## Announcements

- HW7 (project proposals) are due this Friday.
  - Collaborative Kaggle topic is posted on piazza, with comments
- Midterm is two weeks from today

## Today's Lecture

- Bayesian and MAP feature selection
- Sparsity and regularization

# BAYESIAN AND MAP FEATURE SELECTION

WHY FEATURE SELECTION?

- TO BETTER BALANCE $N$ WITH # COMPLEXITY ($d_{VC}$, OR ~ #d.o.f.)
- DISCOVER WHICH FEATURES ARE MOST IMPORTANT FOR PREDICTION.

1. USE CONTINUOUS VARIABLES $\underline{w}$.

EX: APARTMENT RENTS

$x_1$ = LIVING ~~ROOM~~ AREA (sq.f.): $500 \leq x_1 \leq 3500$

$x_2$ = # OF ROOMS: $1 \leq x_2 \lesssim 12$

$$\hat{f}(\underline{x}) = \underline{w}^T \underline{x} = w_0 + w_1 x_1 + w_2 x_2$$

$|w_1| \Leftrightarrow |w_2|$

CAN USE $|w_i| \Leftrightarrow |w_j|$ ONLY IF DATA ($x_i$) ARE APPROPRIATELY NORMALIZED (e.g., STANDARDIZED)

POSE AS A MAP ESTIMATION PROBLEM.

$$p(\underline{w} \mid \mathcal{D}) = \frac{1}{K}\, p(\mathcal{D} \mid \underline{w})\, p(\underline{w})$$

$$= \frac{1}{K}\left[\prod_{i=1}^{N} p(y_i \mid \underline{x}_i, \underline{w})\right] p(\underline{w}) \quad (1)$$

$$= \frac{1}{K}\, p(\underline{y} \mid \underline{\underline{X}}, \underline{w})\, p(\underline{w})$$

IF LINEAR REGRESSION W/ USUAL ASSUMPTION:

$$p(y \mid \underline{x}, \underline{w}) = N(y \mid \underline{w}^T \underline{x}, \sigma^2)$$

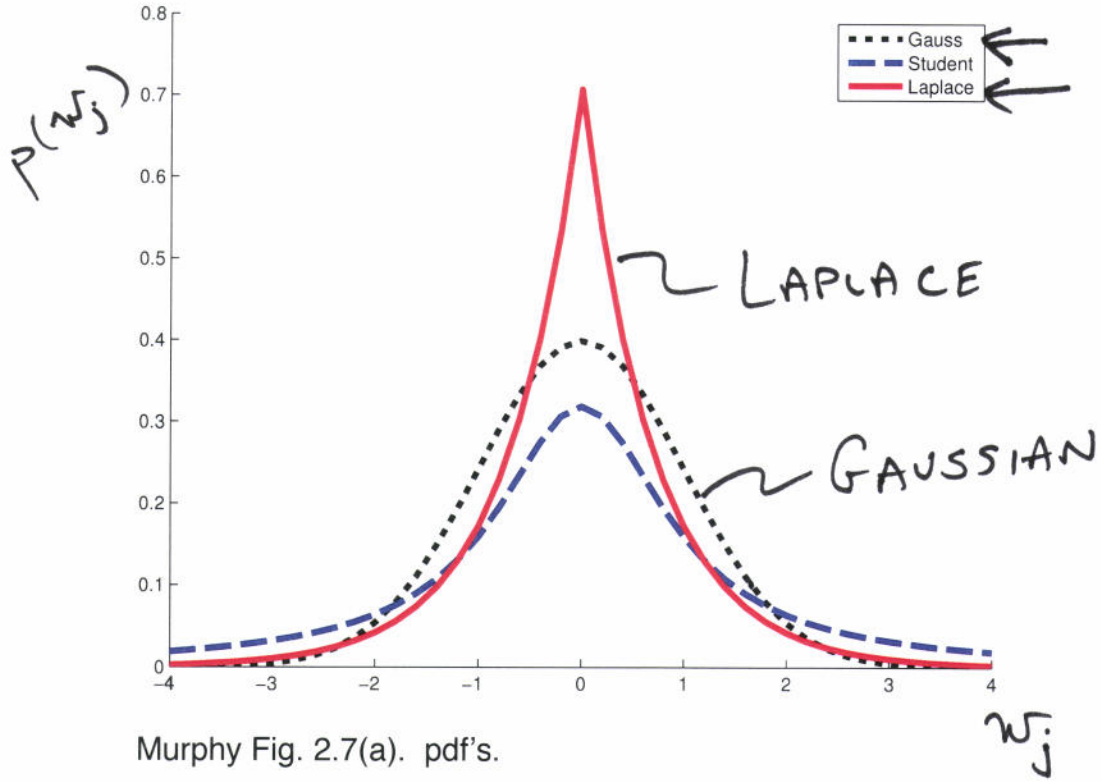
PRIOR TERM:

