Pisa part 2

The Fenchel-Rockafellar duality

The Fenchel conjugate

X Hilbert space

$f : X \to]-\infty, +\infty]$ proper, convex l.s.c.

$f^* : X \to]-\infty, +\infty]$ $\quad f^*(u) = \sup_x \langle u, x \rangle - f(x)$

f^* is proper, convex, l.s.c. and $f^{**} = f$ (Fenchel-Moreau theorem).

Few facts.

1. $f(x) + f^*(u) \geq \langle x, u \rangle$ (Young inequality)
2. $f(x) + f^*(u) = \langle x, u \rangle \iff u \in \partial f(x)$
3. $u \in \partial f(x) \iff x \in \partial f^*(u)$.

$\Gamma_0(X) = \{ f : X \to]-\infty, +\infty] \mid f$ is proper, convex and l.s.c. $\}$

Let $f \in \Gamma_0(X)$, $g \in \Gamma_0(Y)$ and $A : X \to Y$ bounded linear operator. Consider the minimization problem:

$$(P) \qquad \min_{x \in X} f(x) + g(Ax) =: \Phi(x)$$

The dual problem (in the sense of Fenchel-Rockefeller) is

$$(D) \qquad \min_{u \in Y} g^*(u) + f^*(-A^*u) =: \Psi(u)$$

It is of the same form as (P), $A^* : Y \to X$ is the transpose operator of A.

The dual problem can be obtained "formally" via Lagrange duality.

$$(P) \iff \inf_{\substack{(x,y) \\ Ax = y}} f(x) + g(y) \qquad \text{(constrained problem)}$$

The Lagrangian is $L(x, y, u) = f(x) + g(y) + \langle Ax - y, u \rangle$

We have $\sup_u L(x, y, u) = \begin{cases} f(x) + g(y) & \text{if } Ax = y \\ +\infty & \text{if } Ax \neq y. \end{cases}$

Therfore (P) $\Leftrightarrow \inf_{(x,y)} \sup_{u} L(x,y,u)$

Now, if we swap inf and sup we obtain

$$\sup_{u} \inf_{(x,y)} L(x,y,u)$$

$$= \sup_{u} \inf_{x,y} f(x) - \langle x, -A^*u \rangle + g(y) - \langle y, u \rangle$$

$$= \sup_{u} -f^*(-A^*u) - g^*(u)$$

$$= \sup_{u} -\Psi(u) = -\inf_{u} \Psi(u)$$

So, in the end we obtain the dual problem (D).

What is the relation between the primal and the dual problem?

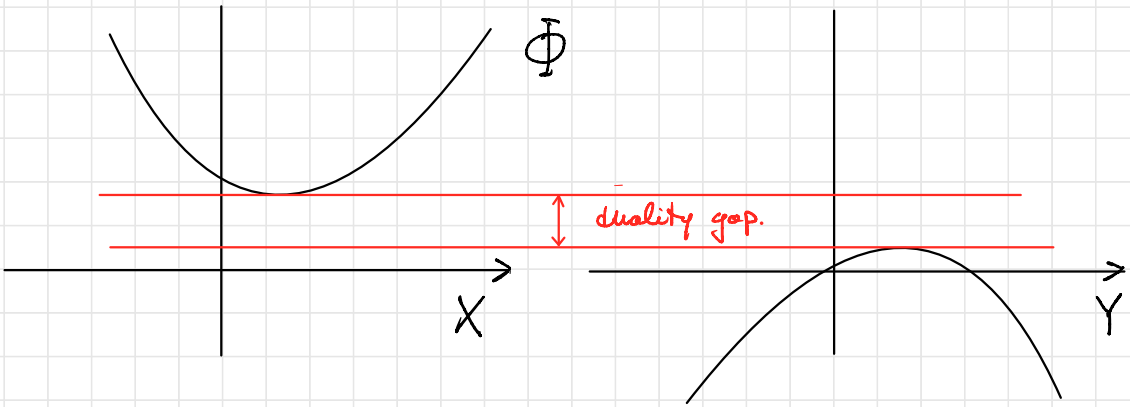
Prop. $\forall x \in X \; \forall u \in Y \quad \Phi(x) \geq -\Psi(u)$

Proof: $\Phi(x) + \Psi(u) = f(x) + f^*(-A^*u)$

$$+ g(Ax) + g^*(u)$$

$$\geq \langle x, -A^* u \rangle + \langle Ax, u \rangle = 0$$

Therefore the situation is as follows.



