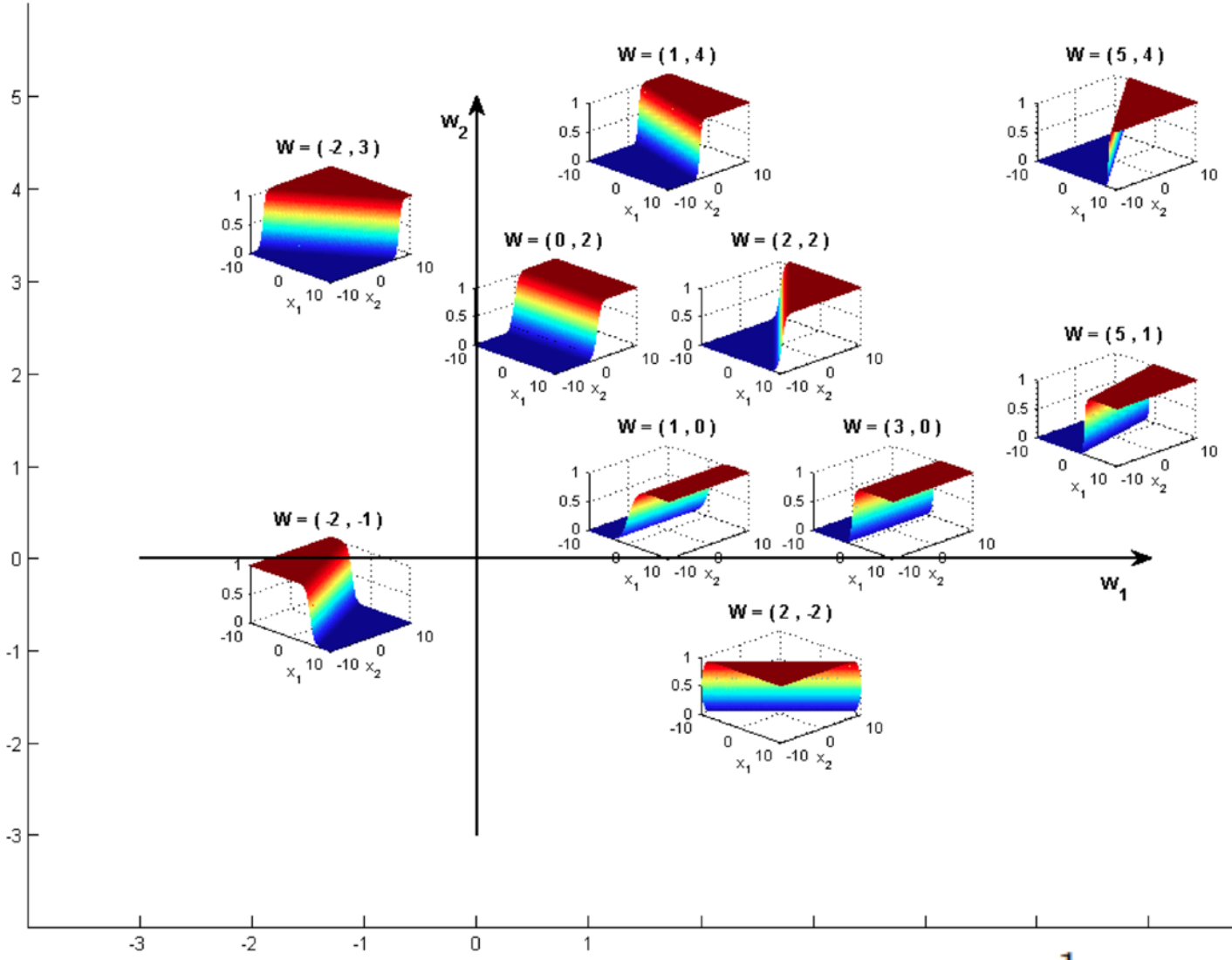
$$\text{sigm}(\eta) := \frac{1}{1+\exp(-\eta)} = \frac{e^{\eta}}{e^{\eta}+1}$$

$$p(y=1|\mathbf{x},\boldsymbol{\theta}) = \sigma(\mathbf{w}^T(\mathbf{x}-\mathbf{x}_0))$$

$$\mathbf{w} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0 = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$$

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\frac{\log(\pi_1/\pi_0)}{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}$$

