Optimization

- Many more complex models in statistics and machine learning do not have analytic solutions to ML estimators.
- In most models parameters are learned by some numerical optimization technique.

$$\min_{\theta} F(\theta)$$

- How many minima are there?
- How do we find optimal θ?
- Are we guaranteed to find the global optimum θ^*, rather just a local one?
- How efficiently can we solve for θ?
- What if there are constraints?

Constrained Optimization

- Optimization problems with constraints, e.g.

$$\begin{aligned} & \min_{\theta \in \mathbb{R}^d} F(\theta) & \\ \text{subject to} \quad & g_i(\theta) \leq 0 & \text{for } i = 1, \ldots, I \\ & h_j(\theta) = 0 & \text{for } j = 1, \ldots, J \end{aligned}$$

  where g_i enforce inequality constraints and h_j equality constraints.
- Can write this succinctly:

$$\begin{aligned} & \min_{\theta \in \mathbb{R}^d} F(\theta) \\ \text{subject to} \quad & g(\theta) \preceq 0 \\ & h(\theta) = 0 \end{aligned}$$

  where $g : \mathbb{R}^d \to \mathbb{R}^I$ is a vector-valued function with $g(\theta)_i = g_i(\theta)$. Similarly $h(\theta) : \mathbb{R}^d \to \mathbb{R}^J$. $x \preceq y$ iff $x_i \leq y_i \forall i$.
- These problems are called **programmes**.

Constrainted Optimization

$$\min_{\theta \in \mathbb{R}^d} F(\theta)$$

$$\text{subject to} \quad g(\theta) \preceq 0$$

$$h(\theta) = 0$$

- We can enforce constraints by using **Lagrange multipliers** or **dual variables** $\lambda \in \mathbb{R}^I$ and $\kappa \in \mathbb{R}^J$.
- The optimization problem can be written as a mini-max optimization of the Lagrangian:

$$\min_{\theta} \max_{\lambda \succeq 0, \kappa} \mathcal{L}(\theta, \lambda, \kappa) = \min_{\theta} \max_{\lambda \succeq 0, \kappa} F(\theta) + \lambda^\top g(\theta) + \kappa^\top h(\theta)$$

- Intuition: For any θ, we have:

$$\max_{\lambda \succeq 0, \kappa} \mathcal{L}(\theta, \lambda, \kappa) = \begin{cases} +\infty & \text{if there is some unsatisfied constraint,} \\ F(\theta) & \text{if all constraints are satisfied.} \end{cases}$$

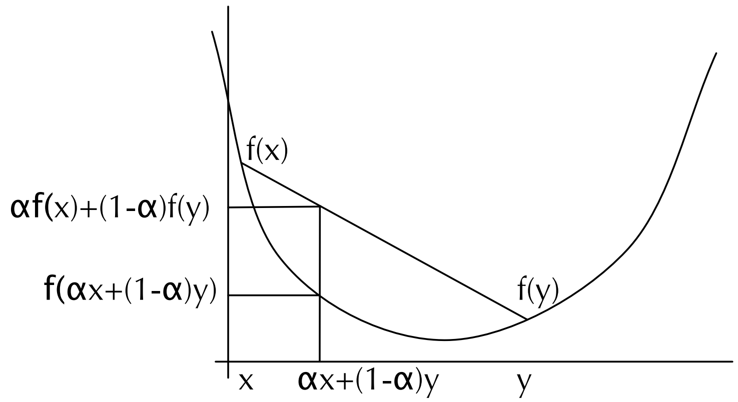
So the outer minimization over θ results in the same optimization problem.

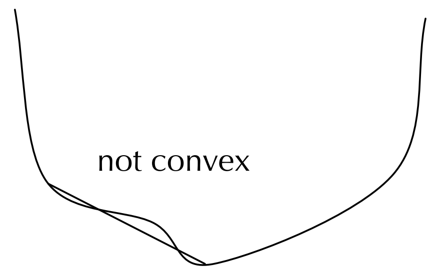
Convex Optimization

- A function $f : \mathbb{R}^d \to \mathbb{R}$ is **convex** if

  $$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y)$$

  for all $x, y \in \mathbb{R}^d$, $\alpha \in [0, 1]$.
- For smooth functions: Equivalent to 2nd derivative (**Hessian**) being positive semidefinite.
- A programme is a **convex programme** if:
  - $F(\theta)$ is convex,
  - $g_i(\theta)$ is convex for each i,
  - $h(\theta) = A\theta + b$ is affine.
- Examples: linear, quadratic, semidefinite programming.
- Convex programmes have a unique minimum (typically), which can be efficiently found.





Convex Duality

- Say the minimum is p^*, and occurred at θ^*.
- The **dual programme** inverts the order of max and min:

$$p^* = \min_{\theta} \max_{\lambda \succeq 0, \kappa} \mathcal{L}(\theta, \lambda, \kappa) \geq \max_{\lambda \succeq 0, \kappa} \min_{\theta} \mathcal{L}(\theta, \lambda, \kappa) = d^*$$

  where the dual optimum is d^*.
- **Karush-Kuhn-Tucker Theorem**: Subject to regularity conditions, a solution θ^* is the optimal solution of a convex programme, if and only if there are λ^* and κ^* (the dual optimal solution) such that:
  - **Primal feasible**: $g(\theta^*) \preceq 0$, $h(\theta^*) = 0$.
  - **Dual feasible**: $\lambda^* \succeq 0$.
  - $(\theta^*, \lambda^*, \kappa^*)$ is a **saddle point** of $\mathcal{L}$: For every $\theta, \lambda \succeq 0, \kappa$, we have

$$\mathcal{L}(\theta^*, \lambda, \kappa) \leq \mathcal{L}(\theta^*, \lambda^*, \kappa^*) \leq \mathcal{L}(\theta, \lambda^*, \kappa^*)$$

  - $\nabla_\theta \mathcal{L}(\theta^*, \lambda^*, \kappa^*) = \nabla_\theta F(\theta^*) + (\lambda^*)^\top \nabla_\theta g(\theta^*) + (\kappa^*)^\top \nabla_\theta h(\theta^*) = 0$
  - **Complementary slackness**: For every i,

$$\lambda_i^* g_i(\theta^*) = 0$$