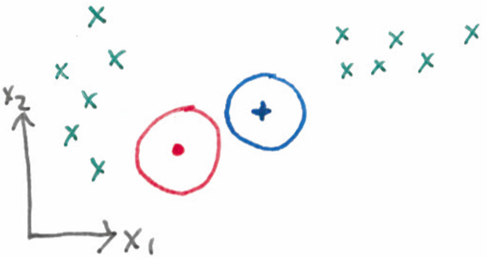
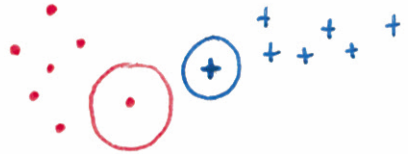
EM for Mixtures of Gaussians



0. Initialize parameters

$$\theta = \{ \underline{\pi}, \{ \underline{\mu}^{(k)}, \Sigma^{(k)} \} \}$$



1. E-Step

Fix soft responsibilities:

$$r_k^{(n)} = P(z^{(n)} = k \mid \underline{x}^{(n)}, \theta)$$

2. M-Step

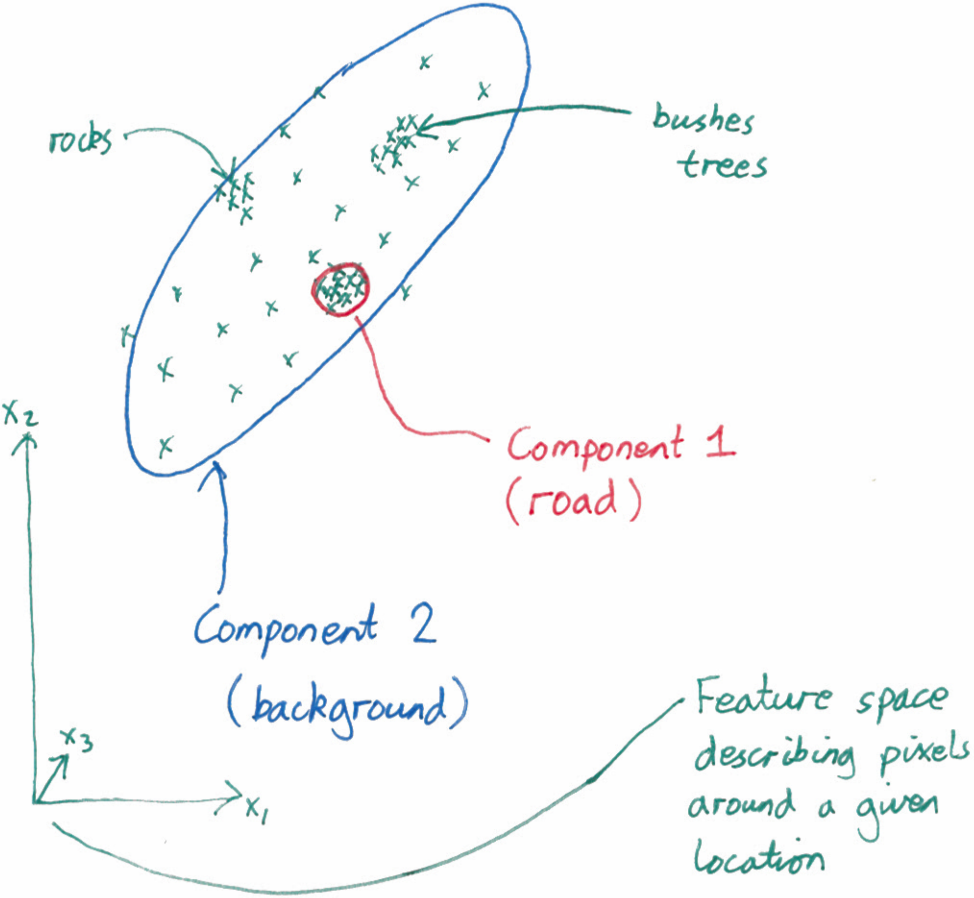
Fit θ using $\{r_k^{(n)}\}$

$$r_k = \sum_n r_k^{(n)}, \quad \pi_k = \frac{r_k}{N}, \quad \underline{\mu}^{(k)} = \frac{1}{r_k} \sum_n r_k^{(n)} \underline{x}^{(n)}$$

$$\Sigma^{(k)} = \frac{1}{r_k} \sum_n r_k^{(n)} (\underline{x}^{(n)} - \underline{\mu}^{(k)})(\underline{x}^{(n)} - \underline{\mu}^{(k)})$$