Logistic Regression

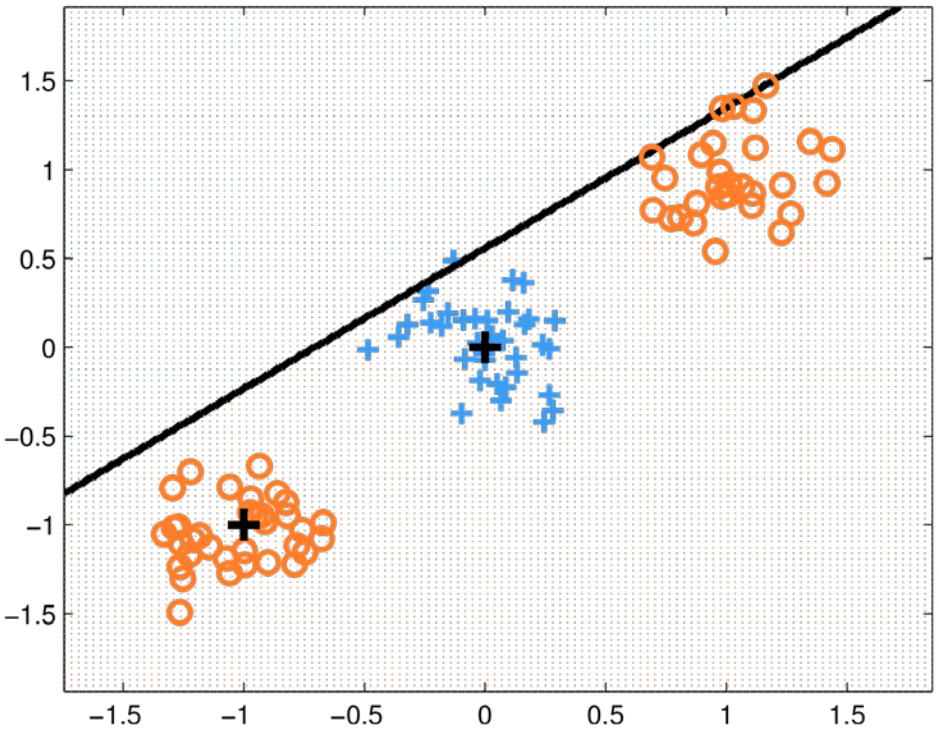


$$p(y|\mathbf{x}, \mathbf{w}) = \mathrm{Ber}(y|\mathrm{sigm}(\mathbf{w}^T\mathbf{x}))$$

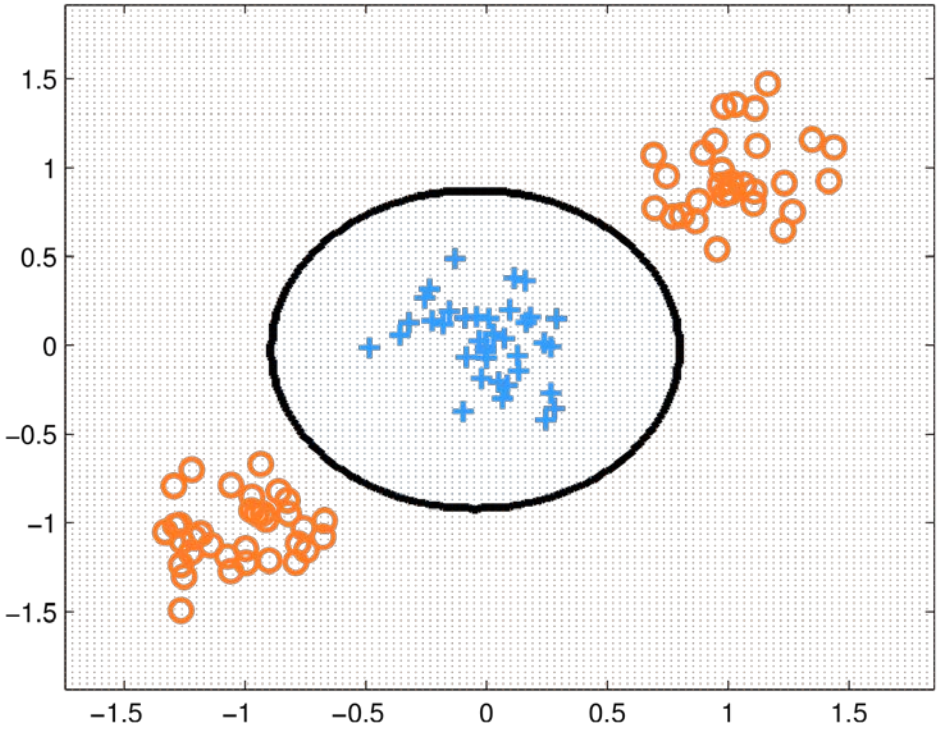
$$\mathrm{sigm}(\eta) := \frac{1}{1+\exp(-\eta)} = \frac{e^{\eta}}{e^{\eta}+1}$$

Decision Boundaries

By assumption, a linear function of input features.

