Supervised Learning (COMP0078)

4. First Order Optimization and ERM

Carlo Ciliberto

University College London
Department of Computer Science

Outline

- Binary Classification via Convex Surrogates
- Logistic Regression
- Gradient Descent
- Sub-gradient Method

Previous Classes

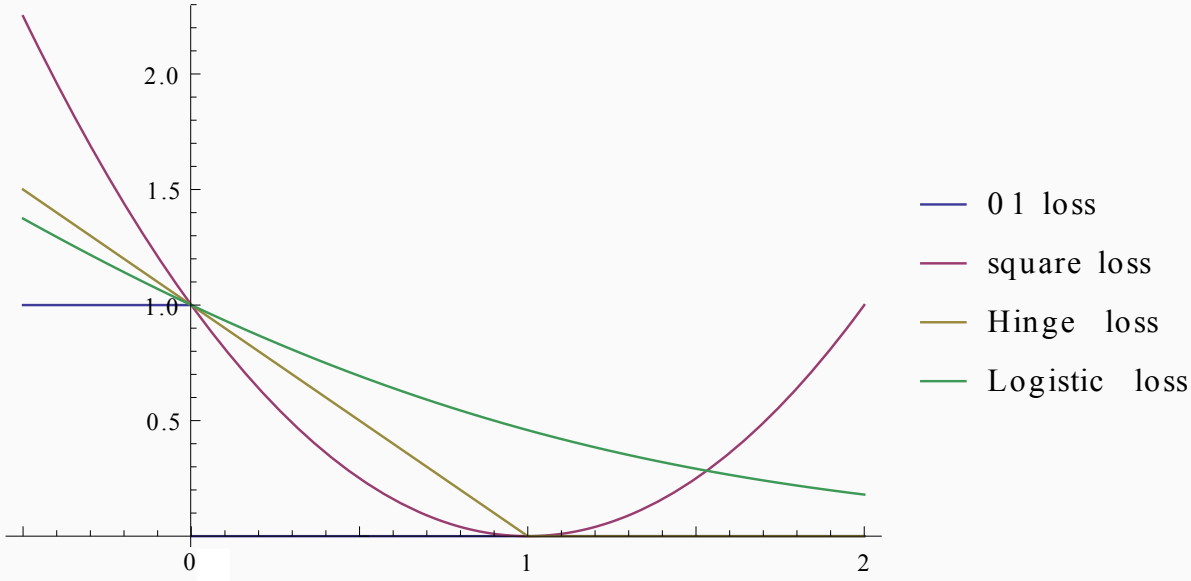
Focusing on (binary) classification, we have observed how different strategies for learning a classifier end up being in some form of ERM (possibly regularised).

$$\min_{w} \frac{1}{m} \sum_{i=1}^{m} \ell(w^\top x_i, y) + \lambda \|w\|^2$$

- **Least Squares:** $\ell(y, y') = (y - y')^2$
- **SVM:** $\ell(y, y') = \max(0, 1 - yy')$
- **Logistic Regression (Today!):** $\ell(y, y') = \log(1 + e^{-yy'})$
- ...

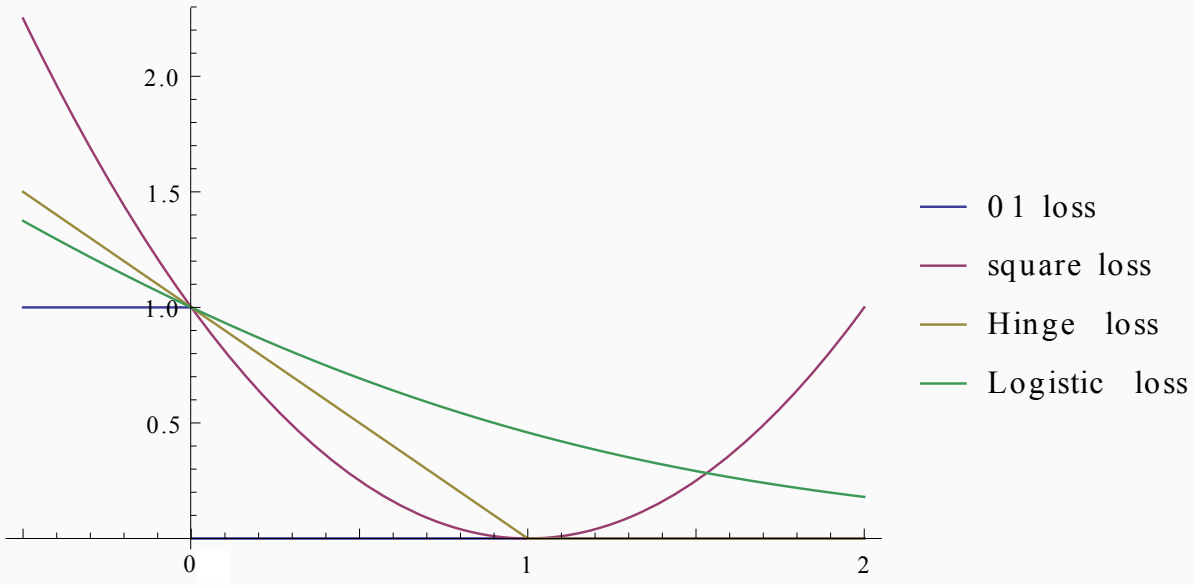
Matching signs

A binary classification loss measures studies the interplay between $f(x)$ and the sign of y as $\ell(f(x), y) = \tilde{\ell}(f(x)y)$



- 0-1 loss $\tilde{\ell}(r) = \mathbf{1}_{\{-r>0\}}$
- Square loss $\tilde{\ell}(r) = (1-r)^2$,
- Hinge-loss $\tilde{\ell}(r) = \max(1-r, 0)$,
- logistic loss $\tilde{\ell}(r) = \log(1+\exp(-r))$,

Convex Surrogates



These loss functions come from different interpretations of the problem, but their main appeal is that they are **convex upperbounds** to the binary-classification loss (which would be intractable to minimize).