3. Go to 1. or stop.

Use in "Stanley" car



Interpretation

We are using an approx. posterior

$$Q(z^{(n)} = k) = r_k^{(n)} = P(z^{(n)} = k \mid \underline{x}^{(n)}, \theta^{(old)})$$

Only correct when $\theta = \theta^{(old)}$

Compare distributions

$$D_{KL}(Q \parallel P) = \sum_{\underline{z}} Q(\underline{z}) \log \frac{Q(\underline{z})}{P(\underline{z} \mid X, \theta)} \geq 0$$

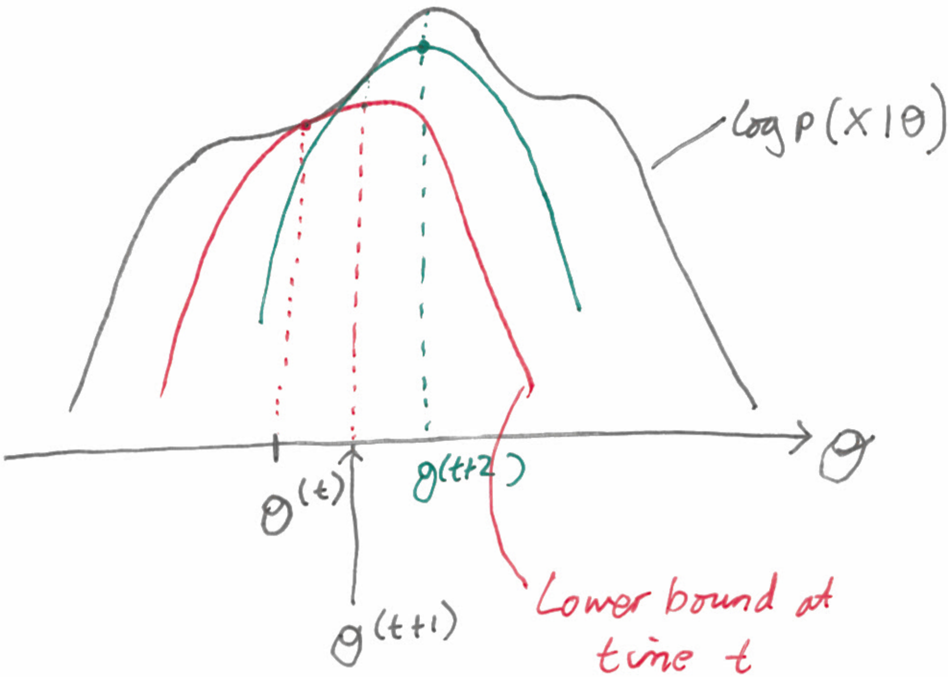
Use $p(\underline{z} \mid X, \theta) = \dfrac{p(\underline{z}, X \mid \theta)}{p(X \mid \theta)}$ ← Likelihood

$$\Rightarrow \sum_{\underline{z}} Q(\underline{z}) \log \frac{Q(\underline{z})}{P(X, \underline{z} \mid \theta)} \geq -\log p(X \mid \theta)$$

