Optimization in Machine Learning

Optimization Problems
Unconstrained problems

Reg. risk with $\lambda = 1$, $\alpha = 0.5$
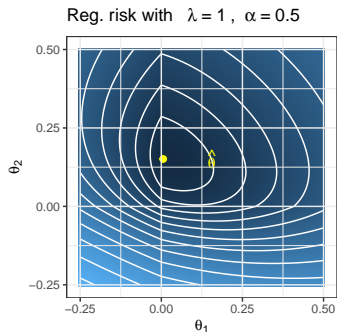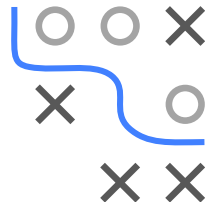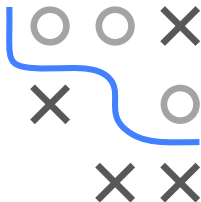
Learning goals

- Definition
- Max. likelihood
- Linear regression
- Regularized risk minimization
- SVM
- Neural network

UNCONSTRAINED OPTIMIZATION PROBLEM

$$\min_{\mathbf{x} \in \mathcal{S}} f(\mathbf{x})$$

with objective function

$$f : \ \mathcal{S} \to \mathbb{R}.$$

The problem is called

- **unconstrained**, if the domain $\mathcal{S}$ is not restricted:

$$\mathcal{S} = \mathbb{R}^d$$

- **smooth** if f is at least $\in \mathcal{C}^1$
- **univariate** if $d = 1$, and **multivariate** if $d > 1$.
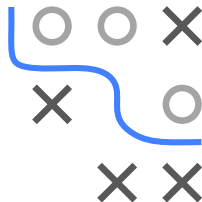- **convex** if f convex function and $\mathcal{S}$ convex set

NOTE: A CONVENTION IN OPTIMIZATION

W.l.o.g., we always **minimize** functions f.

Maximization results from minimizing $-f$.

The solution to maximizing f (left) is equivalent to the solution to minimizing f (right).

EXAMPLE 1: MAXIMUM LIKELIHOOD

$\mathcal{D} = \left(\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)} \right) \stackrel{\text{i.i.d.}}{\sim} f(\mathbf{x} \mid \mu, \sigma)$ with $\sigma = 1$:

$$f(\mathbf{x} \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\mathbf{x} - \mu)^2}{2\sigma^2} \right)$$

Goal: Find $\mu \in \mathbb{R}$ which makes observed data most likely.

Normal density, $\sigma^2 = 1$

EXAMPLE 1: MAXIMUM LIKELIHOOD / 2

- **Likelihood:**
$$\mathcal{L}(\mu \mid \mathcal{D}) = \prod_{i=1}^{n} f\left(\mathbf{x}^{(i)} \mid \mu, 1\right) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}^{(i)} - \mu)^2\right)$$

- **Neg. log-likelihood:**
$$-\ell(\mu, \mathcal{D}) = -\log \mathcal{L}(\mu \mid \mathcal{D}) = \frac{n}{2}\log(2\pi) + \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}^{(i)} - \mu)^2$$
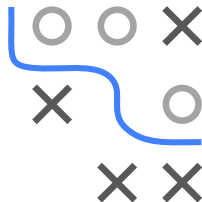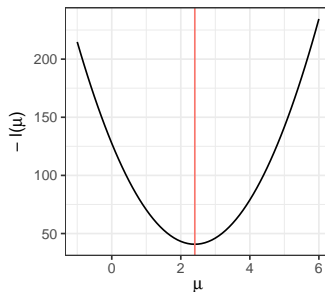