Interpretations

- Square loss: we know that the Bayes estimator is

$$f_*(x) = \mathbb{E}_{y|x}[y] = P(1|x) - P(-1|x)$$

  $\Rightarrow$ the Bayes estimator for the $0-1$ Loss is recovered by taking the sign $\text{sign}(f_*(x)) = \text{argmax}_y P(y|x)$.

- Hinge loss: we have seen in last class how the geometric interpretation of finding the hyperplane with highest margin between the two classes corresponds to ERM with Hinge loss.

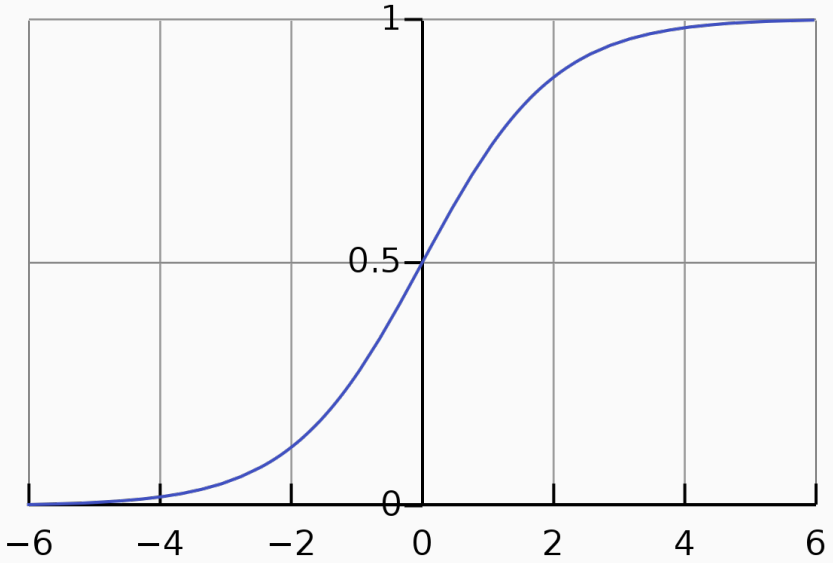
- Logistic loss: ???

Logistic Regression

Logistic regression originates from the parametrization in the scalar setting (i.e. $x \in \mathcal{X} \subset \mathbb{R}$):

$$p_{\mu,s}(x) = P(y = 1|x; \mu, s) = \frac{1}{1 + e^{-(x-\mu)/s}}$$

Namely modeling the probability of y being 1 to follow a sigmoid shape.



Logistic Regression: General Model

By taking $w = 1/s = w$ and $b = -\mu/s$ we have the alternative parametrization

$$\frac{1}{1 + e^{-(x-\mu)/s}} = p_{\mu,s}(x) = p_{w,b}(x) = \frac{1}{1 + e^{-(wx+b)}},$$

which can be further generalized to $\mathcal{X} \subset \mathbb{R}^n$ any feature map $\phi : \mathbb{R}^n \to \mathbb{R}^N$ and w a vector in $\mathbb{R}^N$:

$$p_w(x) = \frac{1}{1 + e^{-w^\top \phi(x)}}$$

For simplicity, in the following we will assume $\phi(x) = x$.

Logistic Regression: Optimization

Given a dataset $S = (x_i, y_i)_{i=1}^m$, we would like to find the model p_w that best fits our data.

Since we are talking about probabilities, we could frame this as finding the w maximising the likelihood of "observing" S:

$$\max_w \; P(y_1, \ldots, y_m | x_1, \ldots, x_m; w)$$

Assuming the data to have been sampled i.i.d., we can decompose the joint probability above as

$$\begin{aligned} P(y_1, \ldots, y_m | x_1, \ldots, x_m; w) &= P(y_1 | x_1; w) \cdots P(y_m | x_m; w) \\ &= \Pi_{i=1}^m P(y_i | x_i; w) \end{aligned}$$

Logistic Regression: Optimization (II)

Since the logarithm is a monotone increasing function, we can reformulate our problem as

$$\max_w \log\left(\Pi_{i=1}^m P(y_i|x_i;w)\right) = \max_w \sum_{i=1}^m \log P(y_i|x_i;w)$$

By definition of the logistic model we have

- if $y=1$, then $P(1|x;w) = \frac{1}{1+e^{-w^\top x}} = \frac{1}{1+e^{-yw^\top x}}$
- if $y=-1$, then $P(-1|x;w) = 1 - \frac{1}{1+e^{-w^\top x}} = \frac{1+e^{-w^\top x}-1}{1+e^{-w^\top x}}$.
  Hence $P(-1|x;w) = \frac{1}{1+e^{w^\top x}} = \frac{1}{1+e^{-yw^\top x}}$

