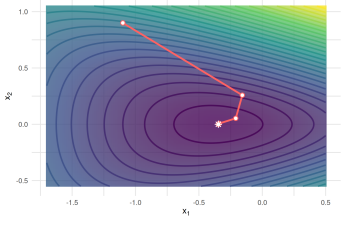
Optimization in Machine Learning

Second order methods
Newton and IRLS

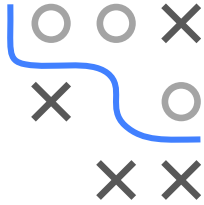


Learning goals

- Setup: Linear predictor + separable loss
- Structure of gradient and Hessian
- Newton as weighted least squares

Slides based on
▸ Aggarwal 2020 (Ch. 5.5)

SETUP: LINEAR PREDICTOR AND SEPARABLE LOSS

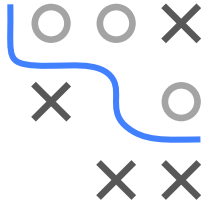


- We now look at L2-regularized ERM problems with a **linear predictor** and a loss that is **separable over the observations**:
  - **(A1) Linearity:** scalar score $f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right) = \boldsymbol{\theta}^\top \boldsymbol{x}^{(i)} = (\mathbf{X}\boldsymbol{\theta})_i$
  - **(A2) Separability:** obs. i only enters additively via its own loss term $L\left(y^{(i)}, f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)$

- This yields objectives of the form

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right)\right) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$$

THREE EXAMPLES FITTING THIS PATTERN



The following three (regularized) models are of exactly this form – they only differ in the point-wise loss $L\left(y^{(i)}, f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)$:

- **Least squares regression** ($y^{(i)} \in \mathbb{R}$): quadratic loss

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{1}{2} \left(f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right) - y^{(i)}\right)^2 + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$$

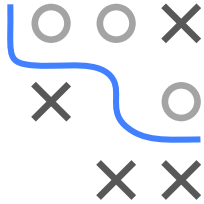
- **Logistic regression** ($y^{(i)} \in \{0, 1\}$): Bernoulli loss with $\pi^{(i)} = s\left(f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)$

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \left[y^{(i)} \log(\pi^{(i)}) + (1 - y^{(i)}) \log(1 - \pi^{(i)})\right] + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$$

- **L2-SVM** ($y^{(i)} \in \{-1, +1\}$): squared hinge loss

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \max\left(0,\, 1 - y^{(i)} f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2 + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$$

NEWTON'S METHOD FOR THIS PROBLEM CLASS



- **Recall** the Newton-Raphson update applied to $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$:

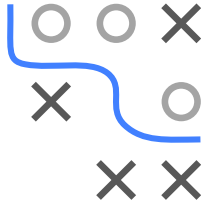
$$\boldsymbol{\theta}^{[t+1]} = \boldsymbol{\theta}^{[t]} - (\nabla^2 \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}^{[t]}))^{-1} \nabla \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}^{[t]})$$

- **Goal:** We will now show that for problems of this kind, each step of Newton's method can be seen as solving a **weighted least squares** problem – with weights and "working responses" that change over the iterations and can be interpreted:

Iteratively Reweighted Least Squares (IRLS)

- **Plan:**
  - **Step 1:** derive the structure of the gradient $\nabla \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$
  - **Step 2:** derive the structure of the Hessian $\nabla^2 \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$
  - **Step 3:** rewrite the Newton step as a weighted LS problem

STEP 1: STRUCTURE OF THE GRADIENT



- Quantity of interest: $\nabla \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) \in \mathbb{R}^{1 \times d}$, with

$$\nabla \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} L\left(y^{(i)}, f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right)\right) + \lambda \boldsymbol{\theta}^{\top}$$

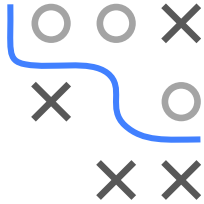
- Chain rule on each loss term:

$$\nabla_{\boldsymbol{\theta}} L\left(y^{(i)}, f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right)\right) = \underbrace{\frac{\partial L\left(y^{(i)}, f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)}{\partial f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right)}}_{=: r_i \in \mathbb{R} \text{ (residual)}} \underbrace{\nabla_{\boldsymbol{\theta}} f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right)}_{= \left(\boldsymbol{x}^{(i)}\right)^{\top} \text{ by (A1)}} \in \mathbb{R}^{1 \times d}$$

- Summing over the observations, with $\mathbf{r} := (r_1, \ldots, r_n)^{\top} \in \mathbb{R}^n$:

$$\boxed{\nabla \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \mathbf{r}^{\top} \mathbf{X} + \lambda \boldsymbol{\theta}^{\top}} \in \mathbb{R}^{1 \times d}$$