Min. neg. log. likelihood
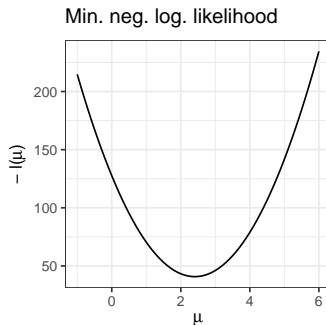
EXAMPLE 1: MAXIMUM LIKELIHOOD / 3

$$\min_{\mu \in \mathbb{R}} -\ell(\mu, \mathcal{D}).$$

can be solved analytically (setting the first deriv. to 0) since it is a quadratic form:

$$-\frac{\partial \ell(\mu, \mathcal{D})}{\partial \mu} = \sum_{i=1}^{n} \left(\mathbf{x}^{(i)} - \mu \right) \quad = \quad 0 \quad \Leftrightarrow \quad \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}^{(i)}$$

Min. neg. log. likelihood

EXAMPLE 1: MAXIMUM LIKELIHOOD

Note: The problem was **smooth**, **univariate**, **unconstrained**, **convex**.
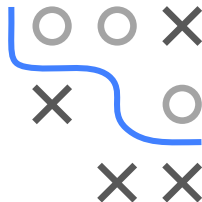
If we had optimized for σ as well

$$\min_{\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+} -\ell(\mu, \mathcal{D}).$$

(instead of assuming it is known) the problem would have been:

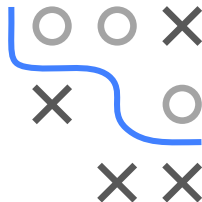- bivariate (optimize over (μ, σ))
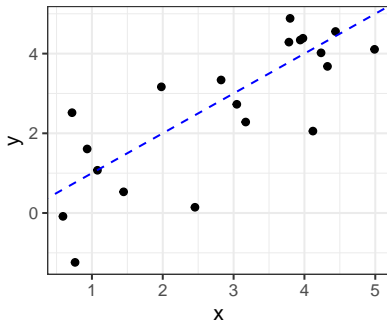- constrained ($\sigma > 0$)

$$\min_{\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+} -\ell(\mu, \mathcal{D}).$$

EXAMPLE 2: NORMAL REGRESSION

Assume (multivariate) data $\mathcal{D} = \left(\left(\mathbf{x}^{(1)}, y^{(1)} \right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)} \right) \right)$
and we want to fit a linear function to it

$$y = f(\mathbf{x}) = \boldsymbol{\theta}^{\top} \mathbf{x}$$

EXAMPLE 2: LEAST SQUARES LINEAR REGR.

Find param vector $\boldsymbol{\theta}$ that minimizes SSE / risk with L2 loss

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \sum_{i=1}^{n} \left(\boldsymbol{\theta}^\top \mathbf{x}^{(i)} - y^{(i)} \right)^2$$
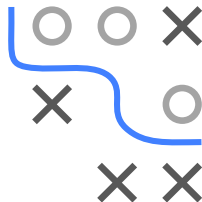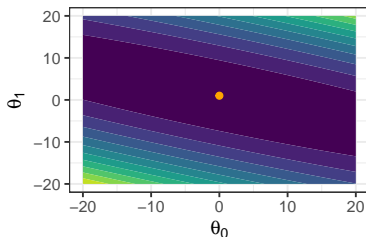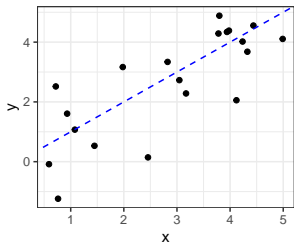
- **Smooth**, **multivariate**, **unconstrained**, **convex** problem
- Quadratic form
- Analytic solution: $\boldsymbol{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}$, where $\mathbf{X}$ is design matrix

RISK MINIMIZATION IN ML

In the above example, if we exchange

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \sum_{i=1}^{n} \left(\boldsymbol{\theta}^\top \mathbf{x}^{(i)} - y^{(i)} \right)^2$$
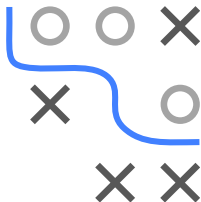
- the linear model $\boldsymbol{\theta}^\top \mathbf{x}$ by an arbitrary model $f(\mathbf{x} \mid \boldsymbol{\theta})$
- the L2-loss $(f(\mathbf{x} \mid \boldsymbol{\theta}) - y)^2$ by any loss $L(y, f(\mathbf{x}))$

we arrive at general **empirical risk minimization** (ERM)