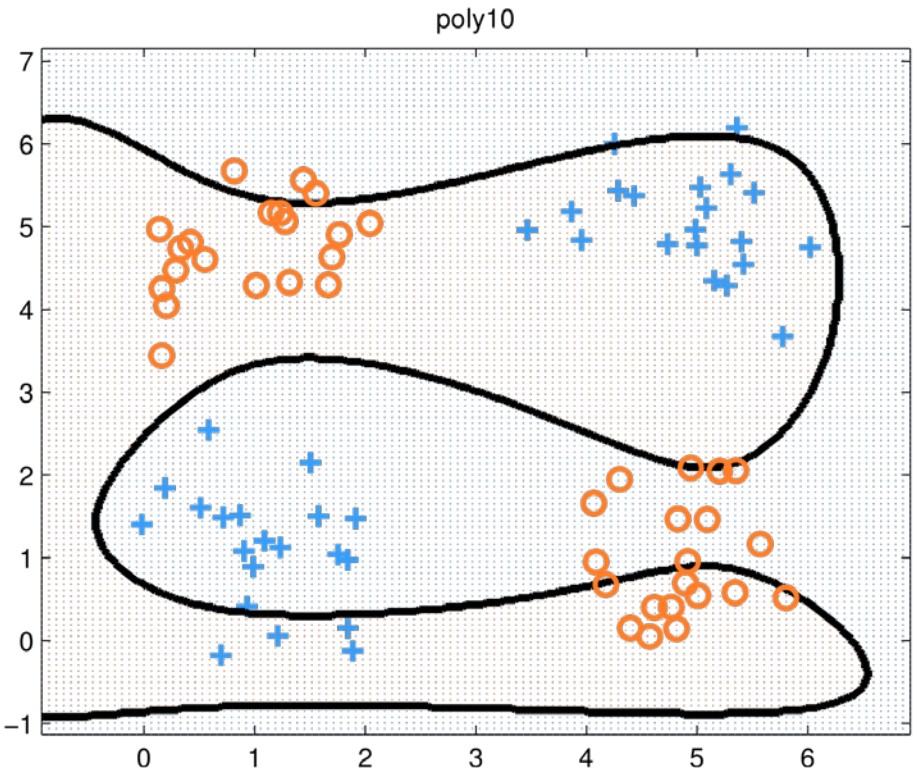


Raw features.

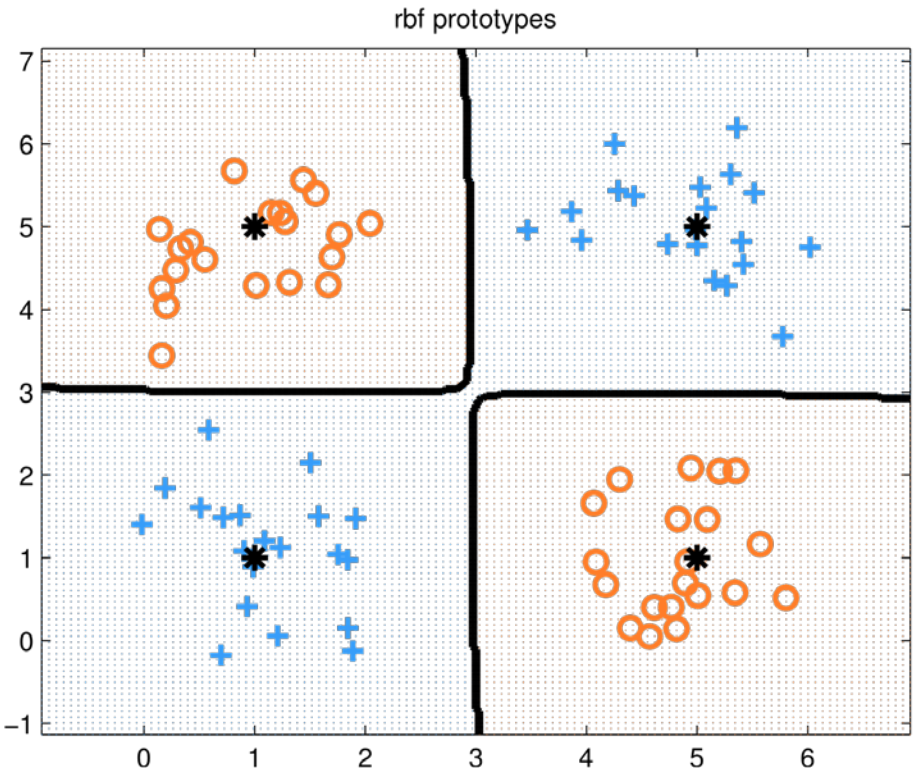


2nd-order polynomial expansion.

