Example (regression $y \in \mathbb{R}$)

\[ l(u, v) = \frac{1}{2} (u - v)^2 \quad \text{square loss / } \ell_2 \]
\[ e(u, v) = |u - v| \quad \text{absolute loss / } \ell_1 \]
\[ l(u, v) = |u - v|^p \quad \text{p-loss / } \ell_p \]

If we take $l = \ell_2$ loss, then

\[ R(f) = \mathbb{E} l(f(x), y) = \mathbb{E} (f(x) - y)^2 \]

mean-squared prediction error of $f$ (MSE)

\[ = \mathbb{E}_x \left[ \mathbb{E}_y [(f(x) - y)^2 | x] \right] \]
Now that

\[ E_y \left[ (f(x) - y)^2 \right | x \] 

\[ = E_y \left[ f(x)^2 - 2f(x)y + y^2 \right | x \] 

\[ = f(x)^2 - 2f(x)E[y | x] + E[y^2 | x] \]

so this minimized when

\[ 2f(x) - 2E[y | x] = 0 \]

\[ f(x) = E[y | x] \]
so the predictor \( f^* \) that minimizes the MSE risk in estimating \( \mathcal{Y} \) is the conditional expectation of \( \mathcal{Y} \) given \( X \),

\[
E[\mathcal{Y}|X]
\]

\( f^* \) can be arbitrarily complicated, and we restrict ourselves to a particular function class, so

\[
f^* = \arg\min_{f \in \mathcal{F}} R(f)
\]

can be thought of as the best estimate of \( E[\mathcal{Y}|X] \) in \( \mathcal{F} \).
Fact: if $d(u,v) = |u - v|

then $f^* = \operatorname{Med}(\mathcal{Q}(X))$
Recall we don’t have access to \( P \)
so we do EM

\[
\hat{f} \in \mathcal{F} = \arg \min_{f \in \mathcal{F}} R_n(f)
\]

\[
= \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} l(f(x_i), y_i)
\]

and we hope

\[
R(\hat{f}) \approx R(f^*) \approx R(f^*)
\]

optimization & generalization error

approximation error:
how good is \( \mathcal{F} \)
An additional concern is that we use numerical optimization, so we don't actually find \( \hat{f}_T \), instead we find \( \hat{f}_T \) say after \( T \) iterations of algorithm \( A \).

Thus the error (risk) of models learned using ERM breaks down as

\[
\mathbb{E}(A, n, f, T) = R(\hat{f}_T) \\
= R(\hat{f}_T) - R(\hat{f}_{T+1}) + R(\hat{f}_{T+1}) - R(f^*_T) + R(f^*_T) - R(f^*) + R(f^*_T) - R(f^*)
\]

approximation error + \( R(f^*_T) \) generalization gap
In logistic regression (Maximum Likelihood Estimation give rise to ERM)

Assume we have predictor variables $x \in \mathbb{R}^d$ and a target $y \in \{\pm 1\}$, and we can model the log odds of $y=1$ linearly in $x$

$$\ln \left( \frac{p}{1-p} \right) = \omega_0 + \sum_{i=1}^{d} \omega_i x_i$$

Interpretation: probability of success goes up if $\omega_i > 0$ and $x_i$ increases goes down if $\omega_i < 0$ and $x_i$ increases
\[ \therefore P = \frac{e^{\omega_0 + \omega^T x}}{1 + e^{\omega_0 + \omega^T x}} \Rightarrow p = \frac{e^{\omega_0 + \omega^T x}}{1 + e^{\omega_0 + \omega^T x}} \]

\[ p = \frac{1}{1 + e^{-\omega_0 - \omega^T x}} \]

\[ p = \sigma(\omega_0 + \omega^T x) \]

where \( \sigma(x) = \frac{1}{1 + e^{-x}} \) is called the sigmoid function.
Now that means

\[ y \sim \text{Bern}(\sigma(w_0 + w^T x)) \]

Goal: recover \( w_0 \) & \( w \) given training data \( x \) so in the future when we observe \( x \) we can predict \( y \).

Use maximum likelihood principle: choose the \( w_0, w \) that maximize the prob. of your training data.
\[ w_{w_0}^* = \arg\max_{w_0 \in \omega} p(y_1, \ldots, y_n) \]

\[ = \arg\max_{w_0 \in \omega} \frac{1}{n} \log \left( \prod_{i=1}^{n} p(y_i | \omega_0, \omega^T x_i) \right) \]

\[ = \arg\max_{w_0 \in \omega} \frac{1}{n} \log \left( \frac{1}{n} \sum_{i=1}^{n} \log (p(y_i)) \right) \]

\[ = \arg\min_{w_0 \in \omega} \frac{1}{n} \sum_{i=1}^{n} - \log (p(y_i)) \]
The quantity

\[-\log p(y_1, \ldots, y_n) = \sum_{i=1}^{n} -\log (p(y_i))\]

is called the negative \underline{log-likelihood} of the model parameters \(\theta\).

\(\text{ML principle} \Rightarrow \text{choose your model parameters by minimizing the negative log-likelihood}\)
\[ w^*, w^* = \arg \min_{w_0, w} \quad -\frac{1}{n} \sum_{i=1}^{n} \log(p(y_i)) \]

\[ p(y_i) = \begin{cases} 
\sigma(w_0 + w^T x_i) & \text{if } y_i = 1 \\
1 - \sigma(w_0 + w^T x_i) & \text{if } y_i = -1
\end{cases} \]

\[ \text{face: } \frac{1}{1 + e^{-x}} + \frac{e^{-x}}{1 + e^{-x}} = 1 \quad \Rightarrow \quad 1 - \sigma(x) = \sigma(-x) \]

\[ \sigma(x) = \frac{1}{1 + e^{-x}} \]

\[ p(y_i) = \begin{cases} 
\sigma(w_0 + w^T x_i) & \text{if } y_i = 1 \\
\sigma(-w_0 + w^T x_i) & \text{if } y_i = -1
\end{cases} \]
\[ p(y_i) = \sigma(y_i (w_0 + w^T x_i)) \]

\[ w_0^*, w^* = \arg\min_{w_0, w} -\frac{1}{n} \sum \log(\sigma(y_i (w_0 + w^T x_i))) \]

\[ -\log(\sigma(x)) = \log\left(\frac{1}{\sigma(x)}\right) = \log\left(1 + e^{-x}\right) \]

\[ \log \text{loss!} \]

\[ \log \text{logistic regression} \]
This is ERH w/ loss

\[ \ell(u, v) = \log(1 + e^{-uv}) \]

Note we took

\[ y \sim \text{Bern}(\sigma(\omega_0 + \omega^T x)) \]

we could say

\[ y \sim \text{Bern}(\sigma(f(x))) \]

for \( f \in \mathcal{F} \)

then we get non-linear logistic regression

\[ f^* = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-y_i \sigma(f(x_i)))} \]
Next time:
overfitting
regularization
convex optimization
(poly regression)
(lasso, ridge regression)