

Optimization Problems 1

Exercise 1: Regression

- (a) Let
- $f : \mathbb{R}^d \rightarrow \mathbb{R}, \theta \mapsto 0.5 \|\mathbf{X}\theta - \mathbf{y}\|_2^2 + 0.5 \cdot \lambda \|\theta\|_2^2, \lambda > 0$

$$\frac{\partial}{\partial \theta} f = \theta^\top \mathbf{X}^\top \mathbf{X} - \mathbf{y}^\top \mathbf{X} + \lambda \theta^\top \stackrel{!}{=} \mathbf{0} \iff \theta^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}) = \mathbf{y}^\top \mathbf{X}$$

$$\Rightarrow \theta = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}.$$

$$\frac{\partial^2}{\partial \theta \partial \theta^\top} f = \underbrace{\mathbf{X}^\top \mathbf{X}}_{\text{p.s.d.}} + \underbrace{\lambda \mathbf{I}}_{\text{p.d. if } \lambda > 0} \text{ is p.d. if } \lambda > 0 \Rightarrow f \text{ is (strictly) convex}$$

- (b) Since the observations and parameters are assumed to be i.i.d. it follows that

$$p_{\theta | \mathbf{x}, \mathbf{y}}(\theta) \propto p_{\mathbf{y} | \mathbf{x}, \theta}(\mathbf{y}) p_{\theta}(\theta) \propto \exp\left(-\frac{(\mathbf{X}\theta - \mathbf{y})^\top \mathbf{I}(\mathbf{X}\theta - \mathbf{y})}{2}\right) \exp\left(-\frac{\theta^\top \mathbf{I} \theta}{2\sigma_w^2}\right).$$

The minimizer of the negative log posterior density is maximizer of posterior density and hence

$$\theta^* = \arg \min_{\theta} -\log\left(\exp\left(-\frac{(\mathbf{X}\theta - \mathbf{y})^\top \mathbf{I}(\mathbf{X}\theta - \mathbf{y})}{2}\right) \exp\left(-\frac{\theta^\top \mathbf{I} \theta}{2\sigma_w^2}\right)\right) = \arg \min_{\theta} 0.5 \|\mathbf{X}\theta - \mathbf{y}\|_2^2 + 0.5 \cdot \frac{1}{2\sigma_w^2} \|\theta\|_2^2.$$

This is ridge regression and the solution follows from a) with $\lambda = \frac{1}{\sigma_w^2}$.

- (c) From b) we see that for the density of interest it must hold that

$$-\log p(\theta) = 0.5 \cdot \lambda \|\theta\|_2^2 + c \text{ with } c \in \mathbb{R} \iff p(\theta) \propto \exp(-0.5 \cdot \lambda \|\theta\|_2^2).$$

$$\Rightarrow \theta \stackrel{\text{i.i.d.}}{\sim} \text{Laplace}(0, 2/\lambda).$$

- (d) Let
- $f : \mathbb{R}^d \rightarrow \mathbb{R}, \theta \mapsto \|\mathbf{X}\theta - \mathbf{y}\|_2^2$
- .

First consider the difference vector between the unregularized solution and the regularized one:

$$\theta_{\text{reg}}^* - \theta^* = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} - \theta^* = ((\mathbf{X}^\top \mathbf{X})^{-1} - (\mathbf{X}^\top \mathbf{X})^{-1} \lambda \mathbf{I} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1}) \mathbf{X}^\top \mathbf{y} - \theta^*$$

$$= \theta^* - (\mathbf{X}^\top \mathbf{X})^{-1} \lambda \mathbf{I} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} - \theta^* = -(\mathbf{X}^\top \mathbf{X})^{-1} \lambda \mathbf{I} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}.$$

This difference is only zero in general if $\lambda = 0 \Rightarrow \theta_{\text{reg}}^* \neq \theta^*$.

Now, assume that $\|\theta^*\|_2 \leq \|\theta_{\text{reg}}^*\|_2$ then it follows that $\theta_{\text{reg}}^* = \arg \min_{\theta} f$ s.t. $\|\theta^*\|_2 \leq \|\theta\|_2 \leq t$ and consequently $\theta_{\text{reg}}^* = \theta^*$ which is a contradiction $\Rightarrow \|\theta_{\text{reg}}^*\|_2 < \|\theta^*\|_2$.