STEP 2: STRUCTURE OF THE HESSIAN



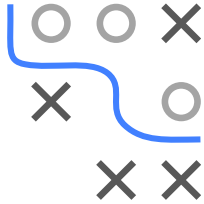
- Differentiate the gradient from step 1, using that the only $\boldsymbol{\theta}$-dependence in $\mathbf{r}^\top \mathbf{X}$ sits inside $\mathbf{r}$:

$$\nabla^2 \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}^\top} \Big(\underbrace{\mathbf{r}^\top \mathbf{X}}_{= \sum_{i=1}^{n} r_i \left(\boldsymbol{x}^{(i)}\right)^\top} + \lambda \boldsymbol{\theta}^\top \Big) = \sum_{i=1}^{n} \frac{\partial r_i}{\partial \boldsymbol{\theta}^\top} \left(\boldsymbol{x}^{(i)}\right)^\top + \lambda \boldsymbol{I}$$

- Differentiating one residual entry:

$$\frac{\partial r_i}{\partial \boldsymbol{\theta}^\top} = \underbrace{\frac{\partial^2 L\left(y^{(i)}, f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)}{\partial f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right)^2}}_{=: d_i \in \mathbb{R}} \underbrace{\frac{\partial f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}^\top}}_{= \boldsymbol{x}^{(i)} \text{ by (A1)}} = d_i \, \boldsymbol{x}^{(i)} \in \mathbb{R}^d$$

STEP 2: STRUCTURE OF THE HESSIAN



- Plugging in yields the Hessian in sum form:

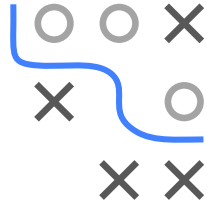
$$\nabla^2 \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} d_i \, \boldsymbol{x}^{(i)} \left(\boldsymbol{x}^{(i)} \right)^\top + \lambda \boldsymbol{I} \in \mathbb{R}^{d \times d}$$

- In matrix notation, stacking the d_i into $\mathbf{D} := \operatorname{diag}(d_1, \ldots, d_n) \in \mathbb{R}^{n \times n}$:

$$\boxed{\nabla^2 \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \mathbf{X}^\top \mathbf{D} \mathbf{X} + \lambda \boldsymbol{I}}$$

- **Note:** the loss only impacts the Hessian through the diagonal entries d_i of $\mathbf{D}$, while the structure $\mathbf{X}^\top \mathbf{D} \mathbf{X}$ itself is fixed by the setup (A1) + (A2); $\mathbf{D}$ is diagonal here because the loss is separable over the observations (A2)

STEP 3: NEWTON STEP AS WEIGHTED LEAST SQUARES



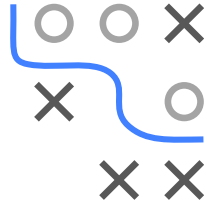
- Plug gradient and Hessian into the Newton-Raphson update, where $\mathbf{r}$ and $\mathbf{D}$ are evaluated at the current iterate $\boldsymbol{\theta}^{[t]}$:

$$\begin{aligned}
\boldsymbol{\theta}^{[t+1]} &= \boldsymbol{\theta}^{[t]} - \left(\mathbf{X}^\top \mathbf{D}\mathbf{X} + \lambda \boldsymbol{I}\right)^{-1} \left(\mathbf{X}^\top \mathbf{r} + \lambda \boldsymbol{\theta}^{[t]}\right) \\
&= \left(\mathbf{X}^\top \mathbf{D}\mathbf{X} + \lambda \boldsymbol{I}\right)^{-1} \left(\left(\mathbf{X}^\top \mathbf{D}\mathbf{X} + \lambda \boldsymbol{I}\right) \boldsymbol{\theta}^{[t]} - \mathbf{X}^\top \mathbf{r} - \lambda \boldsymbol{\theta}^{[t]}\right) \\
&= \left(\mathbf{X}^\top \mathbf{D}\mathbf{X} + \lambda \boldsymbol{I}\right)^{-1} \left(\mathbf{X}^\top \mathbf{D}\mathbf{X}\boldsymbol{\theta}^{[t]} - \mathbf{X}^\top \mathbf{r}\right) \\
&= \left(\mathbf{X}^\top \mathbf{D}\mathbf{X} + \lambda \boldsymbol{I}\right)^{-1} \mathbf{X}^\top \mathbf{D} \underbrace{\left(\mathbf{X}\boldsymbol{\theta}^{[t]} - \mathbf{D}^{-1}\mathbf{r}\right)}_{=: \tilde{\mathbf{f}} \in \mathbb{R}^n}
\end{aligned}$$

- Here, $\tilde{\mathbf{f}}$ is called the **working response** (assuming $d_i > 0$), which will take the role of a pseudo-target; entry-wise

$$\tilde{f}^{(i)} = f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}^{[t]}\right) - r_i / d_i$$

STEP 3: NEWTON STEP AS WEIGHTED LEAST SQUARES

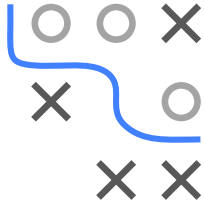


- This Newton iterate solves a **regularized weighted least squares** problem with working responses $\tilde{f}^{(i)}$ as targets and weights d_i:

$$\boldsymbol{\theta}^{[t+1]} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2} \sum_{i=1}^{n} d_i \left(\tilde{f}^{(i)} - \boldsymbol{\theta}^\top \boldsymbol{x}^{(i)} \right)^2 + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2$$