In other words: $P(y|x;w) = \frac{1}{1+e^{-yw^\top x}}$

Logistic Regression: Optimization (III)

Hence, we can finally formulate the optimization problem as

$$\begin{aligned}
\max_w \sum_{i=1}^{m} \log P(y_i|x_i; w) &= \max_w \sum_{i=1}^{m} \log \frac{1}{1 + e^{-y_i w^\top x_i}} \\
&= \max_w \sum_{i=1}^{m} -\log(1 + e^{-y_i w^\top x_i}) \\
&= -\min_w \sum_{i=1}^{m} \underbrace{\log(1 + e^{-y_i w^\top x_i})}_{\text{logistic loss}}
\end{aligned}$$

Logistic Regression and ERM

Therefore, the problem of finding the best model p_w can be interpreted as performing empirical risk minimization (ERM):

$$\min_w \sum_{i=1}^{m} \underbrace{\log(1 + e^{-y_i w^\top x_i})}_{\text{logistic loss}}$$

Like for SVM and the Square loss, the class prediction at a point $x \in \mathcal{X}$ is given by taking the sign of $f(x) = w^\top x$:

This follows by observing that $P(1|x; w) > P(-1|x; w)$ if and only if $\frac{1}{1+e^{-w^\top x}} > \frac{1}{1+e^{w^\top x}}$. Which corresponds to requiring

$$1 + e^{-w^\top x} < 1 + e^{w^\top x}$$

or equivalently $-w^\top x < w^\top x$, which is possible if and only if $\text{sign}(w^\top x)$ is positive.

Solving ERM

So... how do we solve a problem of the form

$$\min_w \frac{1}{m} \sum_{i=1}^{m} \ell(w^\top x_i, y_i) + \lambda \|w\|^2$$

in practice?

Gradient Descent

The importance of being *Smooth*

Let $F : \mathbb{R}^N \to \mathbb{R}$ be a differentiable function (e.g. the square or the logistic loss) that we want to minimize.

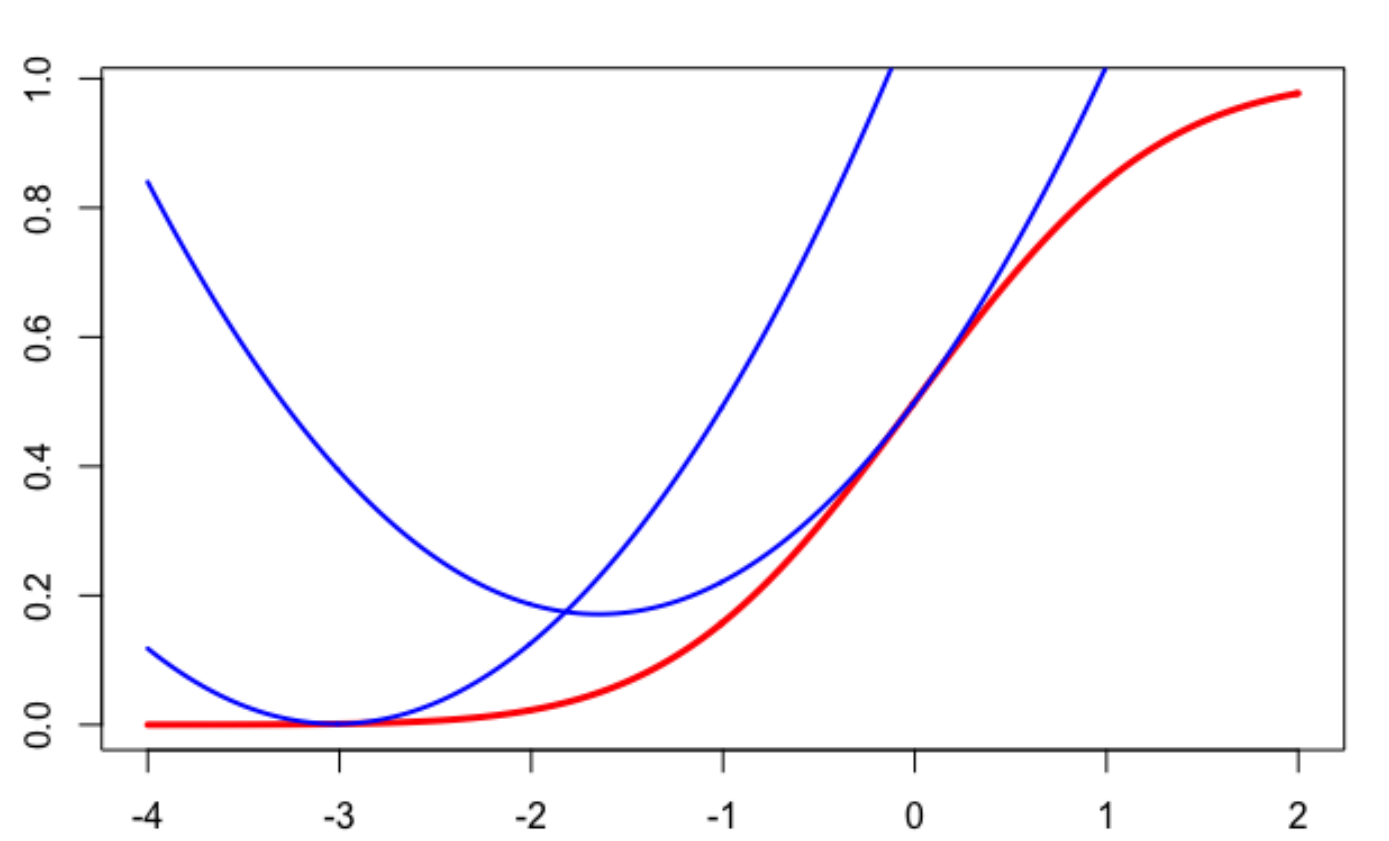
If F is actually smooth[1], we can find an upper bound that will prove very useful!

