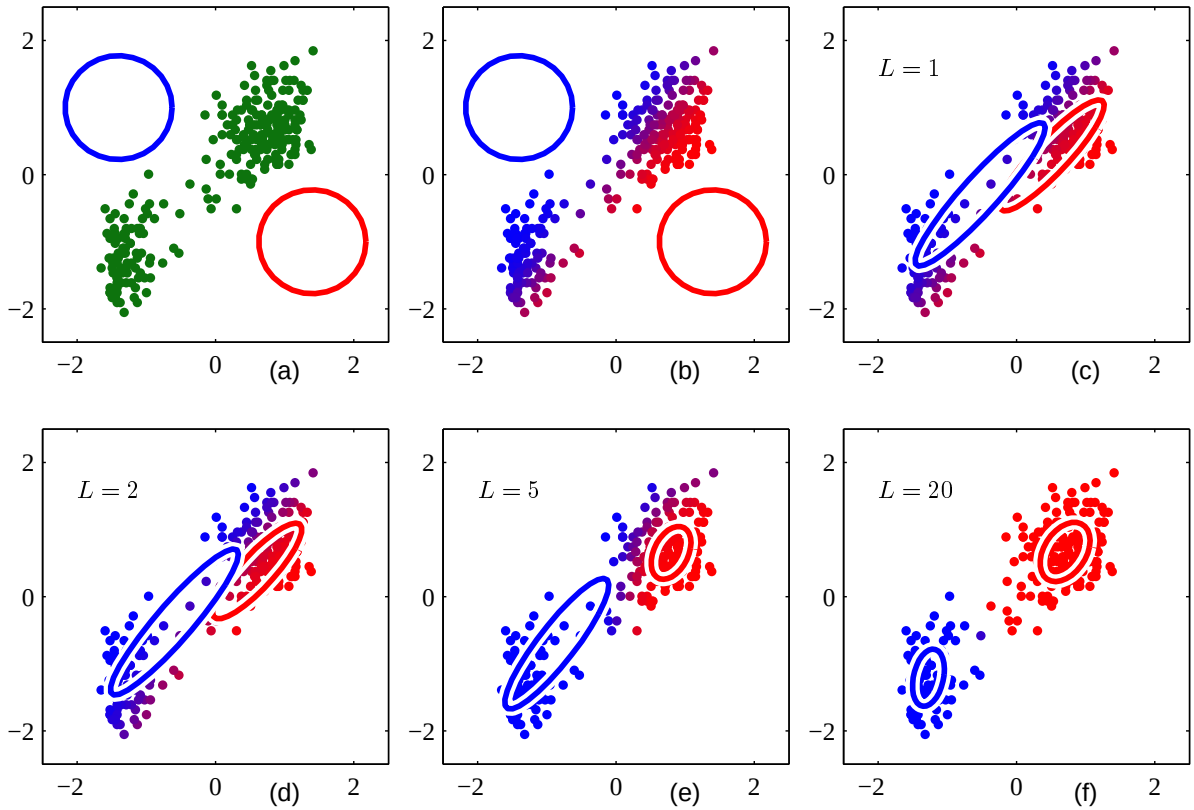
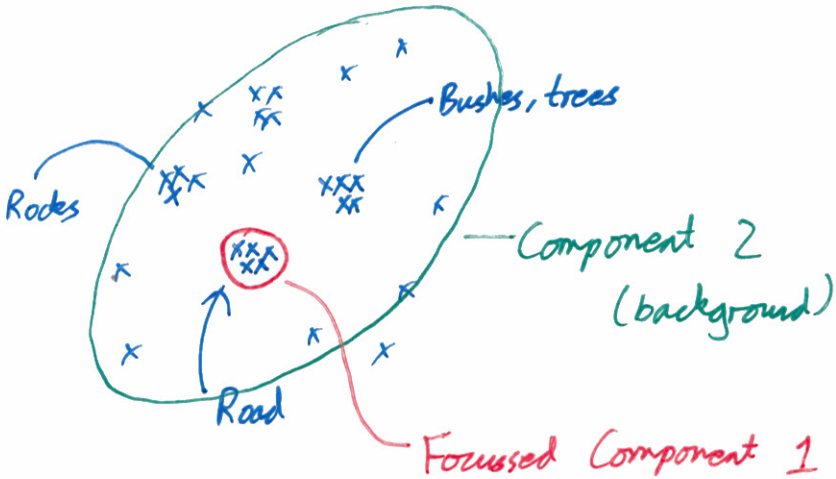
EM algorithm for Gaussian mixtures