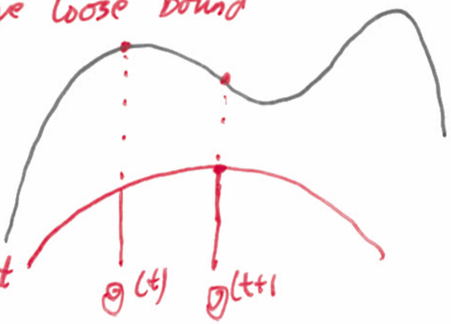
⇒ Bound on log-likelihood

Tight if $\theta = \theta^{(old)}$ used to set Q

Bound-based optimizer



Can't have loose bound if want to guarantee improvement



Newton's Method

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

Gradients $\underline{g} = \nabla_{\underline{w}} E(\underline{w})$

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

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

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

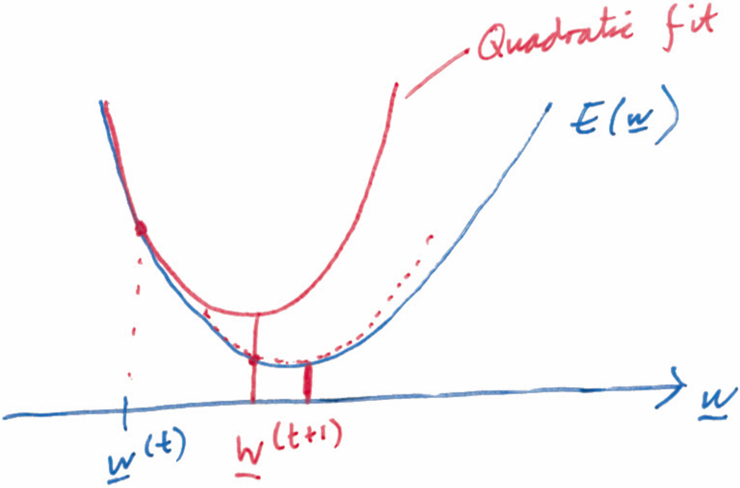
If we have a quadratic cost:

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

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

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

$$= \underline{w}^{(*)}$$



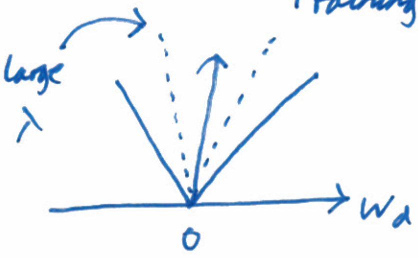
Why use other optimizers?

- Convergence?
- Not needing to tune...
- SGD can't give "sparse" solutions

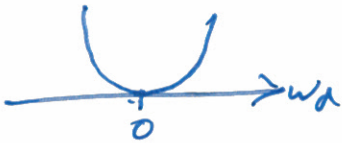
some $w_d = 0$

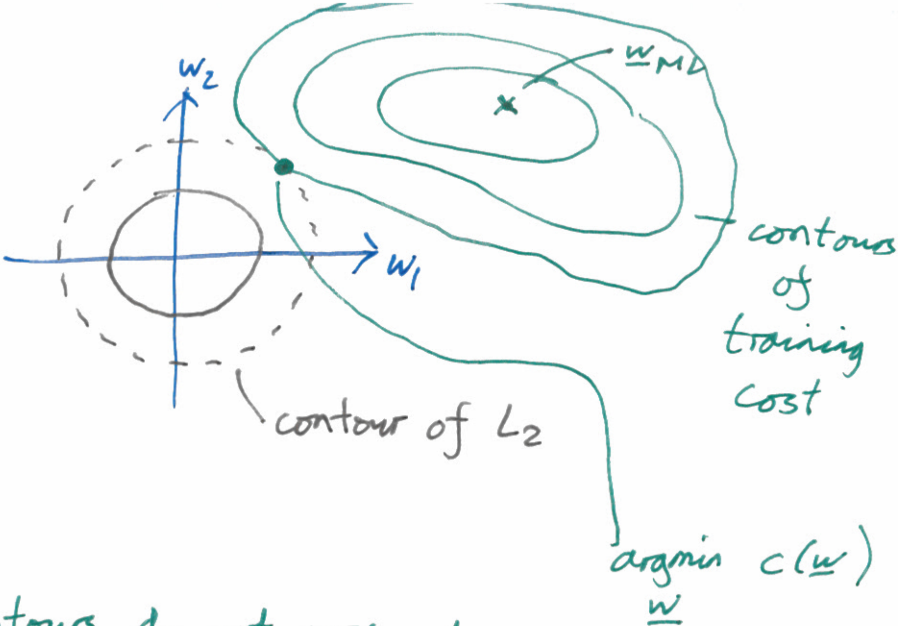
L1 Regularization

$$c(\underline{w}) = \underbrace{E(\underline{w})}_{\text{Training error}} + \underbrace{\lambda \sum_d |w_d|}_{\lambda \|\underline{w}\|_1}$$