Theorem (Quadratic Upper-bound). Let F be M-smooth, with $M > 0$. Then for any $w, z \in \mathbb{R}^N$,

$$F(z) \leq F(w) + \nabla F(w)^\top (z - w) + \frac{M}{2} \|z - w\|^2 .$$

[1] in the context of convex analysis. We will define it in a moment.

Quadratic Upper Bound



$$F(z) \leq F(w) + \nabla F(w)^\top (z - w) + \frac{M}{2} \|z - w\|^2 .$$

Towards Gradient Descent

Why is this inequality useful?

$$F(z) \leq F(w) + \nabla F(w)^\top (z - w) + \frac{M}{2} \|z - w\|^2$$

Because, given $w \in \mathbb{R}^N$, the quadratic upperbound suggests a direct strategy to choose z such that $F(z) < F(w)$!

Let us choose z such that $z - w = \eta \nabla F(w)$ for $\eta \in \mathbb{R}$. In other words...

$$z = w + \eta \nabla F(w)$$

Then, we have

$$\begin{aligned} F(z) &\leq F(w) + \eta \|\nabla F(w)\|^2 + \eta^2 \frac{M}{2} \|\nabla F(w)\|^2 \\ &= F(w) + \eta(1 + \eta \frac{M}{2}) \|\nabla F(w)\|^2 \end{aligned}$$

Towards Gradient Descent

We see that for $\eta \in (-\frac{2}{M}, 0)$, the quantity

$$\eta(1 + \eta\frac{M}{2}) < 0$$

is negative! This implies that $F(z) < F(w)$ is decreasing.

For example, if we take $\eta = -\frac{1}{M}$ (which achieves the minimum wrt η), we have

$$F(z) \leq F(w) - \frac{1}{2M} \|\nabla F(w)\|^2$$

Gradient Descent

Gradient Descent (GD) is the strategy that provides an iterative sequence $(w_k)_{k=1}^K$, starting from an initial w_0, such that

$$w_{k+1} = w_k - \eta \nabla F(w_k)$$

for some *step-size* (sometimes also called learning rate in machine learning) $\eta > 0$.

When does this strategy "work"?

In other words: while it is true that by correctly choosing η we can always ensure a decreasing sequence, does GD converge to a minimizer?

Before tackling this question, let us define smoothness and how to prove the quadratic upper bound.

Smoothness

So, what is smoothness and which functions are smooth?

Definition. A function $F : \mathbb{R}^N \to \mathbb{R}$ is M-smooth if it has Lipschitz gradient, with Lipschitz constant M, namely

$$\|\nabla F(w) - \nabla F(z)\| \leq M \|z - w\| \qquad \forall w, z \in \mathbb{R}^N$$

Which Functions are Smooth?

How can we:

- verify whether a function (in this case the gradient of our objective function F) is Lipschitz?
- estimate the Lipschitz constant?

We will get back to these questions once we better understand **why** answering them is important.

Quadratic Upper Bound - Proof

Proof (Quadratic Upper-bound). Define $h : [0, 1] \to \mathbb{R}$ such that for any $\theta \in [0, 1]$

$$h(\theta) = F(w + \theta(z - w)).$$

Since F is differentiable, then h is differentiable and

- $\int_0^1 h'(\theta)\, d\theta = h(1) - h(0)$, and
- $h'(\theta) = \frac{\partial}{\partial \theta} F(w + \theta(z - w)) = \nabla F(w + \theta(z - w))^\top (z - w)$.

Combining the two and observing that $h(0) = F(w)$ and $h(1) = F(z)$, we have

$$F(z) = F(w) + \int_0^1 \nabla F(w + \theta(z - w))^\top (z - w)\, d\theta$$

Quadratic Upper Bound - Proof (II)

We now add and remove the term $\nabla F(w)^\top (z-w)$

$$F(z) = F(w) + \nabla F(w)^\top (z-w) + \int_0^1 \left(\nabla F(w+\theta(z-w)) - \nabla F(w)\right)^\top (z-w)\, d\theta$$

By recalling the Cauchy-Schwartz inequality ($u^\top v \leq \|u\|\,\|v\|$) we can upper bound the term in the integral by

$$\left(\nabla F(w+\theta(z-w)) - \nabla F(w)\right)^\top (z-w) \leq \underbrace{\|\nabla F(w+\theta(z-w)) - \nabla F(w)\|}_{\leq M\theta\|z-w\| \text{ by smoothness}} \|z-w\|$$

Quadratic Upper Bound - Proof (III)

We conclude that

$$F(z) \leq F(w) + \nabla F(w)^\top (z - w) + M \|z - w\|^2 \underbrace{\int_0^1 \theta \, d\theta}_{1/2},$$

as required. ■

We can now go back to our question on whether/when GD converges.

