Lecture 16 **EE 660** Oct 11, 2018

Announcements

- HW7 (project proposal) is due tomorrow.
- HW8 will be posted.

Today's Lecture

- Bayesian feature selection
  - Spike and slab model
  - Bernoulli-Gaussian model
  - L0 regularization
- p-norm and Bridge Regression

THEN:

$$\log p(\underline{\gamma} \mid \pi_0) = \|\underline{\gamma}\|_0 \log \pi_0 + (D - \|\underline{\gamma}\|_0) \cdot \log(1-\pi_0)$$

$$= \|\underline{\gamma}\|_0 \left[\log \pi_0 - \log(1-\pi_0)\right] + D \log(1-\pi_0)$$

$$= -\lambda \|\underline{\gamma}\|_0 + C', \quad \lambda \triangleq \log\left(\frac{1-\pi_0}{\pi_0}\right)$$

ASSUME: $\bar{x}_j = 0$, $\bar{y} = 0$

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

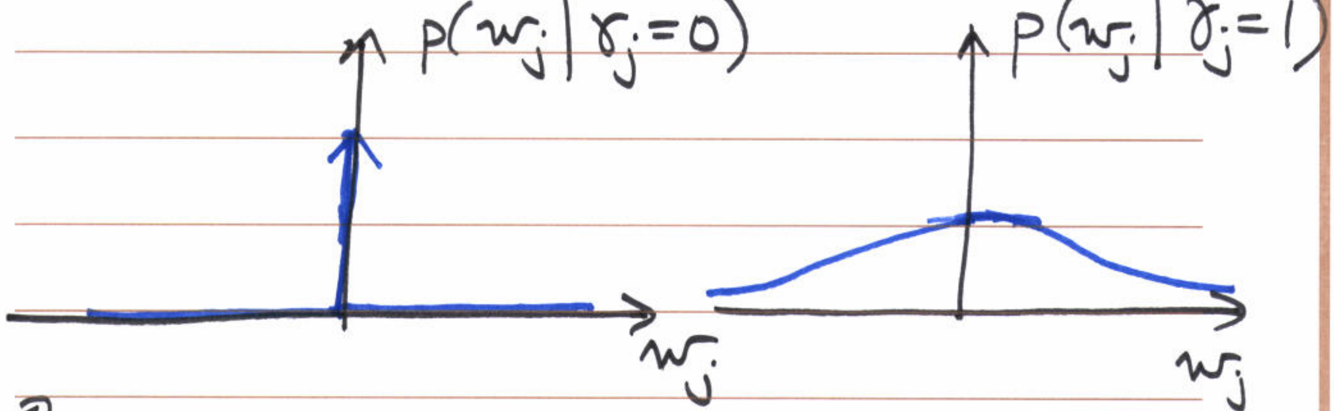
$$= \int p(\underline{y} \mid \underline{\underline{X}}, \underline{w}, \underline{\gamma}) \underbrace{p(\underline{w} \mid \underline{\gamma})}_{\text{prior on } \underline{w}.} d\underline{w}$$

$$p(\underline{w} \mid \underline{\gamma}) = \prod_{j=1}^{D} p(w_j \mid \gamma_j)$$

$$= \prod_{j=1}^{D} \begin{cases} \delta_0(w_j), & \text{IF } \gamma_j = 0 \\ N(w_j \mid 0, \sigma^2 \sigma_w^2), & \text{IF } \gamma_j = 1. \end{cases}$$

σ^2 = VARIANCE OF y.

σ_w^2 = VARIANCE OF RELEVANT FEATURE (w_j) RELATIVE TO σ^2.



"SPIKE AND SLAB" MODEL

— [MURPHY FIG. 13.1 (a), (c)] —

TYPICALLY σ_w^2 = LARGE ($\gg 1$), BUT DEPENDS ON REGULARIZATION DESIRED ON $\gamma_j = 1$ (RELEVANT) FEATURES.

SINCE THIS APPROACH EFFECTIVELY COMPUTES THE FULL $p(\underline{\gamma} \mid \mathcal{D})$, IT IS:

- THOROUGH
- COMPUTATIONALLY DEMANDING, OR
- " APPROXIMATE / SUB-OPTIMAL.

l_0 REGULARIZATION

SLIGHTLY DIFFERENT VIEW:

COMBINE $\underline{\gamma}$ AND $\underline{w}$ INTO 1 VECTOR.

MODEL:

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

WITH: $$\underline{w}_\gamma = \begin{bmatrix} \gamma_1 w_1 \\ \gamma_2 w_2 \\ \vdots \\ \gamma_D w_D \end{bmatrix}$$

($\underline{\gamma}$ IS A BINARY MASK FOR $\underline{w}$).