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta} \right) \right) = \min!$$

Usually, we add a regularizer to counteract overfitting:

$$\mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta} \right) \right) + \lambda J(\boldsymbol{\theta}) = \min!$$

RISK MINIMIZATION IN ML / 2

ML models usually consist of the following components:

ML = $\underbrace{\textbf{Hypothesis Space + Risk + Regularization}}_{\text{Formulating the optimization problem}}$ + $\underbrace{\textbf{Optimization}}_{\text{Solving it}}$

- **Hypothesis Space:** Parametrized function space
- **Risk:** Measure prediction errors on data with loss *L*
- **Regularization:** Penalize model complexity
- **Optimization:** Practically minimize risk over parameter space
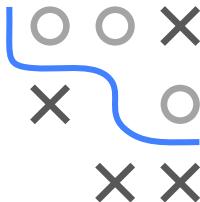
EXAMPLE 3: REGULARIZED LM

ERM with L2 loss, LM, and L2 regularization term:

$$\mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(\boldsymbol{\theta}^\top \mathbf{x}^{(i)} - y^{(i)} \right)^2 + \lambda \cdot \|\boldsymbol{\theta}\|_2^2 \quad \text{(Ridge regr.)}$$
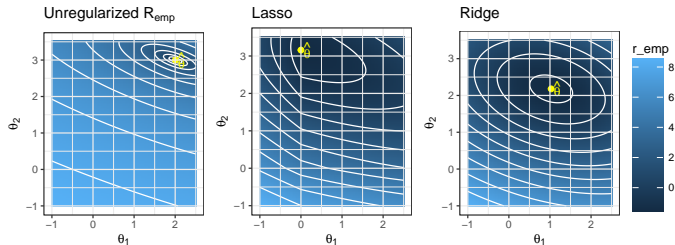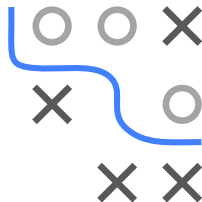
Problem **multivariate**, **unconstrained**, **smooth**, **convex** and has analytical solution
$\boldsymbol{\theta} = (\mathbf{X}^\top \mathbf{X} + \lambda \boldsymbol{I})^{-1} \mathbf{X}^\top \boldsymbol{y}$.

ERM with L2-loss, LM, and L1 regularization:

$$\mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(\boldsymbol{\theta}^\top \mathbf{x}^{(i)} - y^{(i)} \right)^2 + \lambda \cdot \|\boldsymbol{\theta}\|_1 \quad \text{(Lasso regr.)}$$

The problem is still **multivariate**, **unconstrained**, **convex**, but **not smooth**.
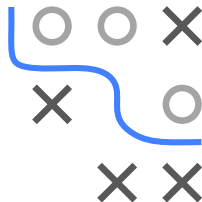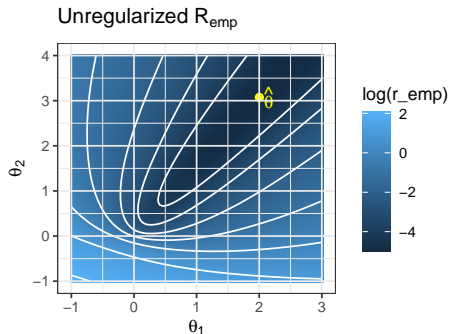
EXAMPLE 4: (REGULARIZED) LOG. REGRESSION

For $y \in \{0, 1\}$ (classification), logistic regression minimizes
log / Bernoulli / cross-entropy loss over data