The images of Φ & Ψ are separated and hence

$$\inf \Phi \geq \sup -\Psi = -\inf \Psi$$

When $\inf \Phi = \sup -\Psi$ (duality gap is zero)
we say that strong duality holds

Strong duality is important because in that case the optimal values of the primal and dual problems are the same and hence solving the dual problem becomes an alternative to solving the primal problem.

In order to ensure strong duality, qualification conditions are needed

The primal objective function is

$$\Phi(x) = f(x) + g(Ax)$$

The minimum requirement for Φ is that it is a proper function.

$$\begin{aligned} \operatorname{dom}\Phi \neq \emptyset &\iff \exists\, x \in X \text{ s.t. } f(x) < +\infty \text{ and } g(Ax) < +\infty \\ &\iff \exists\, x \in \operatorname{dom} f \text{ i.e. } \underbrace{Ax}_{u} \in \operatorname{dom} g \\ &\iff 0 \in \operatorname{dom} g - A(\operatorname{dom} f) \end{aligned}$$

Qualification conditions ask more. One of those is

$$\text{QC:} \quad 0 \in \operatorname{int}(\operatorname{dom} g - A(\operatorname{dom} f))$$

This is satisfied for instance if

g is continuous at some Ax with $x \in \operatorname{dom} f$.

Thm. If QC holds, then strong duality holds and the dual problem admits solutions

(this is usually written as $\inf \Phi = -\min_{u} \Psi$)

Remark:

Since the dual of the dual problem is the primal problem, one can ensure strong duality also by requiring $0 \in \operatorname{int}(\operatorname{dom} f^* + A^*(\operatorname{dom} g^*))$.

Let's address the characterization of primal and dual solutions.

Theorem Let $S = \operatorname{argmin} \Phi$, $S^* = \operatorname{argmin} \Psi$ and $\bar{x} \in X$ and $\bar{u} \in Y$. Then the following are equivalent

a) $\bar{x} \in S$, $\bar{u} \in S^*$ and $\inf \Phi = -\inf \Psi$.

b) $\Phi(\bar{x}) + \Psi(\bar{u}) = 0$.

c) $-A^*\bar{u} \in \partial f(\bar{x})$ and $\bar{u} \in \partial g(A\bar{x})$

d) $\bar{x} \in \partial f^*(-A^*\bar{u})$ and $A\bar{x} \in \partial g^*(\bar{u})$

} KKT conditions.

Proof:

$$\Phi(\bar{x}) + \Psi(\bar{u}) = \underbrace{f(\bar{x}) + f^*(-A^*\bar{u}) - \langle \bar{x}, -A^*\bar{u} \rangle}_{\geq 0}$$

$$+ \underbrace{g(A\bar{x}) + g^*(\bar{u}) - \langle A\bar{x}, \bar{u} \rangle}_{\geq 0}$$

Thus, b) $\iff f(\bar{x}) + f^*(-A^*\bar{u}) = \langle \bar{x}, -A^*\bar{u} \rangle$

and $g(A\bar{x}) + g^*(\bar{u}) = \langle A\bar{x}, \bar{u} \rangle$

$\iff -A^*\bar{u} \in \partial f(\bar{x})$ and $\bar{u} \in \partial g(A\bar{x})$

Remark: As a consequence of the previous theorem, we have that once strong duality is ensured, KKT conditions characterize primal and dual solutions simultaneously

---

Now we address the computational aspects of duality. Ideally one wants to go from a dual solution to a primal solution, so that the primal problem can be solved via the dual problem.