Gradient Descent and Convexity

Convexity

A convex function $F(w)$ is such that for any $w, z \in \mathbb{R}^N$ and any $\theta \in [0, 1]$,

$$F(\theta w + (1 - \theta)z) \leq \theta F(w) + (1 - \theta)F(z)$$

We like convex functions because

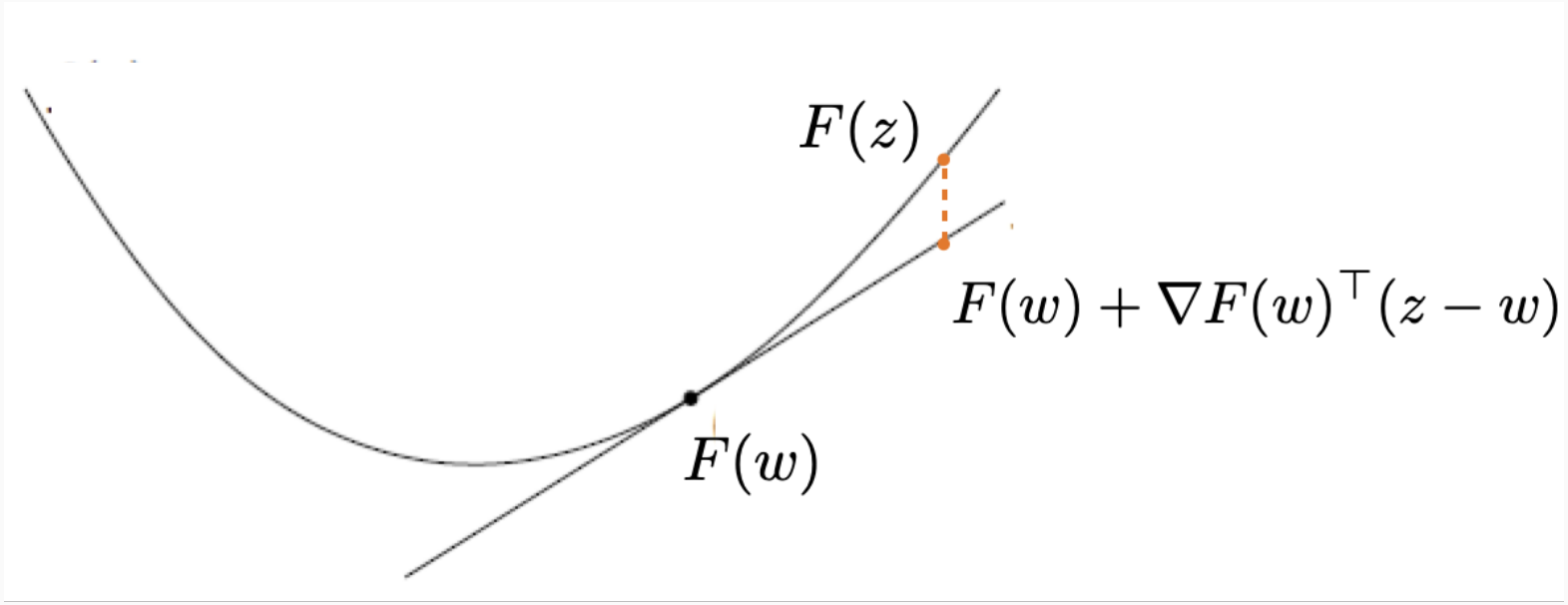
- They do not have local minima[2]
- We have several tools to approximate/find a solution.

[2] other than the global minimum

Convexity and Differentiability

Proposition (First-order condition). Let $F : \mathbb{R}^N \to \mathbb{R}$ be differentiable. Then F is convex if and only if for any $w, z \in \mathbb{R}^N$

$$F(z) \geq F(w) + \nabla F(w)^\top (z - w).$$



Convexity: First-order Condition

$$F(z) \geq F(w) + \nabla F(w)^\top (z - w).$$

Remark 1. If $\nabla F(w_*) = 0$, then for any $z \in \mathbb{R}^N$ we have $F(z) \geq F(w_*)$. Therefore the minimizers of F are all (and only) those such that $\nabla F(w_*) = 0$

Remark 2. While smoothness provides an upper-bound, convexity grants a *lower-bound*, effectively sandwitch-ing our objective F!

Convexity: First-order Condition (II)

Proof. ($\Rightarrow$) Let F be convex and $\theta \in (0, 1]$. From convexity

$$\frac{F(\theta z + (1-\theta)w)}{\theta} \leq F(z) + \frac{(1-\theta)}{\theta} F(w)$$
$$\iff \quad \frac{F(\theta z + (1-\theta)w) - F(w)}{\theta} \leq F(z) - F(w)$$

Since $F(\theta z + (1-\theta)w) = F(w + \theta(z-w))$, we have

$$\lim_{\theta \to 0} \frac{F(w + \theta(z-w)) - F(w)}{\theta} = \nabla F(w)^\top (z-w)$$

by definition of directional derivative. Yielding the desired result.

Convexity: First-order Condition (III)

($\Leftarrow$) Assume now that the first-order condition holds. Let $\theta \in [0, 1]$ and denote $u = \theta w + (1 - \theta)z$. By assumption

- $F(w) \geq F(u) + \nabla F(u)^\top (w - u)$,
- $F(z) \geq F(u) + \nabla F(u)^\top (z - u)$

By multiplying the two inequalities respectively by θ and $(1 - \theta)$ and then summing them up, we conclude

$$\theta F(w) + (1 - \theta)F(z) \geq F(u) + \nabla F(u)^\top \underbrace{(\theta w + (1 - \theta)z - u)}_{=0 \text{ since } u=\theta w+(1-\theta)z}$$

Hence

$$\theta F(w) + (1 - \theta)F(z) \geq F(u) = F(\theta w + (1 - \theta)z)$$

as desired. ■

Convergence of Gradient Descent

We can finally answer the question of whether GD "works". The answer is positive (for convex objectives).

