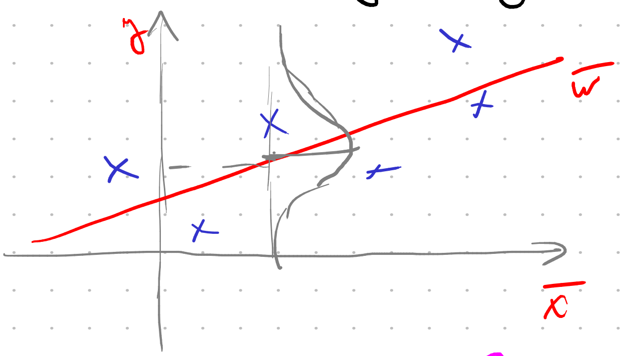
① Geometry of regularization



$y \sim \bar{w}^T \bar{x}$

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

$\log p(\bar{y} \mid \bar{w}, X) = Const - \underbrace{\frac{1}{2\sigma^2} \sum_n (y_n - \bar{w}^T \bar{x}_n)^2}$

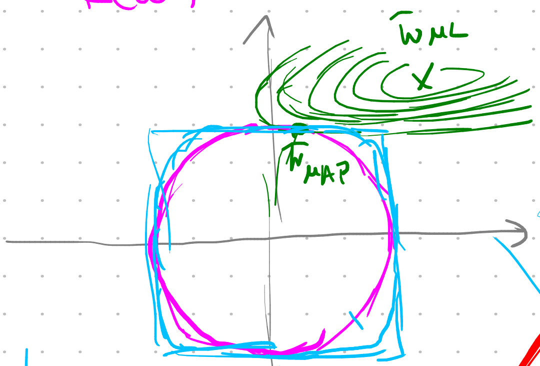
$L = \sum_n (y_n - \bar{w}^T \bar{x}_n)^2 + \alpha \cdot \|\bar{w}\|^2 = \bar{w}^T \bar{w}$ $\quad \xrightarrow{\bar{w}} \min$ ridge regression

$L = \sum_n^N (\_\_)^2 + \alpha \sum_{i=1}^d |w_i|$ $\quad \xrightarrow{\bar{w}} \min$ lasso regression

$\sum_n (\ldots)^2 + \underline{\alpha \cdot R(\bar{w})} \to \min \quad \Leftrightarrow \quad \sum_n (\_)^2 \to \min$