Now, assume that $\|\theta_{\text{reg}}^*\|_2 < t(\lambda) < \|\theta^*\|_2$:

Since, by assumption $\mathbf{X}^\top \mathbf{X}$ is non-singular, f is strictly convex and $f(\theta_{\text{reg}}^*) > f(\theta^*)$.

Consider $\tilde{\theta} = \theta_{\text{reg}}^* + \frac{\theta^* - \theta_{\text{reg}}^*}{\|\theta^* - \theta_{\text{reg}}^*\|_2} \cdot \frac{t(\lambda) - \|\theta_{\text{reg}}^*\|_2}{2}$ then $\tilde{\theta}$ is by construction on the line between θ_{reg}^* and θ^* .

Hence $f(\tilde{\theta}) < f(\theta_{\text{reg}}^*)$ which is a contradiction (θ_{reg}^* should be minimal in the constrained region) since $\|\tilde{\theta}\|_2 < t$ by construction.

$$\Rightarrow \|\theta_{\text{reg}}^*\|_2 = t(\lambda).$$

Exercise 2: Classification

- (a) First observe that
- $1 - \mathbb{P}(y = 1 | \mathbf{x}^{(i)}) = \frac{\exp(-\theta^\top \mathbf{x}^{(i)})}{1 + \exp(-\theta^\top \mathbf{x}^{(i)})} = \frac{1}{1 + \exp(\theta^\top \mathbf{x}^{(i)})} = \mathbb{P}(y = 1 | -\mathbf{x}^{(i)}).$

Define $\sigma(\mathbf{x}) := \mathbb{P}(y = 1 | \mathbf{x}^{(i)})$.

With this we get that $\log(\mathbb{P}(y = y^{(i)} | \mathbf{x}^{(i)})) = \log\left(\mathbb{P}(y = 1 | \mathbf{x}^{(i)})^{y^{(i)}} (1 - \mathbb{P}(y = 1 | \mathbf{x}^{(i)}))^{1-y^{(i)}}\right)$

$$= y^{(i)} \log(\sigma(\mathbf{x}^{(i)})) + (1 - y^{(i)}) \log(1 - \sigma(\mathbf{x}^{(i)}))$$

$$= y^{(i)} (\log(\sigma(\mathbf{x}^{(i)})) - \log(\sigma(-\mathbf{x}^{(i)}))) + \log(\sigma(-\mathbf{x}^{(i)}))$$

$$= y^{(i)} \left(\log\left(\frac{\sigma(\mathbf{x}^{(i)})}{\sigma(-\mathbf{x}^{(i)})}\right) \right) + \log(\sigma(-\mathbf{x}^{(i)}))$$

$$= y^{(i)} \left(\log\left(\frac{1 + \exp(\theta^\top \mathbf{x}^{(i)})}{1 + \exp(-\theta^\top \mathbf{x}^{(i)})}\right) \right) - \log(1 + \exp(\theta^\top \mathbf{x}^{(i)}))$$

$$= y^{(i)} \left(\log\left(\exp(\theta^\top \mathbf{x}^{(i)}) \frac{1 + \exp(-\theta^\top \mathbf{x}^{(i)})}{1 + \exp(-\theta^\top \mathbf{x}^{(i)})}\right) \right) - \log(1 + \exp(\theta^\top \mathbf{x}^{(i)}))$$

$$= y^{(i)} \theta^\top \mathbf{x}^{(i)} - \log(1 + \exp(\theta^\top \mathbf{x}^{(i)}))$$