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(-y^{(i)} \cdot \boldsymbol{\theta}^{\top} \mathbf{x}^{(i)} + \log(1 + \exp\left(\boldsymbol{\theta}^{\top} \mathbf{x}^{(i)}\right)\right)$$
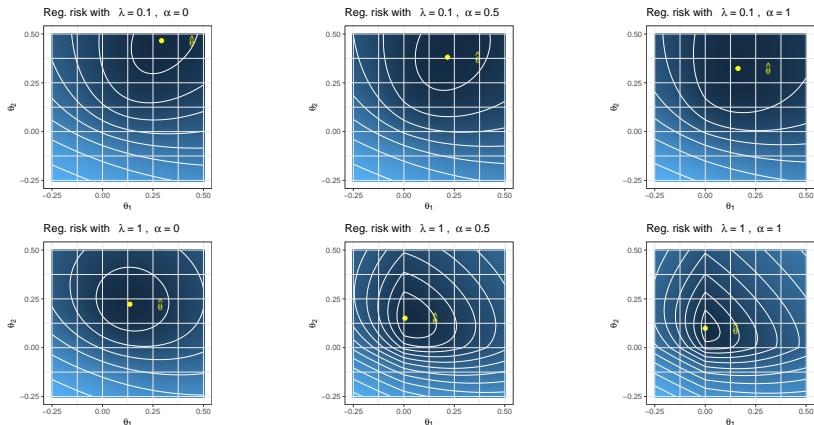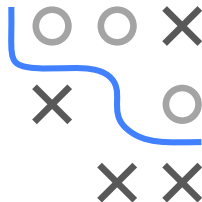
Multivariate, unconstrained, smooth, convex, not analytically solvable.

Unregularized R_{emp}

EXAMPLE 4: (REGULARIZED) LOG. REGRESSION / 2

Elastic net regularization is a combination of L1 and L2 regularization

$$\frac{1}{2n} \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right) + \lambda \left[\frac{1-\alpha}{2}\|\theta\|_2^2 + \alpha\|\theta\|_1\right], \lambda \geq 0, \alpha \in [0, 1]$$

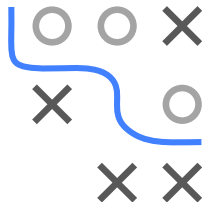The higher λ, the closer to the origin, L1 shrinks coeffs exactly to 0.

EXAMPLE 4: (REGULARIZED) LOG. REGRESSION
/ **3**

$$\frac{1}{2n} \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right) + \lambda \left[\frac{1-\alpha}{2}\|\theta\|_2^2 + \alpha\|\theta\|_1\right], \lambda \geq 0, \alpha \in [0, 1]$$

Problem characteristics:

- Multivariate
- Unconstrained
- If $\alpha = 0$ (Ridge) problem is smooth; not smooth otherwise
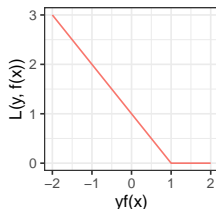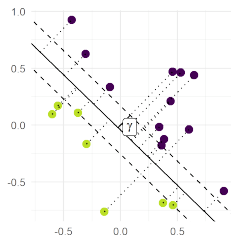- Convex since *L* convex and both L1 and L2 norm are convex

EXAMPLE 5: LINEAR SVM

- $\mathcal{D} = \left((\mathbf{x}^{(i)}, y^{(i)}) \right)_{i=1,\dots,n}$ with $y^{(i)} \in \{-1, 1\}$ (classification)
- $f(\mathbf{x} \mid \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{x} \in \mathbb{R}$ scoring classifier:
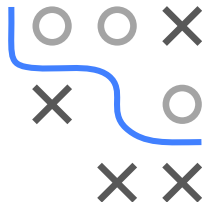 Predict 1 if $f(\mathbf{x} \mid \boldsymbol{\theta}) > 0$ and -1 otherwise.