Bishop Figure 9.8, or see Murphy p353


Bushes, trees
Rocks
Component 2
(background)
Road
Focussed Component 1

Newton's Method

Cost function $E(\underline{w})$

Gradients $\underline{g} = \nabla_{\underline{w}} E(\underline{w})$ $\quad g_d = \frac{\partial E}{\partial w_d}$

Hessian $H_{ij} = \frac{\partial^2 E}{\partial w_i \partial w_j}$

Initialize $\underline{w}^{(0)}$

$$\underline{w}^{(t+1)} = \underline{w}^{(t)} - H^{-1}\underline{g}$$

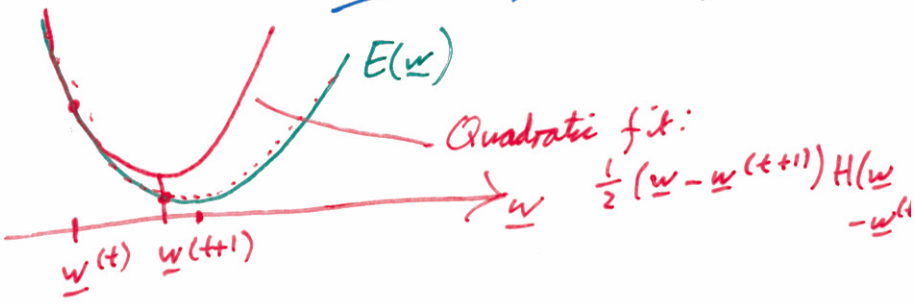
If we had a quadratic cost function:

$$E(\underline{w}) = \frac{1}{2}(\underline{w} - \underline{w}^*)^T H (\underline{w} - \underline{w}^*) + \text{const.}$$

Here: $\underline{g} = H(\underline{w} - \underline{w}^*)$

$$\underline{w}^{(t+1)} = \underline{w}^{(t)} - H^{-1}H(\underline{w}^{(t)} - \underline{w}^*)$$

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



$$\frac{1}{2}(\underline{w} - \underline{w}^{(t+1)}) H(\underline{w} - \underline{w}^{(t}$$

If w is 50,000 parameters:

$8 \times (50 \times 10^3)^2 / 10^9 = 20$ GB RAM

store Hessian

Ref: "Hessian free" versions of Newton's method.

Other methods: L-BFGS

Non-linear Conjugate gradients

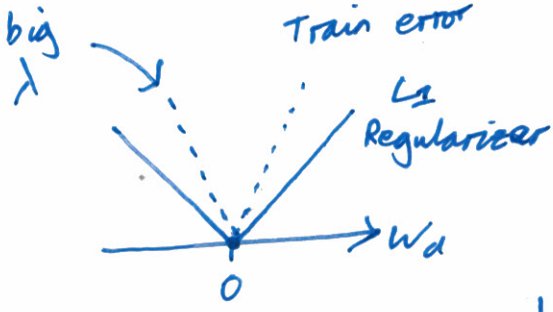
Another reason to find other optimizers

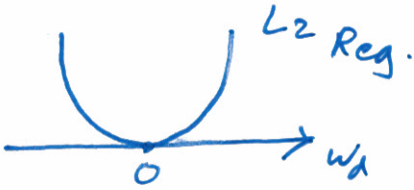
Fit "sparse" solutions, some weights w_d are exactly zero.

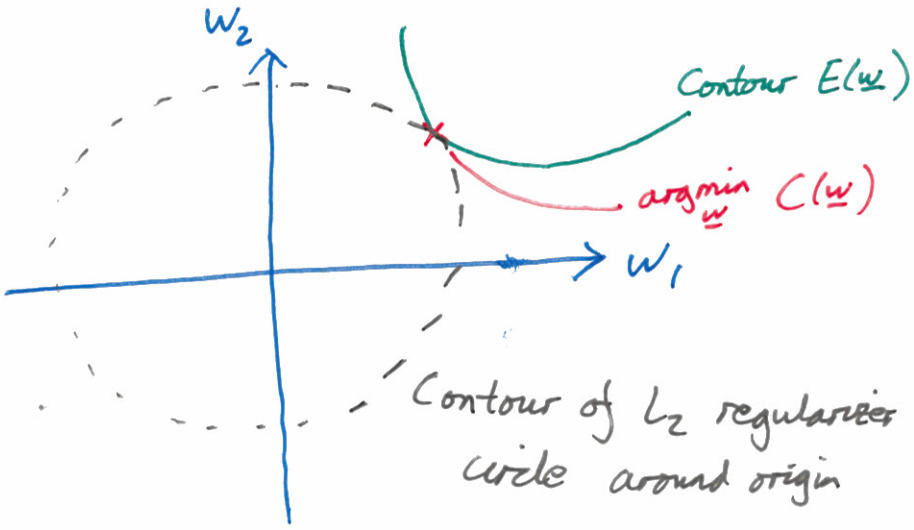
L1 Regularization

$$C(\underline{w}) = E(\underline{w}) + \underbrace{\lambda \Sigma_d |w_d|}_{\lambda \|\underline{w}\|_1}$$