Importance of Features: XOR



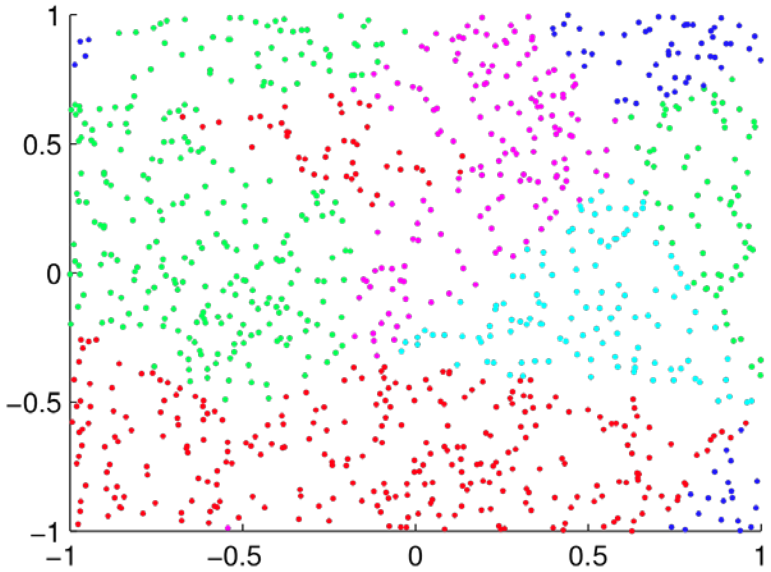
10^{th}-order polynomial.



4 radial basis functions.

Avoid RBF placement via kernel methods…

Multinomial Logistic Regression



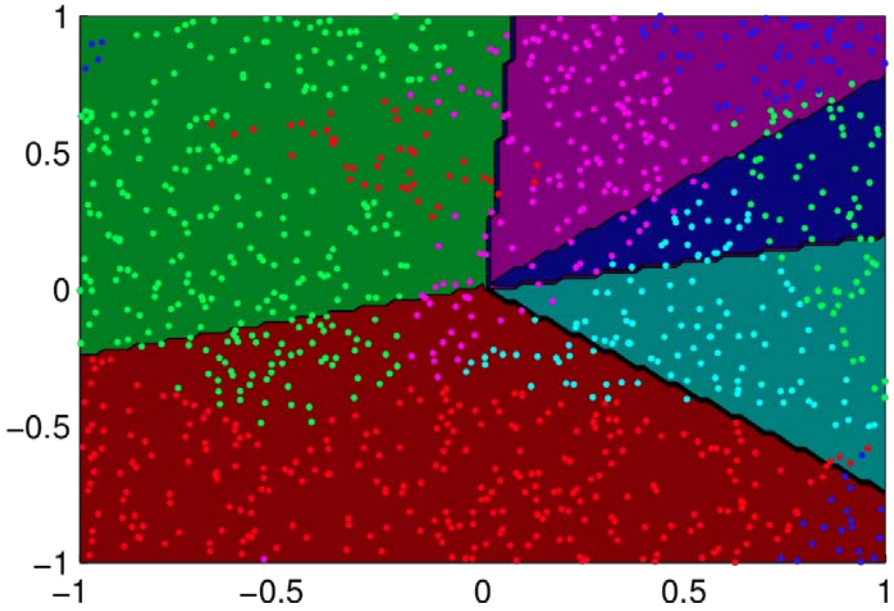
$$p(y|\mathbf{x}, \mathbf{W}) = \mathrm{Cat}(y|\mathcal{S}(\mathbf{W}^T\mathbf{x}))$$