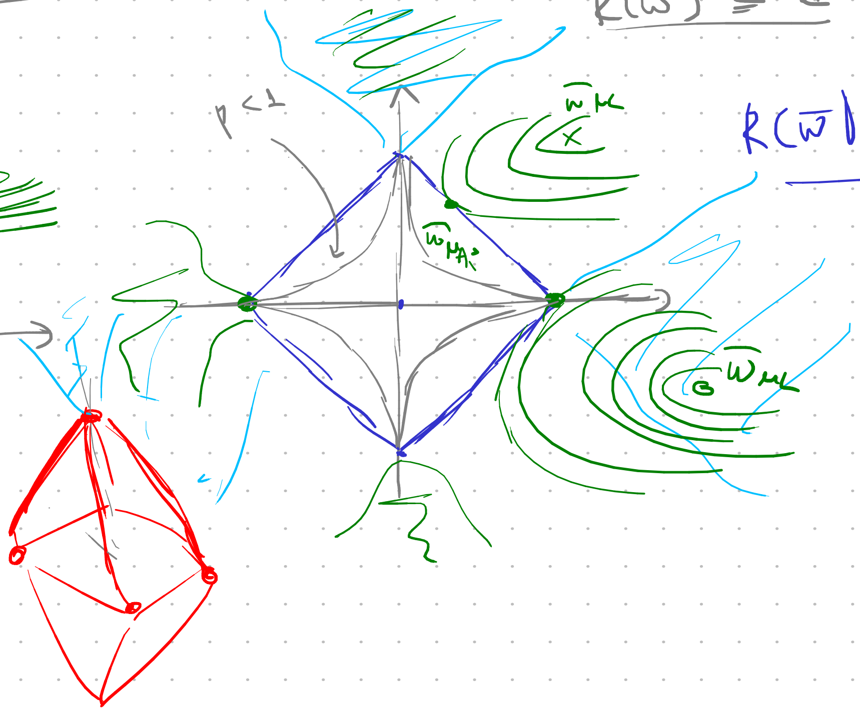
$\underline{R(\bar{w}) \leq c}$

$R(\bar{w}) = \|\bar{w}\|^2$



$L_\infty = \max |w_i|$

$R(\bar{w}) = \sum_n |w_i|^p$



$R(\bar{w}) = \underline{\sum_i |w_i|}$

② Bayesian inference in LR

$$p(\bar{w} \mid D) \propto p(\bar{w}) p(D \mid \bar{w})$$

$$p(D \mid \bar{w}) = \prod_n \mathcal{N}(y_n \mid \bar{w}^T \bar{x}_n, \sigma^2)$$

$$\boxed{p(\bar{w}) = \mathcal{N}(\bar{w} \mid \bar{0}, \sigma_0^2 \cdot I)}$$

$$\log p(\bar{w} \mid D) = Const - \frac{1}{2\sigma^2} \sum_n (y_n - \bar{w}^T \bar{x}_n)^2 - \underline{\frac{1}{2\sigma_0^2} \bar{w}^T I \bar{w}}$$

$$L_1\text{-reg.} \Leftrightarrow p(\bar{w}) = Const \cdot e^{-Const \cdot \sum |w_i|}$$

Conjugate prior: $p(\bar{w}|\bar{\alpha})$

$$p(\bar{w}|\bar{\alpha}) \times p(D|\bar{w}) \propto p(\bar{w}|D,\bar{\alpha}) = p(\bar{w}|\bar{\alpha}')$$

$$\prod_n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_n - \bar{w}^T\bar{x}_n)^2}$$

$$p(\bar{w}|\underbrace{\bar{\mu}_0, \Sigma_0}_{\bar{\alpha}}) = \mathcal{N}(\bar{w}|\bar{\mu}_0, \Sigma_0) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma_0}} e^{-\frac{1}{2}(\bar{w}-\bar{\mu}_0)^T \Sigma_0^{-1}(\bar{w}-\bar{\mu}_0)}$$

$$\log p(\bar{w}|\bar{y}, X, \bar{\mu}_0, \Sigma_0) = \text{Const} - \frac{1}{2\sigma^2} \overbrace{\sum_n (y_n - \bar{w}^T\bar{x}_n)^2}^{=(\bar{y}-X\bar{w})^T(\bar{y}-X\bar{w})} - $$

$$-\frac{1}{2}(\bar{w}-\bar{\mu}_0)^T \Sigma_0^{-1}(\bar{w}-\bar{\mu}_0) =$$

$$= \text{Const} - \frac{1}{2\sigma^2}\left(\cancel{\bar{y}^T\bar{y}} - 2\bar{w}^T(X^T\bar{y}) + \bar{w}^T X^T X \bar{w}\right) -$$

$$-\frac{1}{2}\left(\bar{w}^T \Sigma_0^{-1}\bar{w} - \cancel{\bar{\mu}_0^T \Sigma_0^{-1}\bar{w}} - 2\bar{w}^T(\Sigma_0^{-1}\bar{\mu}_0) + \cancel{\bar{\mu}_0^T \Sigma_0^{-1}\bar{\mu}_0}\right)$$

$$= \text{Const} - \frac{1}{2}\bar{w}^T\left(\Sigma_0^{-1} + \frac{1}{\sigma^2}X^T X\right)\bar{w} + \bar{w}^T \overbrace{\left(\Sigma_0^{-1}\bar{\mu}_0 + \frac{1}{\sigma^2}X^T\bar{y}\right)}^{\Sigma_N^{-1}\bar{\mu}_N}$$