$$p(\underline{w}) = p(\underline{w} \mid \lambda)$$

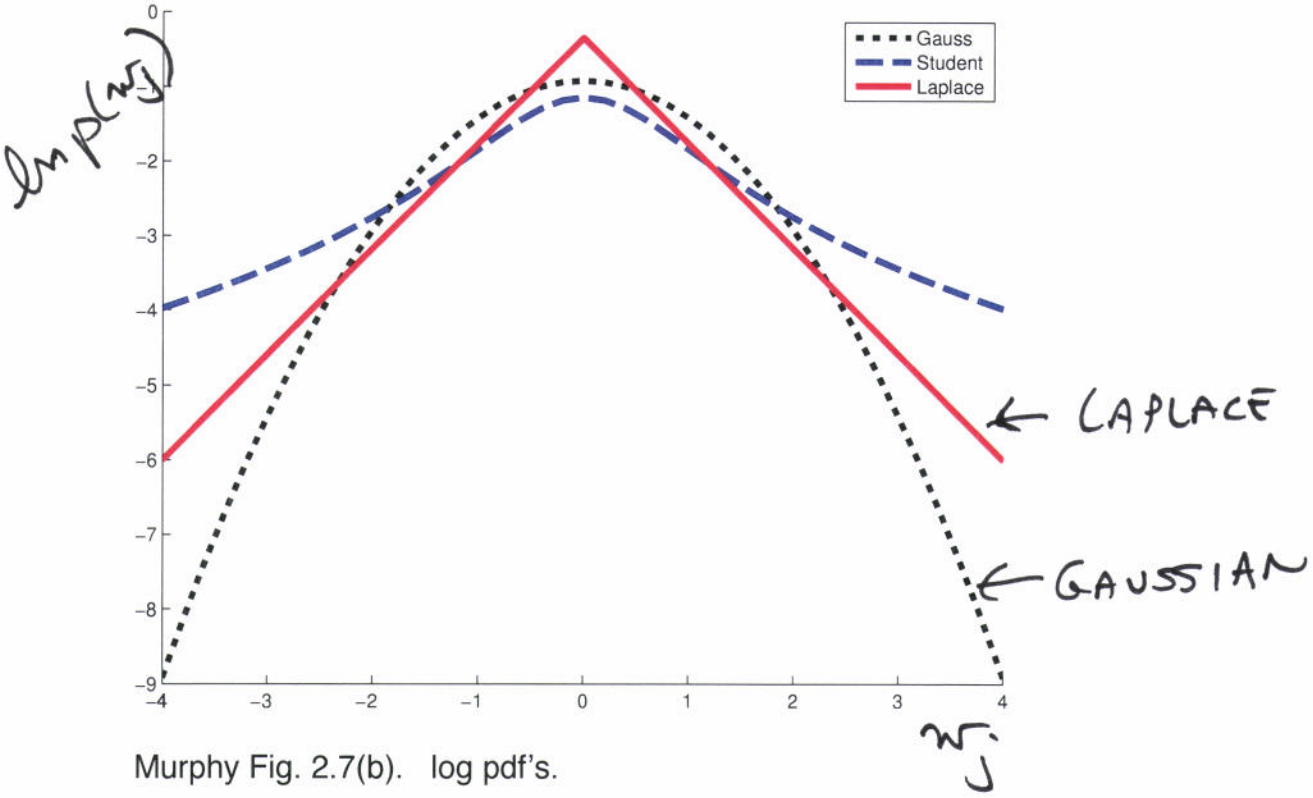
$$= \prod_{j=1}^{D} \text{Lap}\left(w_j \mid 0, \frac{1}{\lambda}\right)$$