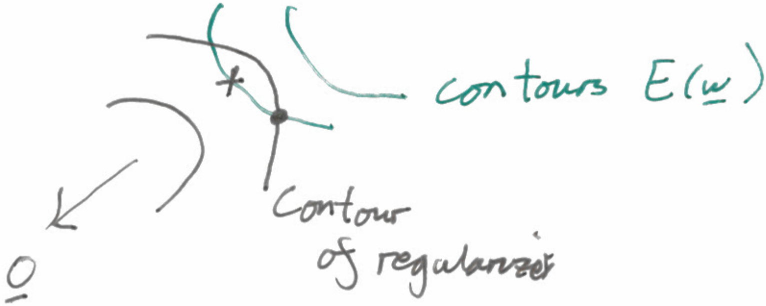


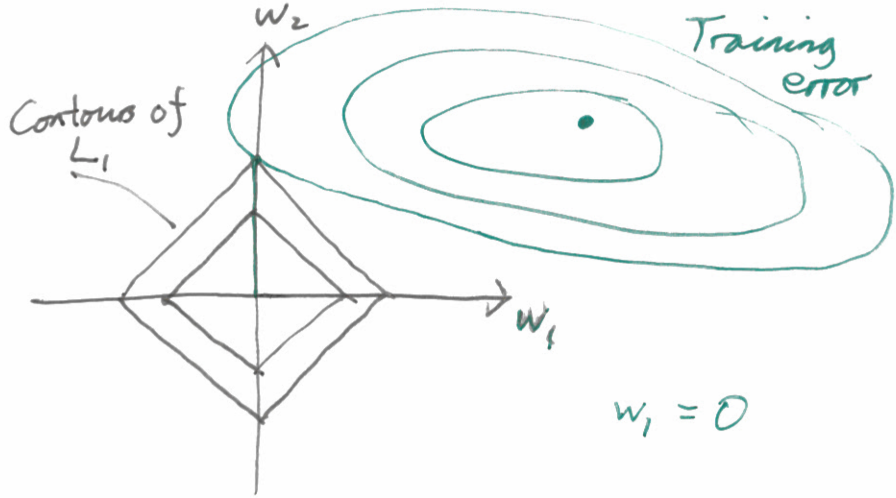
L_2 Regularizer





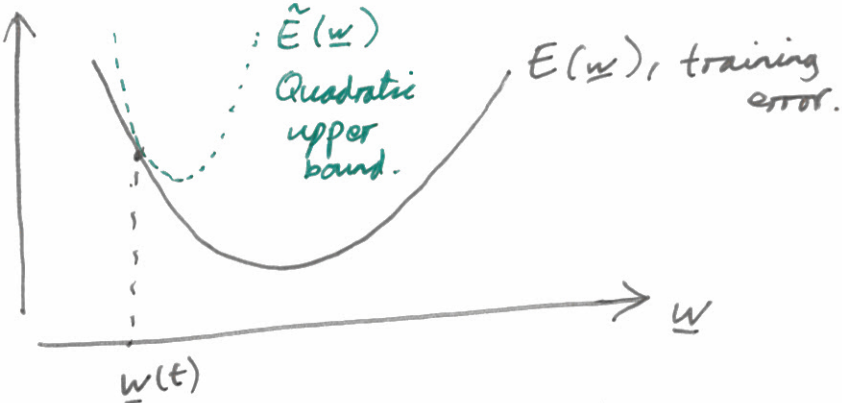
Contours do not cross at optimum:





There are many ways to fit L_1

One way are proximal methods:



$$\tilde{C}(\underline{w}) = \tilde{E}(\underline{w}) + \lambda \|\underline{w}\|_1$$

Can fit in closed form, some $w_d = 0$.

From Bayesian view, predictions are never sparse

$$p(y \mid \underline{x}, D) = \int p(y \mid \underline{x}, \underline{w}) \underbrace{p(\underline{w} \mid D)}_{p(w_d \neq 0 \mid D) > 0} d\underline{w}$$