$$p(\bar{w}|\bar{y}, X, \bar{\mu}_0, \Sigma_0) = \mathcal{N}(\bar{w}|\bar{\mu}_N, \Sigma_N), \text{ где}$$

$$\begin{cases} \Sigma_N^{-1} = \Sigma_0^{-1} + \frac{1}{\sigma^2}X^T X \\ \bar{\mu}_N = \Sigma_N\left(\Sigma_0^{-1}\bar{\mu}_0 + \frac{1}{\sigma^2}X^T\bar{y}\right) \end{cases}$$

$$p(\bar{w}|\bar{\mu}_0,\Sigma_0) \times p(D|\bar{w}) \propto p(\bar{w}|\bar{\mu},\Sigma) \times p(D'|\bar{w}) \propto p(\bar{w}|\bar{\mu}',\Sigma')$$

$$p(y|\bar{x}, D) = \int p(y,\bar{w}|\bar{x},D)\,d\bar{w} = \int p(y|\bar{w},\bar{x}) \cdot p(\bar{w}|D)\,d\bar{w}$$

$$= \mathbb{E}_{p(\bar{w}|D)}\left[p(y|\bar{w},\bar{x})\right] \approx \frac{1}{R}\sum \ldots \bar{w}_r \sim p(\bar{w}|D)$$

③ Predictive distribution in LR

$$p(y|\bar{x},D) = \int p(y|\bar{\omega},\bar{x})\,p(\bar{\omega}|D)\,d\bar{\omega} = \int \mathcal{N}(y|\bar{\omega}^T\bar{x},\sigma^2)\,\mathcal{N}(\bar{\omega}|\bar{\mu}_N,\Sigma_N)\,d\bar{\omega}$$

$$= f(y,\bar{x},D)\ \underbrace{\int \mathcal{N}(\bar{\omega}|\bar{\mu}',\Sigma')\,d\bar{\omega}}_{=1}$$

$$\log p(y|\bar{\omega},\bar{x}) + \log p(\bar{\omega}|D) =$$

$$= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y-\bar{\omega}^T\bar{x})^2 - \frac{d}{2}\log 2\pi - \frac{1}{2}\log\det\Sigma_N - $$

$$- \frac{1}{2}(\bar{\omega}-\bar{\mu}_N)^T\Sigma_N^{-1}(\bar{\omega}-\bar{\mu}_N)$$

$$= \text{Const} - \frac{1}{2\sigma^2}\left(y^2 - 2y\bar{\omega}^T\bar{x} + \bar{\omega}^T\bar{x}\bar{x}^T\bar{\omega}\right) -$$

$$-\frac{1}{2}\left(\bar{\omega}^T\Sigma_N^{-1}\bar{\omega} - 2\bar{\omega}^T(\Sigma_N^{-1}\bar{\mu}_N) + \bar{\mu}_N^T\Sigma_N^{-1}\bar{\mu}_N\right) =$$

$$= \text{Const} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\left(\bar{\omega}^T\left(\Sigma_N^{-1} + \frac{1}{\sigma^2}\bar{x}\bar{x}^T\right)\bar{\omega} - 2\bar{\omega}^T\left(\Sigma_N^{-1}\bar{\mu}_N + \frac{y}{\sigma^2}\bar{x}\right)\right)$$

$$= \text{Const} - \frac{y^2}{2\sigma^2} - \frac{1}{2}(\bar{\omega}-\bar{\mu}')^T\Sigma'^{-1}(\bar{\omega}-\bar{\mu}') + \frac{1}{2}\bar{\mu}'^T\Sigma'^{-1}\bar{\mu}' =$$

$$\Sigma'^{-1} = \Sigma_N^{-1} + \frac{1}{\sigma^2}\bar{x}\bar{x}^T$$

$$\bar{\mu}' = \Sigma'\left(\Sigma_N^{-1}\bar{\mu}_N + \frac{y}{\sigma^2}\bar{x}\right)$$

