

Optimization Problems 1

Solution 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

$$p_{\theta | \mathbf{x}, \mathbf{y}}(\theta) \propto p_{\mathbf{y} | \mathbf{x}, \theta}(\theta) p_{\theta}(\theta) \propto \exp\left(-\frac{(\mathbf{X}\theta - \mathbf{y})^\top \mathbf{I}^{-1} (\mathbf{X}\theta - \mathbf{y})}{2}\right) \exp\left(-\frac{\theta^\top \theta}{2\sigma_w^2}\right).$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

- (c) We can transform the inequalities such that

$$\zeta^{(i)} \geq 1 - y^{(i)} (\theta^\top \mathbf{x}^{(i)} + \theta_0) \quad \text{and} \quad \zeta^{(i)} \geq 0$$

for all $i \in \{1, \dots, n\}$. However, for a minimizer of the first primal form, it has to hold that

$$\zeta^{(i)} = \begin{cases} 1 - y^{(i)} (\theta^\top \mathbf{x}^{(i)} + \theta_0) & \text{if } 1 - y^{(i)} (\theta^\top \mathbf{x}^{(i)} + \theta_0) \geq 0 \\ 0 & \text{if } 1 - y^{(i)} (\theta^\top \mathbf{x}^{(i)} + \theta_0) < 0 \end{cases} = \max(1 - y^{(i)} (\theta^\top \mathbf{x}^{(i)} + \theta_0), 0),$$

since, otherwise, it would not be a minimizer.

Now, we can insert $\zeta^{(i)}$ into the objective function and get

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

Minimizing f is equivalent to minimizing f/C , i.e.,

$$\sum_{i=1}^n \max(1 - y^{(i)}(\boldsymbol{\theta}^\top \mathbf{x}^{(i)} + \boldsymbol{\theta}_0), 0) + \lambda \|\boldsymbol{\theta}\|_2^2$$

for $\lambda = 1/(2C)$.

(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.