With this we find that $\mathcal{R}_{\text{emp}} = -\log \prod_{i=1}^n \mathbb{P}(y = y^{(i)} | \mathbf{x}^{(i)}) = \sum_{i=1}^n \log(1 + \exp(\theta^\top \mathbf{x}^{(i)})) - y^{(i)} \theta^\top \mathbf{x}^{(i)}$

$$(b) \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{R}_{\text{emp}} = \sum_{i=1}^n \frac{\exp(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})}{1 + \exp(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})} \mathbf{x}^{(i)\top} - y^{(i)} \mathbf{x}^{(i)\top}$$

$$\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \mathcal{R}_{\text{emp}} = \sum_{i=1}^n \frac{\exp(\boldsymbol{\theta}^\top \mathbf{x}^{(i)}) (1 + \exp(\boldsymbol{\theta}^\top \mathbf{x}^{(i)}) - \exp(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2)}{(1 + \exp(\boldsymbol{\theta}^\top \mathbf{x}^{(i)}))^2} \mathbf{x}^{(i)} \mathbf{x}^{(i)\top} = \sum_{i=1}^n \underbrace{\frac{\exp(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})}{(1 + \exp(\boldsymbol{\theta}^\top \mathbf{x}^{(i)}))^2}}_{>0} \underbrace{\mathbf{x}^{(i)} \mathbf{x}^{(i)\top}}_{\text{p.s.d.}}$$

is p.s.d. $\Rightarrow \mathcal{R}_{\text{emp}}$ is convex.

(c) We can transform the inequalities such that

$$\zeta^{(i)} \geq 1 - y^{(i)} (\boldsymbol{\theta}^\top \mathbf{x}^{(i)} + \boldsymbol{\theta}_0) \quad \forall i \in \{1, \dots, n\} \text{ and } \zeta^{(i)} \geq 0 \quad \forall i \in \{1, \dots, n\}.$$

We can find the smallest $\zeta^{(i)}$ by assuring that always at least one constraint is active¹ since this means that the value can not be further reduced:

$$\zeta^{(i)} = \begin{cases} 1 - y^{(i)} (\boldsymbol{\theta}^\top \mathbf{x}^{(i)} + \boldsymbol{\theta}_0) & \text{for } 1 - y^{(i)} (\boldsymbol{\theta}^\top \mathbf{x}^{(i)} + \boldsymbol{\theta}_0) \geq 0 \\ 0 & \text{for } 1 - y^{(i)} (\boldsymbol{\theta}^\top \mathbf{x}^{(i)} + \boldsymbol{\theta}_0) < 0 \end{cases} = \max(1 - y^{(i)} (\boldsymbol{\theta}^\top \mathbf{x}^{(i)} + \boldsymbol{\theta}_0), 0)$$

Inserting these $\zeta^{(i)}$ into the objective function results in $f(\boldsymbol{\theta}) = \frac{1}{2} \|\boldsymbol{\theta}\|_2^2 + C \sum_{i=1}^n \max(1 - y^{(i)} \boldsymbol{\theta}^\top \mathbf{x}^{(i)} + \boldsymbol{\theta}_0, 0)$.

Minimizing f is equivalent to minimizing $\frac{1}{2C} \|\boldsymbol{\theta}\|_2^2 + \sum_{i=1}^n \max(1 - y^{(i)} (\boldsymbol{\theta}^\top \mathbf{x}^{(i)} + \boldsymbol{\theta}_0), 0) \Rightarrow \lambda = \frac{1}{2C}$.

(d) First we show that $g : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto \max(x, 0)$ is convex:

$g(x) = 0.5|x| + 0.5x \Rightarrow \max(x, 0)$ is convex since it is the sum of two convex functions.

Also g is increasing $\Rightarrow \max(1 - y^{(i)} \boldsymbol{\theta}^\top \mathbf{x}^{(i)} + \boldsymbol{\theta}_0, 0)$ is convex since $1 - y^{(i)} \boldsymbol{\theta}^\top \mathbf{x}^{(i)} + \boldsymbol{\theta}_0$ is convex (linear).

With this we can conclude that $\sum_{i=1}^n \max(1 - y^{(i)} (\boldsymbol{\theta}^\top \mathbf{x}^{(i)} + \boldsymbol{\theta}_0), 0) + \lambda \|\boldsymbol{\theta}\|_2^2$ is convex since it is the sum of convex functions.

¹the \geq constraint is fulfilled with equality