ERM with LM, hinge loss, and L2 regularization:

$$\mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \max\left(1 - y^{(i)} f^{(i)}, 0 \right) + \lambda \boldsymbol{\theta}^\top \boldsymbol{\theta}, \quad f^{(i)} := \boldsymbol{\theta}^\top \mathbf{x}^{(i)}$$
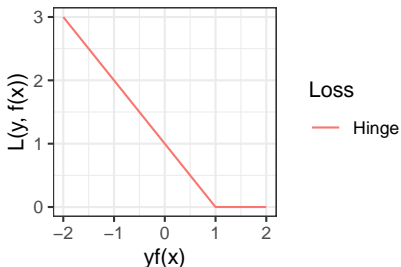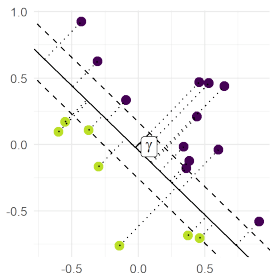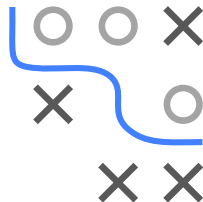

This is one formulation of the **linear SVM**. Problem is: **multivariate**, **unconstrained**, **convex**, but **not smooth**.

EXAMPLE 5: LINEAR SVM / 2

Understanding hinge loss $L(y, f(\mathbf{x})) = \max(1 - y \cdot f, 0)$

y	$f(\mathbf{x})$	Correct pred.?	$L(y, f(\mathbf{x}))$	Reason for costs
1	$(-\infty, 0)$	N	$(1, \infty)$	Misclassification
-1	$(0, \infty)$	N	$(1, \infty)$	Misclassification
1	$(0, 1)$	Y	$(0, 1)$	Low confidence / margin
-1	$(-1, 0)$	Y	$(0, 1)$	Low confidence / margin
1	$(1, \infty)$	Y	0	–
-1	$(-\infty, -1)$	Y	0	–

EXAMPLE 6: KERNELIZED SVM

Kernelized formulation of the primal[(*)] SVM problem:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} L\left(y^{(i)}, \boldsymbol{K}_i^\top \boldsymbol{\theta}\right) + \lambda \boldsymbol{\theta}^\top \boldsymbol{K} \boldsymbol{\theta}$$

with $k(\cdot, \cdot)$ pos. def. kernel function, and
$\boldsymbol{K}_{ij} := k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$, $n \times n$ psd kernel matrix, $\boldsymbol{K}_i$ i-th column of K.

Kernelization

- allows introducing nonlinearity through projection into higher-dim. feature space
- without changing problem characteristics (convexity!)
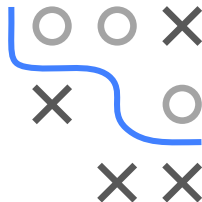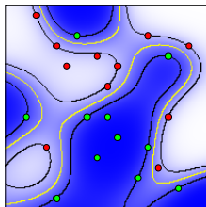
[(*)] There is also a dual formulation to the problem (comes later!)

EXAMPLE 6: NEURAL NETWORK

Normal loss, but complex *f* defined as computational feed-forward
graph. Complexity of optimization problem

$$\arg \min_{\boldsymbol{\theta}} \mathcal{R}_{\text{reg}}(\boldsymbol{\theta}),$$

so smoothness (maybe) or convexity (usually no) is influenced by loss,
neuron function, depth, regularization, etc.
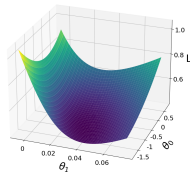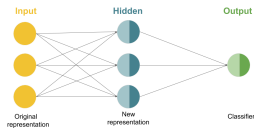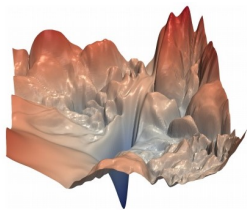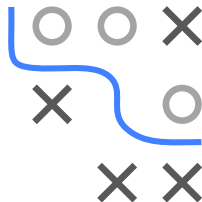
Loss landscapes of ML problems.

Left: Deep learning model ResNet-56, right: Logistic regression with cross-entropy loss

Source: https://arxiv.org/pdf/1712.09913.pdf