$$\mathcal{S}(\eta)_c = \frac{e^{\eta_c}}{\sum_{c'=1}^{C} e^{\eta_{c'}}}$$

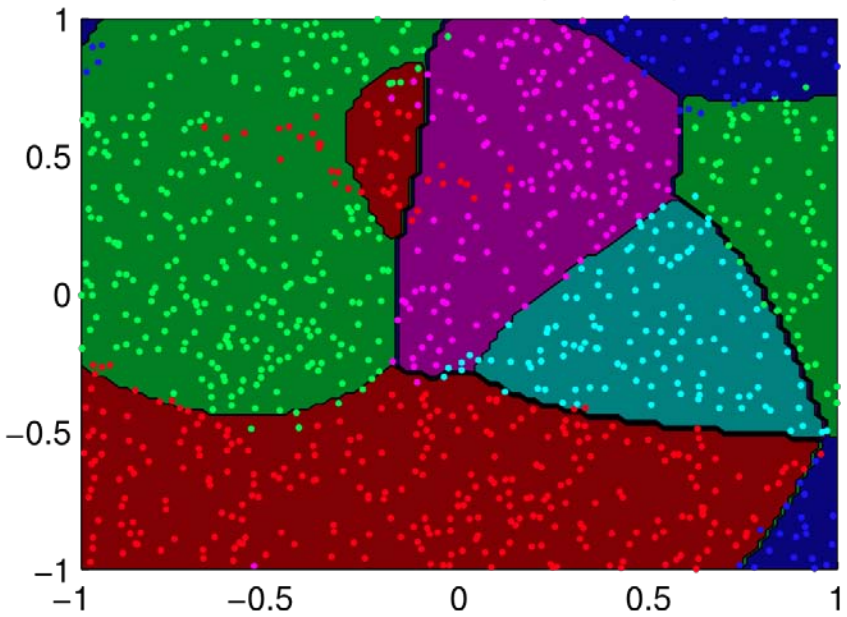
as $T \to 0$

$$\mathcal{S}(\eta/T)_c = \begin{cases} 1.0 & \text{if } c = \arg\max_{c'} \eta_{c'} \\ 0.0 & \text{otherwise} \end{cases}$$