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

LIKELIHOOD $p(\mathcal{D} \mid \underline{w}, \underline{\gamma})$

$$= N(\underline{y} \mid \underline{\underline{X}}\, \underline{w}_\gamma, \sigma^2 \underline{\underline{I}})$$

$$= \prod_{i=1}^{N} N(y_i \mid \underline{w}_\gamma^T x_i, \sigma^2)$$

JOINT PRIOR $p(\underline{w}, \underline{\gamma}) = ?$

$$p(\underline{w}, \underline{\gamma}) = p(\underline{w})\, p(\underline{\gamma})$$

$$= \mathcal{N}(\underline{w} \mid \underline{0}, \sigma_w^2 \underline{\underline{I}})\, \pi_0^{\|\underline{\gamma}\|_0} (1-\pi_0)^{D - \|\underline{\gamma}\|_0}$$

BERNOULLI-GAUSSIAN MODEL

$$f_{obj}(\underline{w}, \underline{\gamma}) \propto -\log p(\underline{w}, \underline{\gamma} \mid \mathcal{D})$$

(DROP CONST.'S OF $\underline{w}, \underline{\gamma}$).

TAKE $\sigma_w^2 \to \infty$, SO DON'T REGULARIZE $\gamma_j = 1$ WEIGHTS

CAN SHOW:

$$f_{obj}(\underline{w}, \underline{\gamma}) = \|\underline{y} - \underline{\underline{X}}\, \underline{w}_\gamma\|_2^2 + \lambda \|\underline{\gamma}\|_0$$

$$f_{obj}(\underline{w}_\gamma) = \|\underline{y} - \underline{\underline{X}}\, \underline{w}_\gamma\|_2^2 + \lambda \|\underline{w}_\gamma\|_0$$

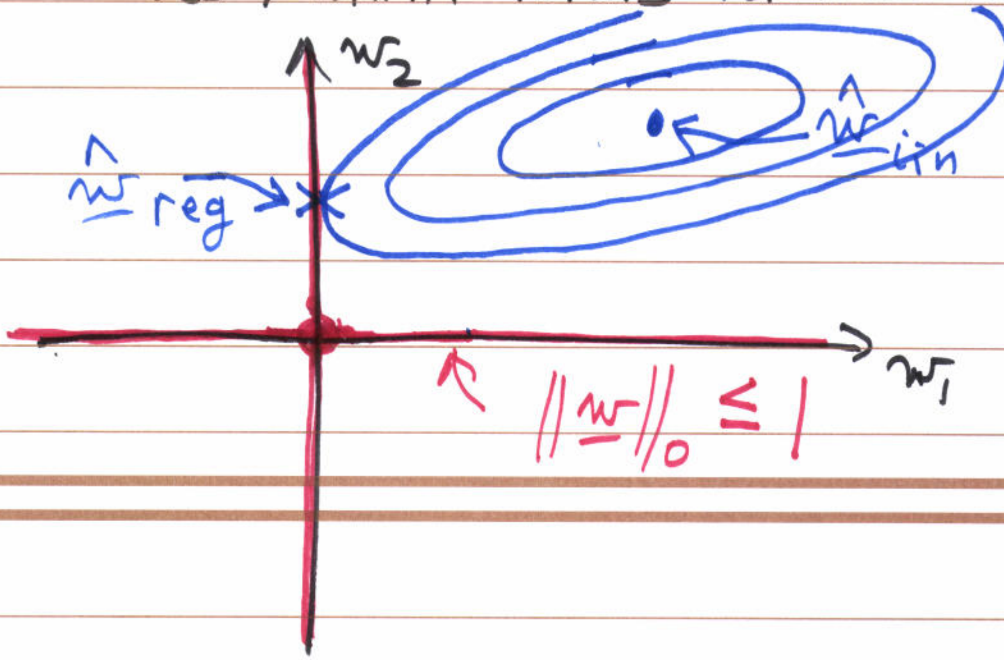
REDEFINE $\underline{w} \triangleq \underline{w}_\gamma$.

THEN:

$$\boxed{f_{obj}(\underline{w}) = \|\underline{y} - \underline{\underline{X}}\, \underline{w}\|_2^2 + \lambda \|\underline{w}\|_0}$$

COMMENTS:

1. MIN. OVER CONTINUOUS VARIABLES $\underline{w}$ NOW.
2. IS f_{obj} CONVEX?
   - → NO, $\|\underline{w}\|_0$ IS NOT CONVEX.
   - → SEE MURPHY 13.2.3 FOR MORE INFO [N.R.F.]



BRIDGE REGRESSION

GENERALIZE l_0, l_1, l_2 REGULARIZERS/CONSTRAINTS.

EST. $\underline{w}$ USING MAP WITH AN EXPONENTIAL POWER DISTRIBUTION AS PRIOR:

$$\mathrm{ExpPwr}(w_j \mid \mu_j, a, b) = \frac{b}{2a\,\Gamma\left(\frac{1}{b}\right)} \exp\left\{ - \frac{|x_j - \mu_j|^b}{a^b} \right\}$$

→ Murphy Eq. (13.132) may have errors.

LEADS TO AN ESTIMATE:

$$\hat{\underline{w}} = \underset{\underline{w}}{\operatorname{argmin}} \left\{ NLL(\underline{w}) + \lambda \underbrace{\sum_{j=1}^{D} |w_j|^b}_{} \right\},\ b \geq 0.$$

LIKE AN "l_b" REGULARIZER, OR "p-norm" BASED REGULARIZER.