Theorem (GD Rates). Let F be convex and M-smooth. Assume F admits a minimum in $w_* \in \mathbb{R}^N$. Let $(w_k)_{k=1}^K$ be a sequence produced by GD with $\eta = 1/M$. Then

$$F(w_K) - F(w_*) \leq \frac{M}{2K} \|w_0 - w_*\|^2$$

The more iterates we make, the better the quality of our approximation of the minimum!

Convergence of GD (II)

Proof. By the first-order condition on convexity applied to the points w_k and w_*, we know that

$$F(w_*) \geq F(w_k) + \nabla F(w_k)^\top (w_* - w_k)$$

or equivalently

$$F(w_k) \leq F(w_*) + \nabla F(w_k)^\top (w_k - w_*)$$

Combining with what observed for GD with step-size $\eta = 1/M$,

$$\begin{aligned} F(w_{k+1}) &\leq F(w_k) - \frac{1}{2M} \|\nabla F(w_k)\|^2 \\ &\leq F(w_*) + \nabla F(w_k)^\top (w_k - w_*) - \frac{1}{2M} \|\nabla F(w_k)\|^2 \end{aligned}$$

Convergence of GD (III)

We can now add and remove the term $\frac{M}{2}\|w_k - w_*\|$ to "complete the square" and use $w_{k+1} = w_k - \eta\nabla F(w_k)$

$$F(w_{k+1}) \leq F(w_*) + \frac{M}{2}(\|w_k - w_*\|^2 - \|w_{k+1} - w_*\|^2)$$

Or, equivalently

$$F(w_{k+1}) - F(w_*) \leq \frac{M}{2}(\|w_k - w_*\|^2 - \|w_{k+1} - w_*\|^2)$$

Implying that the gap between $F(w_k)$ and the minimum becomes smaller and smaller the more the iterates w_k get closer to the minimizer.

Now, does this happen?

Convergence of GD (Conclusion)

Let's sum over all $k = 1, \ldots, K$

$$\sum_{k=1}^{K} (F(w_k) - F(w_*)) \leq \frac{M}{2} \sum_{k=1}^{K} \left(\|w_{k-1} - w_*\|^2 - \|w_k - w_*\|^2 \right)$$
$$\leq \frac{M}{2} \|w_0 - w_*\|^2$$

Since w_k is a decreasing sequence $F(w_k) \geq F(w_K)$ for any $k = 1, \ldots, K$, we get

$$K(F(w_K) - F(w_*)) \leq \frac{M}{2} \|w_0 - w_*\|^2$$

From which we conclude

$$F(w_K) - F(w_*) \leq \frac{M}{2K} \|w_0 - w_*\|^2$$

as required.

The Importance of Being Smooth

From the discussion above it is clear that the constant M of smoothness of a function plays a central role:

- To choose the step size η in practice.
- For convergence rates.

Therefore we can go back to our earlier question:

How do we estimate the Lipschitz constant of a function?

Example: Square Loss

Let $x \in \mathbb{R}^N$ and $y \in \mathbb{R}$. We want to evaluate the M-smoothness of $\ell(w) = \ell(w^\top x, y) = (w^\top x - y)^2$.

First, the gradient of ℓ is

$$\nabla \ell(w) = 2(w^\top x - y)x$$

Now, for any $w, z \in \mathbb{R}^n$

$$\nabla \ell(w) - \nabla \ell(z) = 2(xx^\top)^\top (w - z)$$

Hence

$$\|\ell(w) - \ell(z)\| \leq 2 \|x\|^2 \|w - z\|,$$

which implies that the square loss is $2 \|x\|^2$ Lipschitz.

Example: Square Loss (II)

Consider now $\mathcal{E}_S(w) = \frac{1}{m}\sum_{i=1}^m (w^\top x_i - y_i)^2 + \lambda \|w\|^2$:

$$\nabla \mathcal{E}_S(w) = 2\frac{1}{m}\sum_{i=1}^m (x_i x_i^\top) w + 2\lambda w - \frac{2}{m}\sum_{i=1}^m y_i x_i$$

$$= 2(C_S + \lambda I)w - \frac{2}{m}\sum_{i=1}^m y_i x_i$$

With $C_S = \frac{1}{m}\sum_{i=1}^m (x_i x_i^\top)$ the uncentered covariance. Hence

$$\|\nabla \mathcal{E}_S(w) - \nabla \mathcal{E}_S(z)\| \leq 2 \underbrace{\|C_S + \lambda I\|}_{\substack{\text{Operator norm} \\ \text{(largest singular value)}}} \|w - z\|$$

From which we conclude that $\mathcal{E}_S$ is $2\|C_S + \lambda I\|$-smooth.

Note: this is a value that can be computed from data!

Estimating the Lipschitz Constant

What about other functions, such as the Logistic loss?

We can use more general strategies. In particular...

Proposition. Let $F : \mathbb{R}^n \to \mathbb{R}$ be differentiable and M-Lipschitz. Then $M = \sup_w \|\nabla F(w)\|$.