In order to do so, we need an additional assumption.

Assumption f is μ-strongly convex

As a consequence of this assumption we have

1) the primal problem has a unique solution

2) $\operatorname{dom} f^* = X$ and f^* is $\frac{1}{\mu}$-Lipschitz smooth.

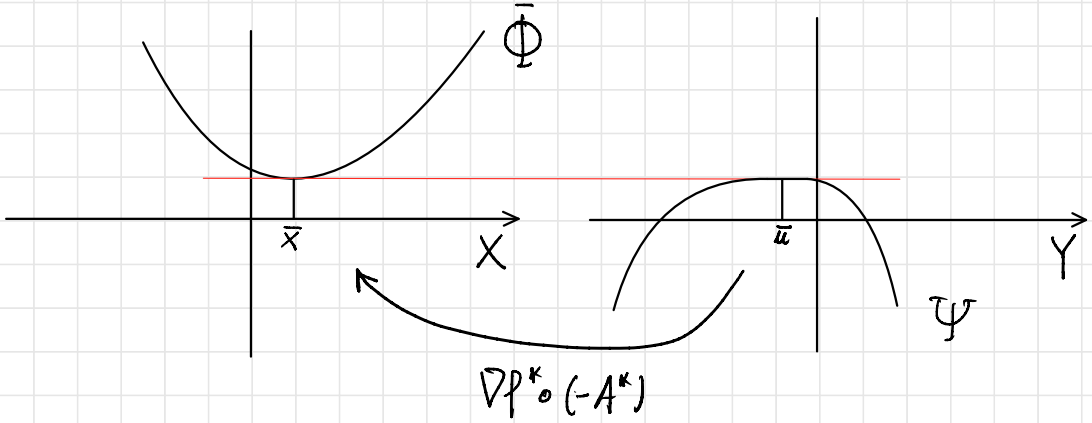
Then the KKT conditions become

$$\bar{x} = \nabla f^*(-A^*\bar{u}) \quad \text{and} \quad A\bar{x} \in \partial g^*(\bar{u})$$

The first of the KKT determine uniquely the primal solution from a dual solution. So, the map.

$$u \mapsto \nabla f^*(-A^*u)$$

provides a way to go from the dual variable to the primal variable.



Proposition

Let $\bar{x} = \nabla f^*(-A^*\bar{u})$ and $x = \nabla f^*(-A^*u)$.

Then $$\frac{\mu}{2}\|x - \bar{x}\|^2 \leq \Psi(u) - \Psi(\bar{u})$$

Proof.

$$f(\bar{x}) + f^*(-A^*\bar{u}) = \langle \bar{x}, -A^*\bar{u} \rangle$$

$$f(x) + f^*(-A^*u) = \langle x, -A^*u \rangle$$

$$\Psi(u) - \Psi(\bar{u}) = g^*(u) - g^*(\bar{u}) + f^*(-A^*u) - f^*(-A^*\bar{u})$$

$$= g^*(u) - g^*(\bar{u}) + \langle x, -A^*u \rangle - f(x) + f(\bar{x}) + \langle \bar{x}, A^*\bar{u} \rangle$$

$$\geq \underbrace{\langle u - \bar{u}, A\bar{x} \rangle}_{A\bar{x} \in \partial g^*(\bar{u})} - \langle Ax, u \rangle + \langle A\bar{x}, \bar{u} \rangle + f(\bar{x}) - f(x)$$

$$= \langle x - \bar{x}, -A^*u \rangle + f(\bar{x}) - f(x)$$

since $-A^*u \in \partial f(x)$ and f is μ-strongly convex

$$f(\bar{x}) - f(x) - \langle \bar{x} - x, -A^*u \rangle \geq \frac{\mu}{2} \|x - \bar{x}\|^2. \qquad \square$$

Remark The previous result shows that if $\Psi(u_n) \to \min \Psi$ then $x_n = \nabla f^*(-A^*u_n) \to \bar{x}$.