Ex: $b=0$: ℓ_0 REG., (IF WE DEFINE $|0|^0=0$).
$b=1$: ℓ_1 REG.,
$b=2$: ℓ_2 REG.

$$\text{p-norm} = \left[\sum_{j=1}^{D} |w_j|^p\right]^{1/p}$$

(IS A NORM FOR $p \geq 1$).

$$\sum_{j=1}^{D} |w_j|^b$$

IS CONVEX FOR $b \geq 1$.

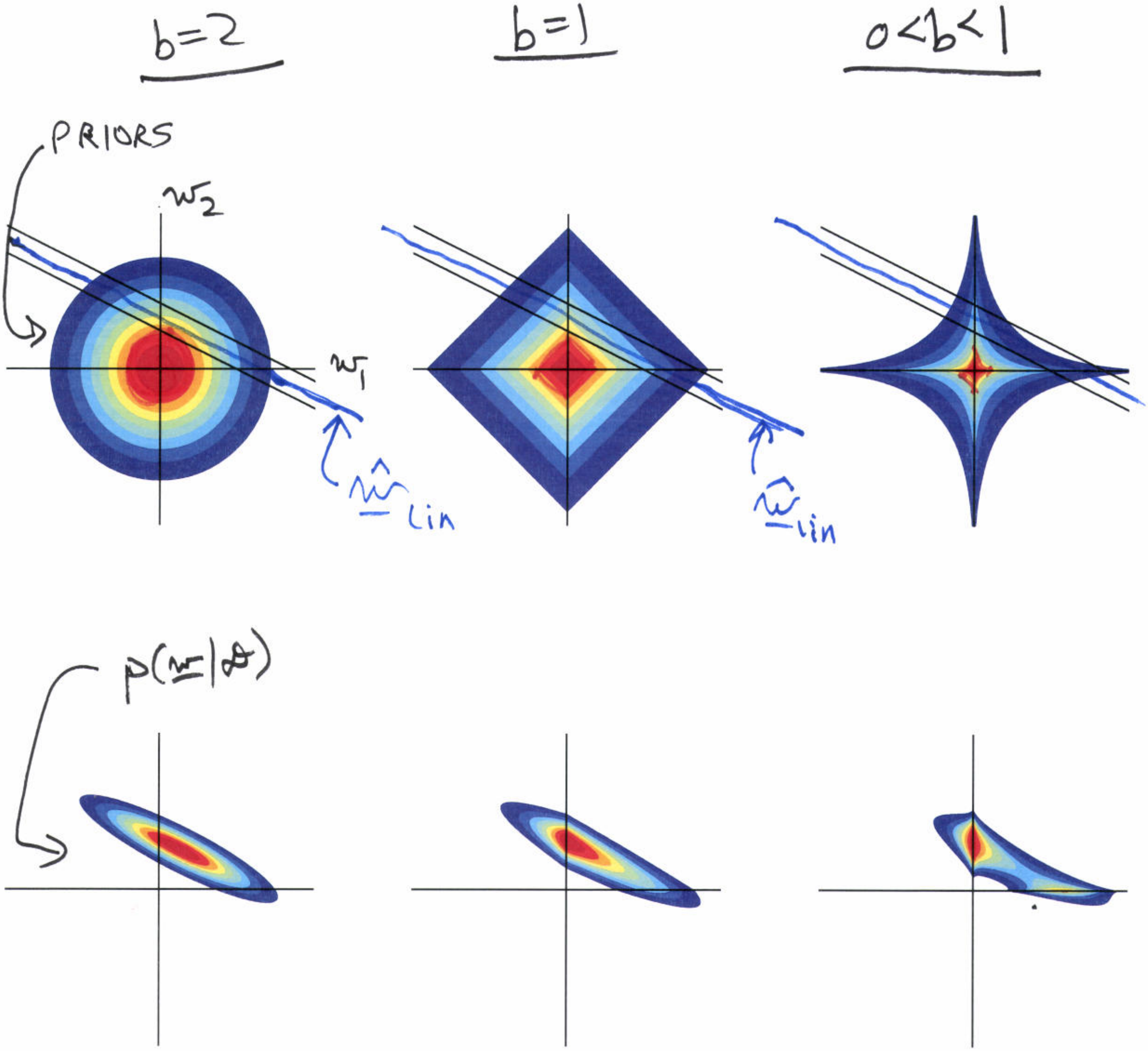
Ex: $f_{obj}(\underline{w}) = -\log p(\underline{w}|\mathcal{D})$

$$\propto \|\underline{y} - \underline{\underline{X}}\,\underline{w}\|_2^2 + \lambda \sum_{j=1}^{D} |w_j|^b$$

$\mathcal{D}$: ONE POINT $\underline{x}_1$, y_1

$$\Rightarrow f_{obj}(\underline{w}) \propto \left[y_1 - (w_1 x_{11} + w_2 x_{12})\right]^2 + \lambda\left[|w_1|^b + |w_2|^b\right].$$

given data point



Murphy Fig. 13.7