- Since **r** and **D** depend on the current iterate $\boldsymbol{\theta}^{[t]}$, the weights d_i and working responses $\tilde{f}^{(i)}$ change in every iteration
- The iterative Newton method thus repeatedly solves *reweighted* LS problems – hence the name **Iteratively Reweighted Least Squares (IRLS)**

EXAMPLE: LEAST SQUARES REGRESSION



- Quadratic loss $L\left(y^{(i)}, f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right)\right) = \frac{1}{2}\left(f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right) - y^{(i)}\right)^2$, so

$$r_i = f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right) - y^{(i)}, \qquad d_i = 1$$

- Weights and working responses **collapse** to $\mathbf{D} = \boldsymbol{I}$ and $\tilde{\mathbf{f}} = \boldsymbol{y}$, since

$$\tilde{f}^{(i)} = f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}^{[t]}\right) - \frac{f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}^{[t]}\right) - y^{(i)}}{1} = y^{(i)}$$

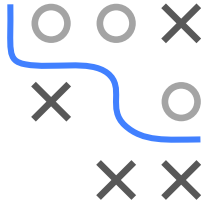
- Plugging into the Newton step:

$$\boldsymbol{\theta}^{[t+1]} = \left(\mathbf{X}^\top \mathbf{D} \mathbf{X} + \lambda \boldsymbol{I}\right)^{-1} \mathbf{X}^\top \mathbf{D} \tilde{\mathbf{f}} = \left(\mathbf{X}^\top \mathbf{X} + \lambda \boldsymbol{I}\right)^{-1} \mathbf{X}^\top \boldsymbol{y}$$

  recovering the well-known (regularized) LS estimator

- Since, in this specific case, neither $\mathbf{D}$ nor $\tilde{\mathbf{f}}$ depends on $\boldsymbol{\theta}^{[t]}$, the problem is solved in **one** Newton step (from any starting point)

EXAMPLE: LOGISTIC REGRESSION



- Using $s'(f) = s(f)(1 - s(f))$, for the Bernoulli loss it follows:

$$r_i = -\left(\frac{y^{(i)}}{\pi^{(i)}} - \frac{1 - y^{(i)}}{1 - \pi^{(i)}}\right) s'\left(f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right)\right) = \pi^{(i)} - y^{(i)},$$

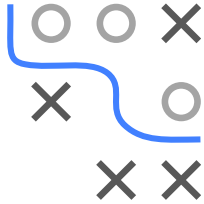
$$d_i = \frac{\partial r_i}{\partial f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right)} = s'\left(f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right)\right) = \pi^{(i)}\left(1 - \pi^{(i)}\right),$$

$$\tilde{f}^{(i)} = f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}^{[t]}\right) - \frac{\pi^{(i)} - y^{(i)}}{\pi^{(i)}(1 - \pi^{(i)})}$$

(for a detailed derivation see ▸ Aggarwal 2020)

- **Interpretation:** Weights $d_i = \pi^{(i)}(1 - \pi^{(i)}) \in (0, \frac{1}{4}]$ are small for confidently classified obs. – they barely influence the update

EXAMPLE: L2-SVM

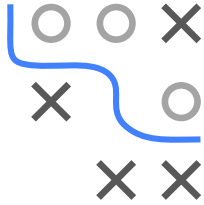


- Define the active set of **margin violators** at $\theta^{[t]}$:
  $A := \{i : y^{(i)} f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}^{[t]}\right) < 1\}$, which changes with the iterate $\boldsymbol{\theta}^{[t]}$
- Residuals and weights:

$$r_i = \begin{cases} -2y^{(i)}\left(1 - y^{(i)} f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}\right)\right) & i \in A \\ 0 & \text{otherwise} \end{cases}, \qquad d_i = \begin{cases} 2 & i \in A \\ 0 & \text{otherwise} \end{cases}$$

  (for a detailed derivation see ▸ Aggarwal 2020)
- For $i \notin A$: $d_i = 0$, so the obs. simply gets weight zero in the WLS

EXAMPLE: L2-SVM



- For $i \in A$, the working response collapses to the label (using $(y^{(i)})^2 = 1$):

$$\tilde{f}^{(i)} = f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}^{[t]}\right) + y^{(i)} \left(1 - y^{(i)} f\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}^{[t]}\right)\right) = y^{(i)}$$

- Consequently, the Newton step is a ridge regression on the current margin violators with their true labels as targets ($\mathbf{X}_A$, $\boldsymbol{y}_A$: rows/entries with $i \in A$):

$$\boldsymbol{\theta}^{[t+1]} = \left(2\mathbf{X}_A^\top \mathbf{X}_A + \lambda \boldsymbol{I}\right)^{-1} 2\mathbf{X}_A^\top \boldsymbol{y}_A$$