$$\frac{1}{2}(\ )^T\cancel{\Sigma'}\underbrace{\cancel{\left(\Sigma_N^{-1} + \frac{1}{\sigma^2}\bar{x}\bar{x}^T\right)}}_{=\Sigma'^{-1}}\Sigma'(\_\_)$$

$$= \frac{1}{2}(\_\_)^T\Sigma'(\_\_)$$

$$= \text{Const} - \frac{y^2}{2\sigma^2} + \frac{1}{2}\left(\Sigma_N^{-1}\bar{\mu}_N + \frac{y}{\sigma^2}\bar{x}\right)^T\Sigma'\left(\Sigma_N^{-1}\bar{\mu}_N + \frac{y}{\sigma^2}\bar{x}\right) - \frac{1}{2}(\_\_\_\_)$$

$$p(y|\bar{x}, D) = \mathcal{N}(y|\mu_{pred}, \sigma^2_{pred})$$

$$(\;)^T\left(\Sigma' \Sigma_N^{-1}\bar{\mu}_N + \underbrace{\frac{y}{\sigma^2}\Sigma'\bar{x}}\right) = \;=\frac{\mu_{pred}}{\sigma^2_{pred}}\;\|$$

$$= \mu_N^T \Sigma_N^{-1}\Sigma'\Sigma_N^{-1}\bar{\mu}_N + \frac{y}{\sigma^2}\underbrace{\bar{x}^T\Sigma'\Sigma_N^{-1}\bar{\mu}_N} + \frac{y}{\sigma^2}\underbrace{\bar{\mu}_N^T\Sigma_N^{-1}\Sigma'\bar{x}} + \frac{y^2}{\sigma^4}\underbrace{\bar{x}^T\Sigma'\bar{x}}$$

$$\frac{1}{\sigma^2_{pred}} = \frac{1}{\sigma^2}\left(1 - \frac{\bar{x}^T\underbrace{\Sigma'\bar{x}}}{\sigma^2}\right)$$

$$\Sigma'^{-1}\underbrace{\Sigma_N\bar{x}} = \bar{x} + \frac{1}{\sigma^2}\bar{x}\left(\bar{x}^T\Sigma_N\bar{x}\right) = \left(1+\frac{1}{\sigma^2}\bar{x}^T\Sigma_N\bar{x}\right)\bar{x}$$

$$\Sigma_N\bar{x} = \left(1+\frac{1}{\sigma^2}\bar{x}\Sigma_N\bar{x}\right)\Sigma'\bar{x}$$

$$\Sigma'\bar{x} = \frac{\Sigma_N\bar{x}}{1+\frac{1}{\sigma^2}\bar{x}^T\Sigma_N\bar{x}}$$

$$\frac{1}{\sigma^2_{pred}} = \frac{1}{\sigma^2}\left(1 - \frac{\frac{1}{\sigma^2}\bar{x}^T\Sigma_N\bar{x}}{1+\frac{1}{\sigma^2}\bar{x}^T\Sigma_N\bar{x}}\right) = \frac{1}{\sigma^2+\bar{x}^T\Sigma_N\bar{x}}$$

noise

uncertainty in $\bar{w}$

$$\boxed{\sigma^2_{pred} = \sigma^2 + \bar{x}^T\Sigma_N\bar{x}}$$

$$\boxed{\mu_{pred} = \bar{\mu}_N^T\bar{x}}$$

$$\frac{\mu_{pred}}{\sigma^2+\bar{x}^T\Sigma_N\bar{x}} = \frac{1}{\sigma^2}\bar{\mu}_N^T\cdot\cancel{\Sigma_N^{-1}}\cdot\frac{\cancel{\Sigma_N}\bar{x}}{1+\frac{1}{\sigma^2}\bar{x}^T\Sigma_N\bar{x}}$$