The theorem can be generalized also to vector valued functions (such as the gradient!)

Estimating the Lipschitz Constant (II)

Proof. For any $\theta > 0$ and $v \in \mathbb{R}^N$. Since F is Lipschitz

$$\frac{|F(w + \theta v) - F(w)|}{\theta} \leq M.$$

Taking the limit for $\theta \to 0$ we have $\nabla F(w)^\top v \leq M$ for any $v \in \mathbb{R}^N$, implying $\|\nabla F(w)\| \leq M$. Then $\sup_w \|\nabla F(w)\| \leq M$.

However, for any two vectors $w, z \in \mathbb{R}^n$, we have

$$\begin{aligned} F(w) - F(z) &= \int_0^1 \nabla F(w + \theta(z - w))^\top (z - w)\, d\theta \\ &\leq \left(\int_0^1 \|\nabla F(w + \theta(z - w))\|\, d\theta \right) \|z - w\| \\ &\leq \left(\sup_v \|\nabla F(v)\| \right) \|z - w\| \end{aligned}$$

Hence $\sup_v \|\nabla F(v)\| \geq M$, which concludes the proof. ■

Example: Least Squares (Again!)

Recall that the gradient of of $F(w) = \nabla \mathcal{E}_S(w) = 2(C_S + \lambda I)w$.

Its gradient corresponds to the Hessian of $\mathcal{E}_S(w)$, which is

$$\nabla F(w) = \nabla^2 \mathcal{E}_S = 2(C_S + \lambda I).$$

Taking its (operator) norm over all possible w is

$$\sup_w \|\nabla F(w)\| = 2\,\|C_S + \lambda I\|$$

as we already observed via the more direct derivation.

Sub-gradient Method

Convexity without Differentiability

So... if our Empirical risk has a differentiable (and convex) loss, we can adopt GD to find a minimizer (or at least approximate it).

What about the Hinge loss however?

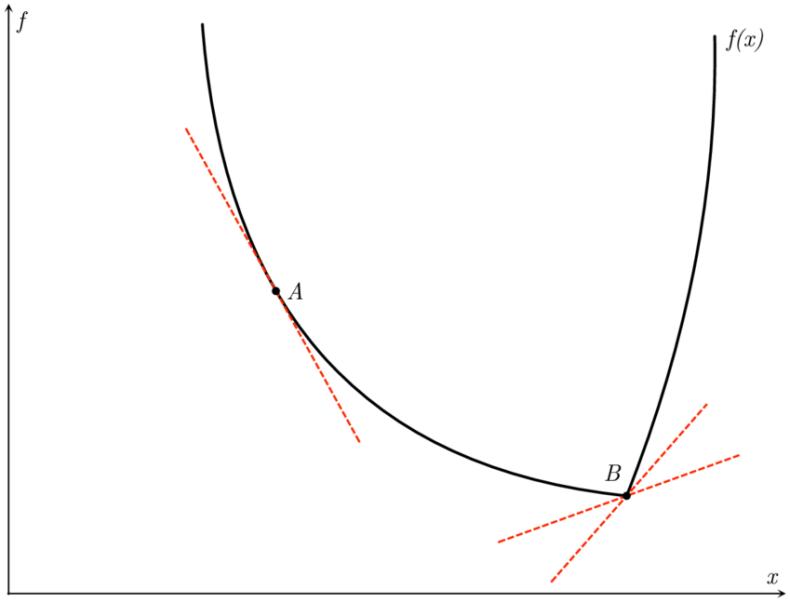
$$F(w) = \max(0, 1 - yw^\top x)$$

is non-differentiable in the points $\bar{w} \in \mathbb{R}^N$ such that $y\bar{w}^\top x = 1$.

We cannot use gradient descent!
(we don't have a gradient there)

Sub-gradients

Looking at a (convex) non-differentiable function in the point of non-differentiability, the problem does not seem like that the first order condition does not hold...



Actually, is holds for many linear lower bounds (rather than a single one!).

Sub-gradients (II)

We might be able to generalize the notion of gradient for a convex function from the first-order condition...

$$F(z) \geq F(w) + \nabla F(w)^\top (z - w)$$

...that of a "sub-gradient":

Definition. We say that $g \in \mathbb{R}^N$ is a sub-gradient of F in w if

$$F(z) \geq F(w) + \nabla F(w)^\top (z - w)$$

for any $z \in \mathbb{R}^N$. The set of all sub-gradients of F in w is the *sub-differential* and is denoted $\partial F(w)$

Sub-gradient Method

Assuming to have a way to compute them (more in the following)...

... Are these sub-gradients useful?

Sub-gradient ~~Descent~~ Method. Starting from a $w_0 \in \mathbb{R}^N$, we consider the sequence $(w_k)_{k=1}^K$ such that

$$w_{k+1} = w_k - \eta_k g_k$$

where $g_k \in \partial F(w_k)$ is a (any!) sub-gradient for F in w_k.

Note that $\eta_k > 0$ here depends on the iteration k.

Question. Does this strategy work?