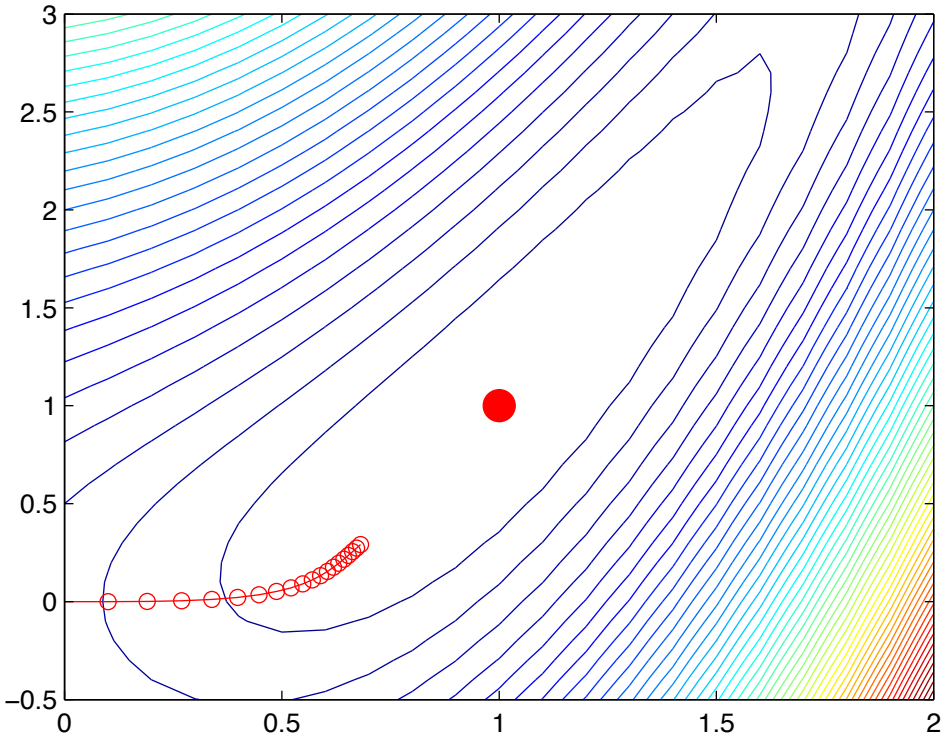
Linear Multinomial Logistic Regression



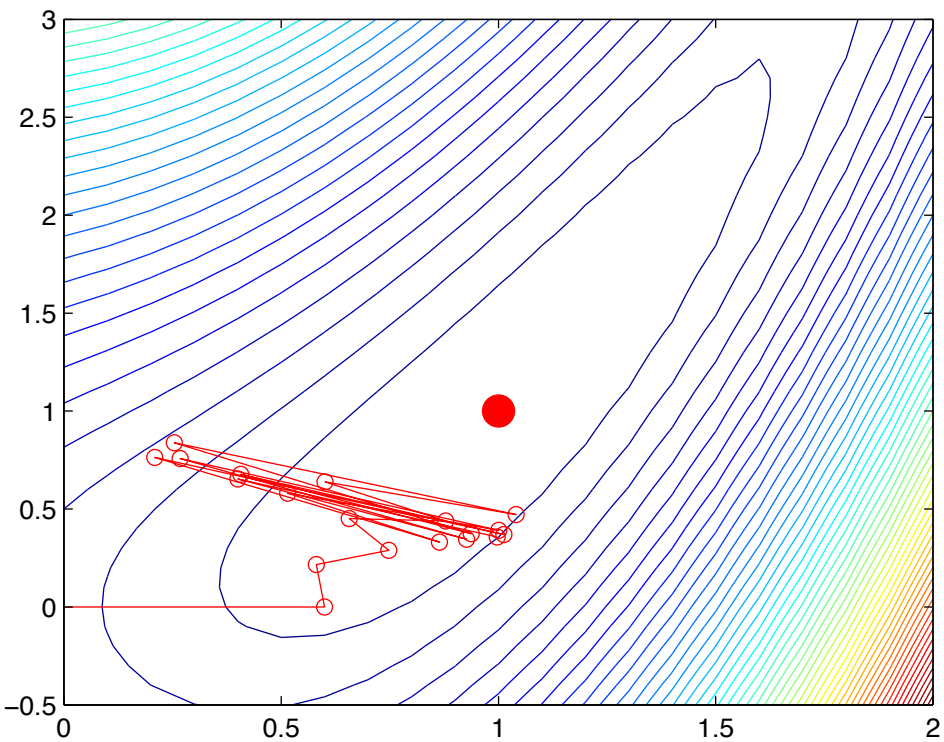
Kernel−RBF Multinomial Logistic Regression