$$= \prod_{j=1}^{D} \frac{\lambda}{2} e^{-\lambda |w_j|} \propto \prod_{j=1}^{D} e^{-\lambda |w_j|}$$

$$p(w_0) \propto 1 \quad \text{(UNIFORM)}.$$




Murphy Fig. 2.7(a). pdf's.




Murphy Fig. 2.7(b). log pdf's.

(1) $\Rightarrow$

$$J(\underline{w}) = -\ln p(\underline{y} \mid \underline{X}, \underline{w}) - \ln p(\underline{w} \mid \lambda)$$

$$= NLL(\underline{w}) + \lambda \|\underline{w}\|_1,$$

$$\|\underline{w}\|_1 \triangleq \sum_{j=1}^{D} |w_j|.$$

FOR LINEAR MODEL:

$$= \sum_{i=1}^{N} \frac{1}{2\sigma^2} \left[y_i - (w_0 + \underline{w}^T \underline{x}_i)\right]^2 + \lambda \|\underline{w}\|_1$$

$$J(\underline{w}) \propto RSS(\underline{w}) + \lambda' \|\underline{w}\|_1, \quad \lambda' = 2\lambda\sigma^2$$

BASIS PURSUIT DENOISING (BPDN) [M13.3.0]

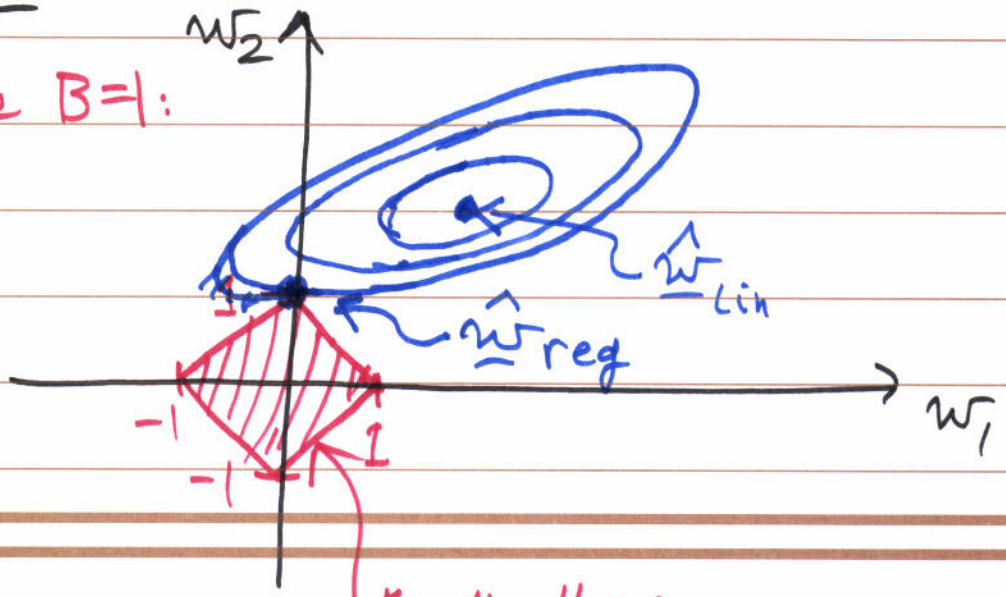
OR LASSO MINIMIZES THIS $J(\underline{w})$ USING SUBGRADIENTS (M 13.3.2)