“Convergence” of the Sub-gradient Method

Theorem. Let $F : \mathbb{R}^N \to \mathbb{R}$ be L-Lipschitz[3]. Let $(w_k)_{k=1}^K$ be a sequence obtained via sub-gradient method. Then

$$F(w_k^{best}) - F(w_*) \leq \frac{\|w_0 - w_*\|^2 + L^2 \sum_{k=1}^K \eta_k^2}{2 \sum_{k=1}^K \eta_k}$$

where

$$w_k^{best} = \underset{w \in \{w_1, \ldots, w_K\}}{\arg\min} F(w).$$

[3] using L instead of M here to highlight that we are requiring F to be Lipschitz and not smooth

Implications for the Sub-gradient Method

Convergence is trickier than for GD: let $\eta_k = \eta$ be a constant step size. Then

$$F(w_K^{best}) - F(w_*) \leq \frac{\|w_0 - w_*\|^2}{2\eta K} + \frac{L^2\eta}{2}$$

If we choose $\eta = \frac{1}{\sqrt{K}}$ we have

$$F(w_K^{best}) - F(w_*) \leq \frac{1}{2\sqrt{K}}(\|w_0 - w_*\| + L^2)$$

...which is much slower than GD!

(Also, we need to choose K a-priori and cannot "do a few more steps" as needed, differently to what we would do with GD).

“Convergence” of the Sub-gradient Method (II)

Proof. By definition of sub-gradient

$$F(w_*) - F(w_k) \geq g_k^\top (w_* - w_k)$$

Therefore

$$\begin{aligned}
\|w_{k+1} - w_*\|^2 &= \|(w_k - w_*) - \eta_k g_k\|^2 \\
&= \|w_k - w_*\|^2 - 2\eta_k g_k^\top (w_k - w_*) + \eta_k^2 \|g_k\|^2 \\
&\leq \|w_k - w_*\|^2 - 2\eta_k (F(w_k) - F(w_*)) + \eta_k^2 \|g_k\|^2
\end{aligned}$$

By applying the inequality above across all iterates

$$\|w_K - w_*\|^2 \leq \|w_0 - w_*\|^2 - 2\sum_{k=1}^{K} \eta_k (F(w_k) - F(w_*)) + \sum_{k=1}^{K} \eta_k^2 \|g_k\|^2$$

“Convergence” of the Sub-gradient Method (Conclusion)

Since $\|w_K - w_*\| \geq 0$, we have

$$2\sum_{k=1}^{K} \eta_k (F(w_k) - F(w_*)) \leq \|w_0 - w_*\|^2 + \sum_{k=1}^{K} \eta_k^2 \|g_k\|^2$$

Additionally, since $F(w_K^{best}) \leq F(w_k)$ for any $k = 1, \ldots, K$,

$$2(\sum_{k=1}^{K} \eta_k) \left(F(w_K^{best}) - F(w_*)\right) \leq \|w_0 - w_*\|^2 + \sum_{k=1}^{K} \eta_k^2 \|g_k\|^2$$

Observing[4] that $\|g_k\| \leq L$ the result follows. ■

[4] a similar reasoning to the smooth case applies. Exercise!

Sub-gradients for a single variable

Ok, so sub-gradients are useful.

How do we compute them in practice?

Let's start from a *single* variable: it is easy to show that for $F : \mathbb{R} \to \mathbb{R}$ convex, the sub-differential is

$$\partial F(w) = [a, b]$$

with

- $a = \lim_{\theta \to 0^-} \frac{F(w+\theta) - F(w)}{\theta}$
- $b = \lim_{\theta \to 0^+} \frac{F(w+\theta) - F(w)}{\theta}$

Example: Hinge Loss

Consider $F(w) = \ell(w, y) = \max(0, 1 - yw)$. Then

$$\partial F(w) = \begin{cases} -y & \text{if } w < 1/y \\ [-y, 0] & \text{if } w = 1/y \\ 0 & \text{if } w > 1/y \end{cases}$$

What about extending to more than one variable?

Non-smooth loss + Linear model

Denote $F(w) = \ell(w^\top x, y)$ for $w, x \in \mathbb{R}^N$ and $y \in \mathbb{R}$. Then, for any $z \in \mathbb{R}^N$

$$F(z) - F(w) = \ell(z^\top x, y) - \ell(w^\top x, y)$$

By definition of sub-gradient of $\ell(\cdot, y)$,

$$\ell(z^\top x, y) - \ell(w^\top x, y) \geq g(z^\top x - w^\top x)$$

for any subgradient $g \in \partial\ell(w^\top x, y)$ of $\ell(\cdot, y)$ in $w^\top x$. Hence

$$F(z) - F(w) \geq \underbrace{gx^\top}_{\text{subgradient for } F!} (w - z)$$

Implying that $\partial F(w) = \partial\ell(w^\top x, y)x = \{gx \mid g \in \partial\ell(w^\top x, y)\}$

Hinge Loss and Linear Models

Let $F(w) = \max(0, yw^\top x)$ for $w, x \in \mathbb{R}^N$

Then

$$\partial F(w) = \begin{cases} -yx & \text{if } w^\top x < 1/y \\ -rx \text{ with } r \in [0, y] & \text{if } w^\top x = 1/y \\ 0 & \text{if } w^\top x > 1/y \end{cases}$$

Summary

- We observed that several strategies exist to tackle binary classification. Many of them boil down to variants of ERM with a different loss.
- We asked the question of how to solve ERM in practice.
- Focusing on convex + smooth loss functions we studied GD as a practical strategy to do so.
- For non-smooth (yet convex) loss functions – such as the hinge loss for SVM – we have studied the sub-gradient method.

Now, go! And try them in practice!