Gradient (Steepest) Descent

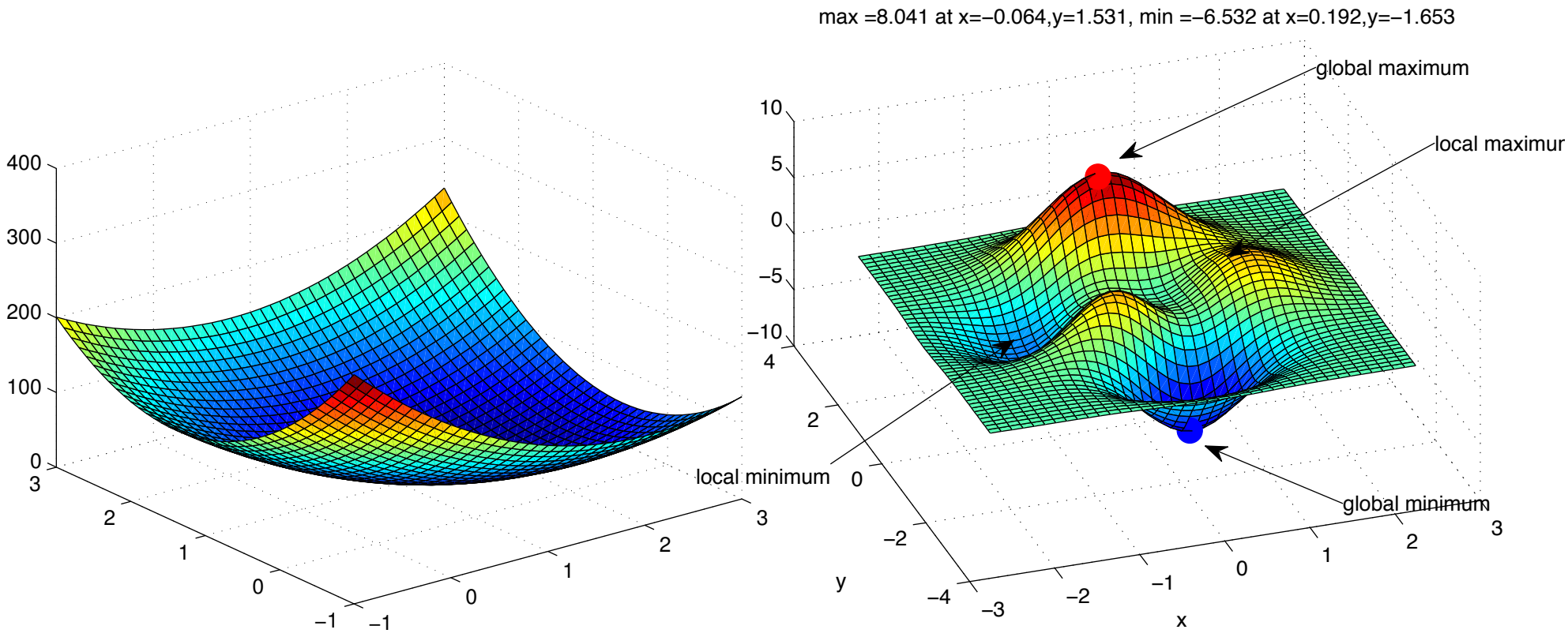


stepsize = 0.1



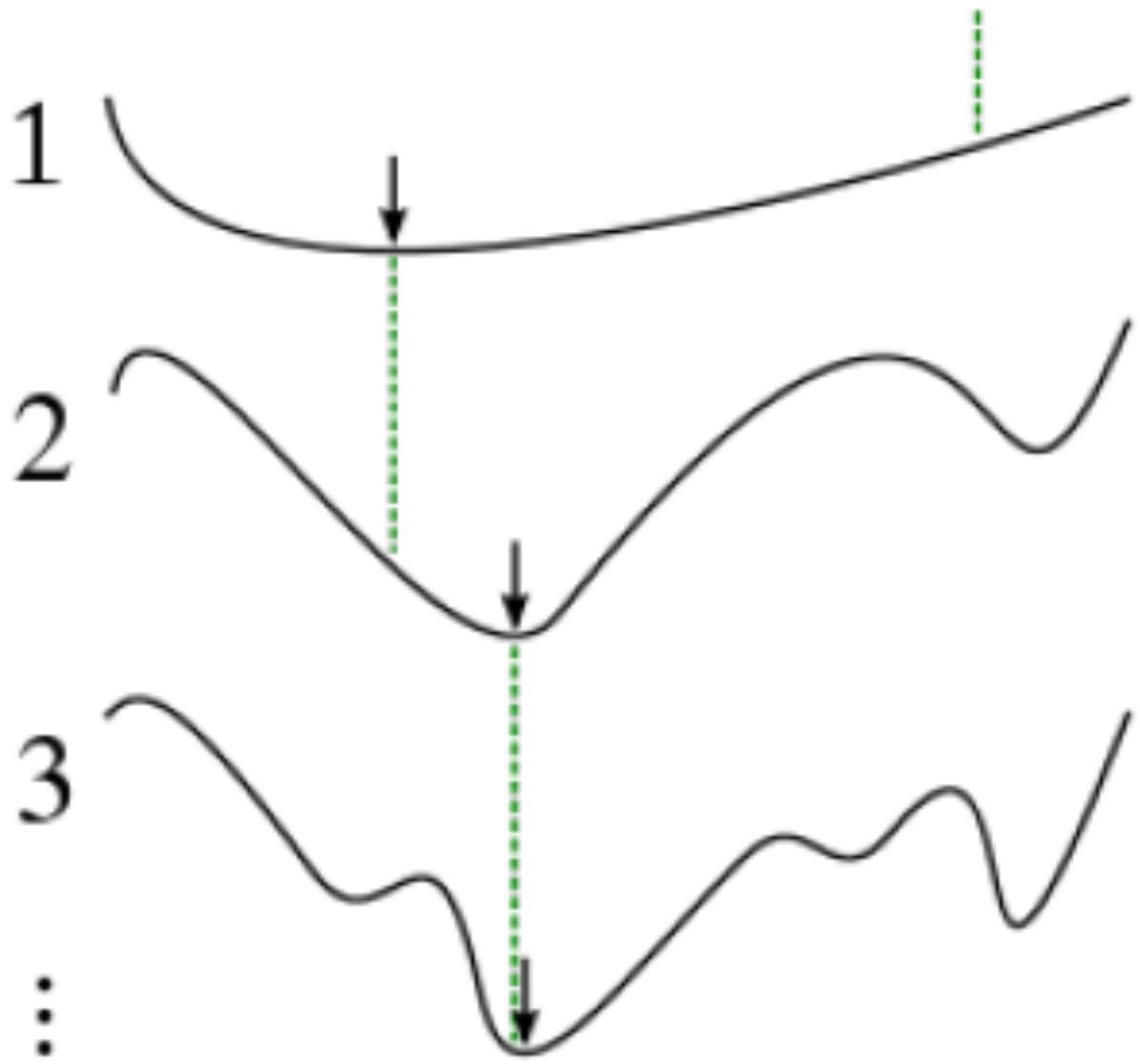
stepsize = 0.6

Global & Local Optima

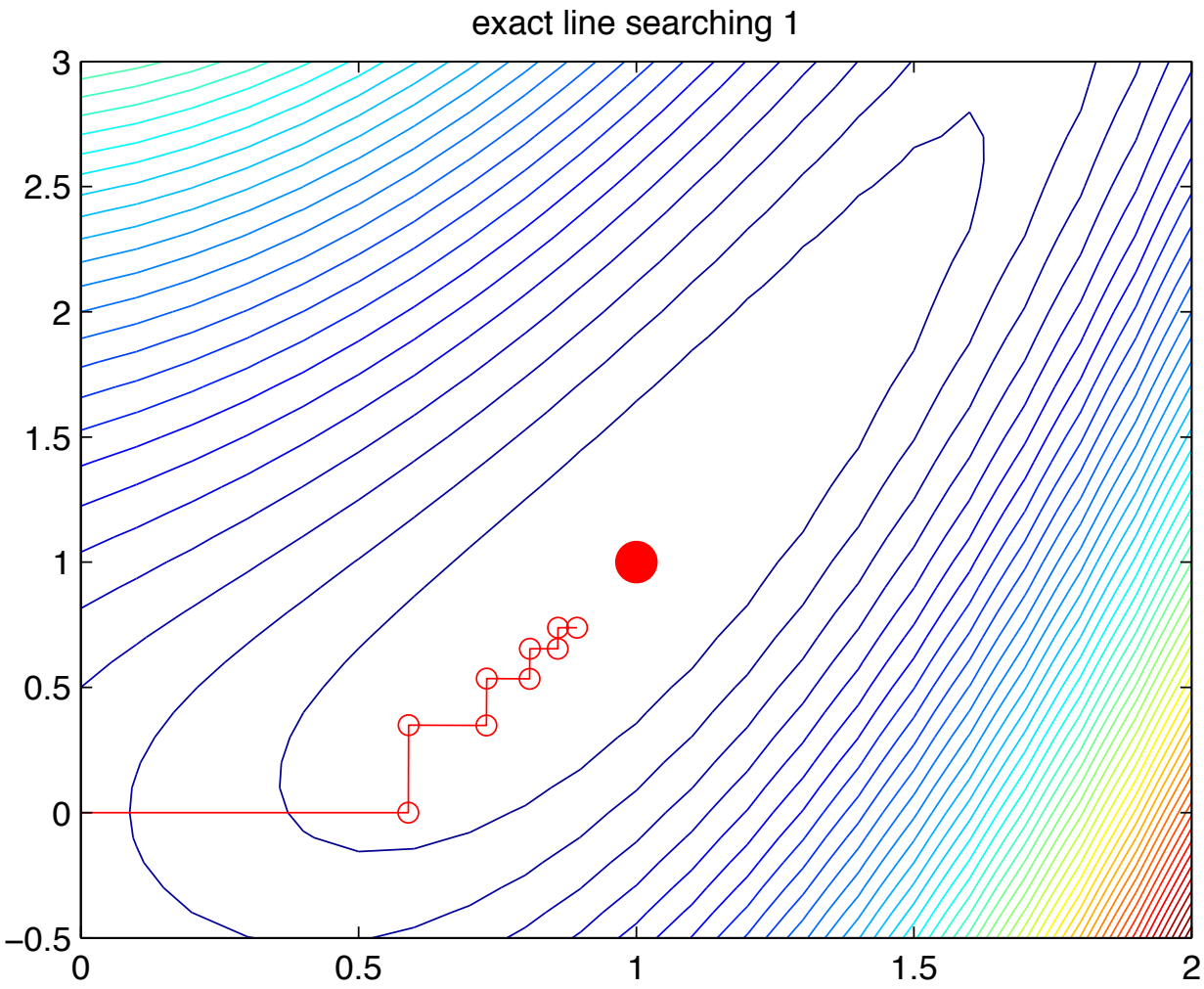


- A *globally convergent* algorithm is guaranteed to converge to a *local optimum* from any initialization
- Convergence to a *global optimum* is another issue…

Deterministic Annealing



Descent via Line Search



Newton's Method