One can solve the primal problem via the dual problem

A dual approach to regularized empirical risk minimization

We first introduce the hypothesis space

Let H be a Hilbert space and $\Lambda : \mathcal{X} \to H$ be a map (called feature map).

Then this map serves to define real functions defined in $\mathcal{X}$ and parametrized in H.

$$f : \mathcal{X} \to \mathbb{R} \qquad f(x) = \langle w, \Lambda(x) \rangle \quad , \quad w \in H.$$

The set of all these functions define a Hilbert space of functions

$$\mathcal{H} = \{ f : \mathcal{X} \to \mathbb{R} \mid f = \langle w, \Lambda(\cdot) \rangle \quad w \in H \}$$

$$\| f \|_{\mathcal{H}} = \inf \{ \|w\| \mid w \in H \text{ and } f = \langle w, \Lambda(\cdot) \rangle \}$$

If $\overline{\text{span}\{ \Lambda(x) \mid x \in \mathcal{X} \}} = H$, then there is

a unique $w \in H$ s.t. $f = \langle w, \Lambda(\cdot) \rangle$ and $\|f\|_{\mathcal{H}} = \|w\|$

One can prove that $\mathcal{H}$ is a RKHS with kernel

$$K(x, x') = \langle \Lambda(x), \Lambda(x') \rangle.$$

Given

$(x_1, y_1) \dots (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$ training points,

$\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ loss function (convex wrt the first v.),

the regularized empirical risk minimization problem is

$$\min_{w \in H} \frac{1}{n\lambda} \sum_{i=1}^{n} \ell(\langle w, \Lambda(x_i) \rangle, y_i) + \frac{1}{2}\|w\|^2 \qquad \text{R-ERM}$$

We will put the problem in the framework of the Fenchel-Rockafellar duality.

1. $A : H \to \mathbb{R}^n \qquad Aw = \begin{bmatrix} \langle w, \Lambda(x_1) \rangle \\ \vdots \\ \langle w, \Lambda(x_n) \rangle \end{bmatrix}$

2. $f(w) = \frac{1}{2}\|w\|^2$ (which is 1-strongly convex)

3. $g : \mathbb{R}^n \to \mathbb{R}$ $\quad g(z) = \frac{1}{n\lambda} \sum_{i=1}^{n} \ell(-z_i, y_i)$

Then problem R-ERM is written as

$$\min_{w \in H} g(-Aw) + f(w).$$

Thus, the dual problem is

$$\min_{\alpha \in \mathbb{R}^n} g^*(\alpha) + f^*(A^*\alpha).$$

Now we note that

1. $A^*\alpha = \sum_{i=1}^{n} \alpha_i \Lambda(x_i)$
2. $f^*(v) = \frac{1}{2}\|v\|^2$

Therefore $f^*(A^*\alpha) = \frac{1}{2}\left\| \sum_{i=1}^{n} \alpha_i \Lambda(x_i) \right\|^2$

$$= \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j \langle \Lambda(x_i), \Lambda(x_j) \rangle = \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j)$$

The Gram matrix $K_{ij} = K(x_i, x_j)$,

$$f^*(A^*\alpha) = \frac{1}{2} \langle K\alpha, \alpha \rangle = \frac{1}{2} \alpha^T K \alpha$$

3. g can be written as

$$g(z) = \sum_{i=1}^{n} g_i(z_i)$$

where $g_i(t) = \frac{1}{n\lambda}\ell(-t, y_i)$ $\quad g_i : \mathbb{R} \to \mathbb{R}$.