# $\ell_1$ REGULARIZATION AND SPARSITY [M 13,3.1]

→ LOOK AT AS CONSTRAINED MIN. PROBLEMS:

$$\min_{\underline{w}} RSS(\underline{w}) \quad \text{S.T.} \quad \|\underline{w}\|_1 \leq B \quad \text{(LASSO)}$$

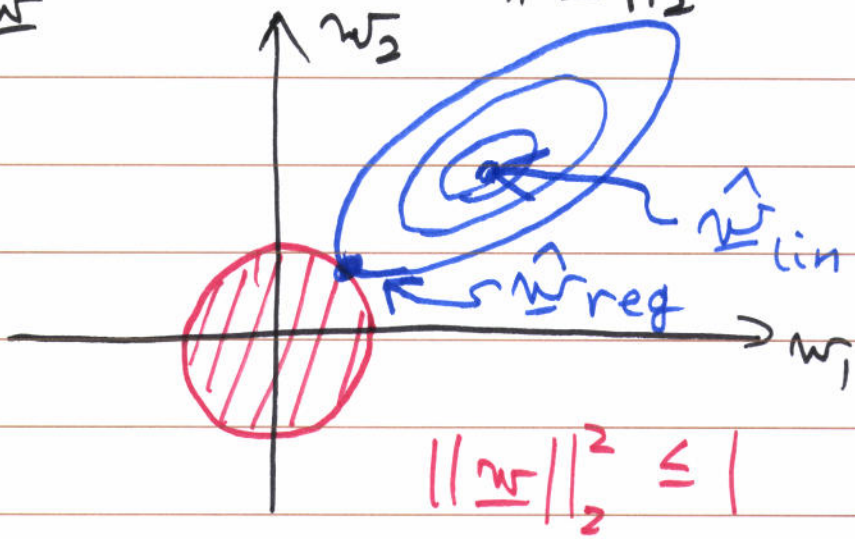
choose $B=1$:




$\|\underline{w}\|_1 \leq 1$

COMPARE WITH:

$$\min_{\underline{w}} RSS(\underline{w}) \quad \text{S.T.} \quad \|\underline{w}\|_2^2 \leq B \quad \text{(RIDGE)}$$




$\|\underline{w}\|_2^2 \leq 1$

EXAMPLE: PREDICTING PSA FROM VARIOUS PARAMETERS.

$\mathcal{D}$: SET OF DATA FROM PROSTATE CANCER PATIENTS (97).

– MURPHY TABLE 13.1 –

– HASTIE et al., TABLE 3.3 –

## 2. BAYESIAN FEATURE (VARIABLE) SELECTION [M 13.2]

$\underset{D\times 1}{\underline{\gamma}}$ : $\gamma_j = \begin{cases} 1 & \text{FEATURE } j \text{ IS RELEVANT} \\ 0 & \text{OTHERWISE} \end{cases}$

FIND $p(\underline{\gamma} \mid \mathcal{D})$.

$$p(\underline{\gamma} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \underline{\gamma})\, p(\underline{\gamma})}{p(\mathcal{D})}$$

PRIOR TERM $p(\underline{\gamma})$?

ASSUME: $p(\underline{\gamma}) = \prod_{j=1}^{D} p(\gamma_j)$

COMMON CHOICE:

$$p(\gamma_j) = \text{Ber}(\gamma_j \mid \pi_0) = \pi_0^{\gamma_j} (1-\pi_0)^{(1-\gamma_j)}$$

$\pi_0$ = PROBABILITY THAT A FEATURE IS RELEVANT.

$$\Rightarrow p(\underline{\gamma}) = \pi_0^{\sum \gamma_j} (1-\pi_0)^{D - \sum \gamma_j}$$

LET $\|\underline{\gamma}\|_0 \triangleq \sum_{j=1}^{D} \gamma_j$ = # OF NONZERO ELEMENTS IN $\underline{\gamma}$.

$= \ell_0$ PSEUDO-NORM.

$$p(\underline{\gamma}) = \pi_0^{\|\underline{\gamma}\|_0} (1-\pi_0)^{D - \|\underline{\gamma}\|_0}$$