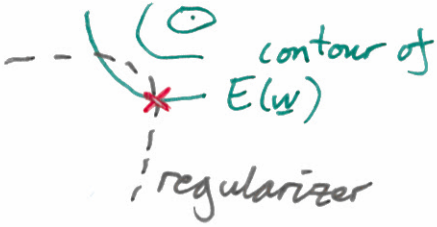
$E(\underline{w})$: Train error

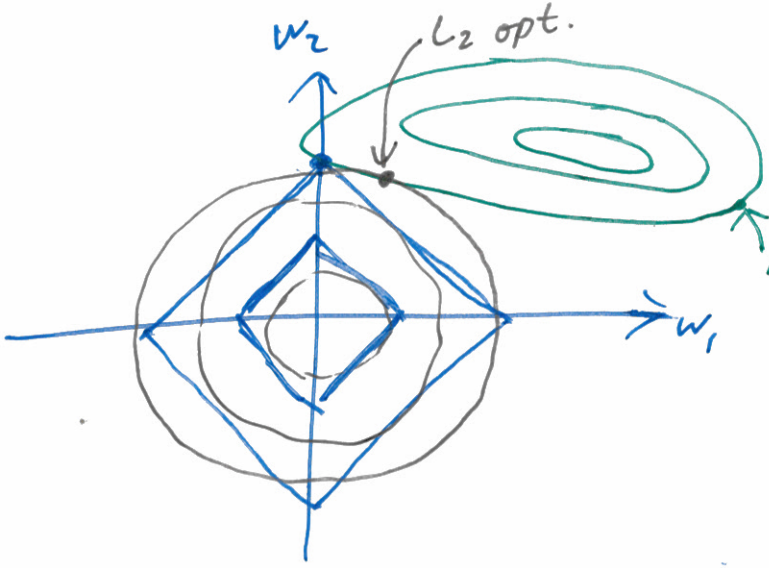






Contours don't cross at optimum.



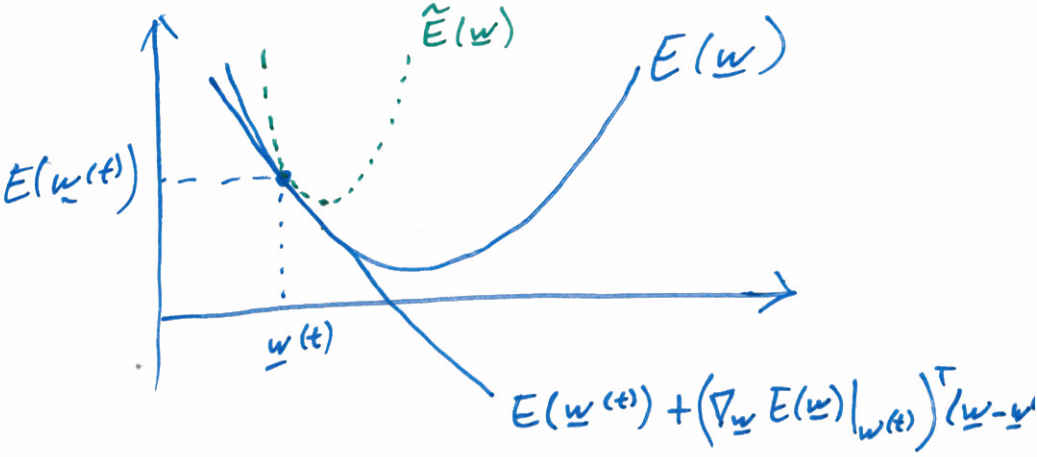


From a Bayesian view, we can't ignore features.

prediction

$$p(y \mid \underline{x}, D) = \int p(y \mid \underline{x}, \underline{w})\, p(\underline{w} \mid D)\, d\underline{w}$$

we end up using all the features.



1D Taylor series

$$f(x+z) \approx f(x) + z f'(x) + \frac{z^2}{2} f''(x) \cdots$$

z: small. $f'(x)$: $\frac{\partial f}{\partial x}$

$$\approx f(\underline{x}) + \underline{z}^T \nabla_{\underline{x}} f(\underline{x}) + (\ ) H (\ ) \cdots$$