Therefore $g^*(a) = \sum_{i=1}^{n} g_i^*(a_i)$

Then we need to compute g_i^*

By definition

$$g_i^*(s) = \sup_t st - g_i(t) = \sup_t st - \frac{1}{n\lambda}\ell(-t, y_i)$$

$$= \frac{1}{n\lambda} \sup_t (-sn\lambda)(-t) - \ell(-t, y_i)$$

$$= \frac{1}{n\lambda}\ell^*(-sn\lambda, y_i)$$

Putting everything together we obtain

$$\min_{a \in \mathbb{R}^n} \frac{1}{2}\langle Ka, a\rangle + \frac{1}{n\lambda}\sum_{i=1}^{n} \ell^*(-a_i n\lambda, y_i)$$

The KKT conditions in general are

$$\bar{w} = \underbrace{\nabla f^*(A^*\bar{a})}_{I} \quad \text{and} \quad -A\bar{w} \in \partial g^*(\bar{a})$$

and in this case they become

$$\bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i \Lambda(x_i) \qquad (1)$$

$$-\langle \bar{w}, \Lambda(x_i) \rangle \in \partial \varphi_i^*(\bar{\alpha}_i)$$

$$\bar{\alpha}_i \in \partial \varphi_i(-\langle \bar{w}, \Lambda(x_i) \rangle)$$

$$-\bar{\alpha}_i \in \frac{1}{n\lambda} \partial \ell(\langle \bar{w}, \Lambda(x_i), y_i \rangle) \qquad (2)$$

We note that the first of the KKT conditions is nothing more than the representer theorem

The corresponding function is

$$f_{\bar{w}}(x) = \langle \bar{w}, \Lambda(x) \rangle = \sum_{i=1}^{n} \bar{\alpha}_i \langle \Lambda(x_i), \Lambda(x) \rangle$$

$$= \sum_{i=1}^{n} \bar{\alpha}_i K(x_i, x)$$

which shows that the estimator can be expressed in terms of the Kernel function and the dual solution. This aspect is what makes this approach attractive and is known as the Kernel trick

This framework defines the so called Kernel methods which has the following nice features:

1. The dual problem is a finite dimensional convex problem
2. The dual objective function is expressed in terms of the Kernel functions, the data and the loss function. Therefore it is completely known.
3. The final estimator (function) is expressed in terms of the Kernel function, the data and the dual solution.

Now, concerning the algorithmic approach to R-ERM, We note that the mapping linking the dual variable to the primal variable is

$$\alpha \mapsto \sum_{i=1}^{n} \alpha_i \Lambda(x_i)$$

and for every α if we define $w = \sum_{i=1}^{n} \alpha_i \Lambda(x_i)$

we have

$$\frac{1}{2}\|w-\bar{w}\|^2 \leq \Psi(\alpha) - \min \Psi$$

This inequality is the basis of approaching the R-ERM problem via dual algorithms

Indeed if $f = \langle w, \Lambda(\cdot)\rangle$ and $\bar{f} = \langle \bar{w}, \Lambda(\cdot)\rangle$, we have

$$|f(x) - \bar{f}(x)| = |\langle w-\bar{w}, \Lambda(x)\rangle| \leq \|w-\bar{w}\| \, \|\Lambda(x)\|$$
$$\leq \sqrt{2(\Psi(\alpha) - \min\Psi)} \sqrt{K(x,x)}.$$

If $(\alpha^{(k)})_{k\in\mathbb{N}}$ is such that $\Psi(\alpha^{(k)}) \to \inf \Psi$ and if we define $w_k = \sum_{i=1}^{n} \alpha^{(k)} \Lambda(x_i)$ and the corresponding function $f_k = \langle w_k, \Lambda(\cdot)\rangle \in \mathcal{H}$, we have that at a new point $x \in \mathcal{X}$

$$|f_k(x) - \bar{f}(x)| \leq \sqrt{2\underbrace{(\Psi(\alpha^{(k)}) - \min\Psi)}_{\downarrow 0}} \sqrt{K(x,x)} \to 0$$

In case $\sup_x K(x,x) < +\infty$, then $\|f_k - \bar{f}\|_\infty \to 0$.

Example

1. The square loss $\ell(t,y) = \frac{1}{2}(t-y)^2$

$$\ell^*(s,y) = \frac{1}{2}s^2 + y\cdot s. \quad \text{Hence}$$

$$\frac{1}{n\lambda}\sum_{i=1}^{n}\ell^*(-\alpha_i n\lambda, y_i) = \frac{1}{n\lambda}\sum_{i=1}^{n}\left(\frac{1}{2}(\alpha_i n\lambda)^2 - y_i\,\alpha_i n\lambda\right)$$

$$= \frac{n\lambda}{2}\|\alpha\|^2 - \langle y,\alpha\rangle$$

The dual problem is

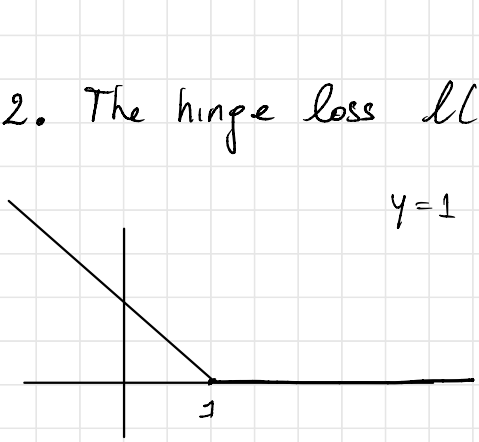
$$\min_{\alpha\in\mathbb{R}^n} \frac{1}{2}\alpha^T K\alpha + \frac{n\lambda}{2}\|\alpha\|^2 - \langle y,\alpha\rangle$$

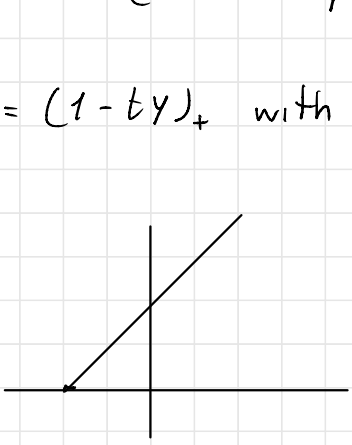
which has solution

$$K\alpha + n\lambda\alpha - y = 0 \iff (K + n\lambda I)\alpha = y$$

$$\iff \alpha = (K + n\lambda)^{-1} y$$

2. The hinge loss $\ell(t,y) = (1-ty)_+$ with $y\in Y=\{-1,1\}$.





$$\ell(t,y) = \chi(ty) \quad \text{where} \quad \chi(r) = (1-r)_+$$

$$\ell^*(s,y) = \sup_t st - \chi(ty) = \sup_t (sy)(ty) - \chi(ty)$$

$$= \chi^*(sy)$$

Moreover the function χ is the function $(-r)_+$ translated by -1, therefore $\chi(r) = [(-s)_+]_{s=r-1}$

$$\chi^*(q) = q + i_{[-1,0]}(q) \qquad [(-s)_+]^*(q) = i_{[-1,0]}(q).$$

$$\ell^*(s,y) = sy + i_{[-1,0]}(sy)$$

$$\frac{1}{n\lambda} \sum_{i=1}^{n} \ell^*(-2in\lambda, y_i) = \frac{1}{n\lambda} \sum_{i=1}^{n} -2in\lambda y_i + i_{[-1,0]}(-2in\lambda y_i)$$

$$0 \le 2in\lambda y_i \le 1 \quad \Leftrightarrow \quad 0 \le 2vy_i \le \frac{1}{n\lambda}$$

$$\boxed{\min_{2 \in \mathbb{R}^n} \frac{1}{2} \langle K2, 2 \rangle - 2^T y + i_{[0, \frac{1}{n\lambda}]^n}(2 \odot y)}$$

This is a constrained quadratic minimization problem.