CR12: Statistical Learning & Applications

Kernel methods and Stochastical Learning

Lecturer: Joseph Salmon *Scribes: Antoine Plet and Bertrand Simon*

1 Stochastic learning (second part)

1.1 Strong convexity

Part of the elements in this part can be found in [HUL93a, HUL93b]

Definition (Strong convexity)**.** *A function f is said λ-strongly convex if the function $\left(f - \frac{\lambda}{2}\|\cdot\|^2\right)$ is convex.*

Remark. *The norm used in this document is always the Euclidean norm: $\|\cdot\| = \|\cdot\|_2$.*

Proposition. *f is λ-strongly convex if and only if $\forall w_1 w_2, \forall \alpha \in [0,1]$:*

$$f\left(\alpha w_1 + (1-\alpha)w_2\right) \leq \alpha f(w_1) + (1-\alpha)f(w_2) - \frac{\lambda}{2}\alpha(1-\alpha)\|w_1 - w_2\|^2 \tag{1}$$

We will then use w_α denoting $\alpha w_1 + (1-\alpha)w_2$.

Proof. Firstly, we remark that for all $\alpha \in [0,1]$,

$$\|w_\alpha\|^2 = \alpha^2\|w_1\|^2 + (1-\alpha)^2\|w_2\|^2 + 2\alpha(1-\alpha)\langle w_1, w_2\rangle$$

Then, we prove the equivalence of the two statements. f is strongly convex if and only if the following inequality is valid for all $\alpha \in [0,1]$.

$$f(w_\alpha) - \frac{\lambda}{2}\|w_\alpha\|^2 \leq \alpha f(w_1) + (1-\alpha)f(w_2) - \frac{\lambda}{2}\left(\alpha\|w_1\|^2 + (1-\alpha)\|w_2\|^2\right)$$

Then, by replacing $\|w_\alpha\|$ by its expression, we obtain:

$$f(w_\alpha) \leq \alpha f(w_1) + (1-\alpha)f(w_2) - \frac{\lambda}{2}\left(\left(\alpha - \alpha^2\right)\|w_1\|^2 + \left((1-\alpha) - (1-\alpha)^2\right)\|w_2\|^2 + 2\alpha(1-\alpha)\langle w_1, w_2\rangle\right)$$

By reducing and factorizing, we get the following inequalities

$$f(w_\alpha) \leq \alpha f(w_1) + (1-\alpha)f(w_2) - \frac{\lambda}{2}\alpha(1-\alpha)\left(\|w_1\|^2 + \|w_2\|^2 - 2\langle w_1, w_2\rangle\right)$$

$$f(w_\alpha) \leq \alpha f(w_1) + (1-\alpha)f(w_2) - \frac{\lambda}{2}\alpha(1-\alpha)\|w_1 - w_2\|^2$$

This last inequality is (1), and valid for all $\alpha \in [0,1]$, so the two statements are equivalent. □

Theorem. *f is λ-strongly convex if and only if $\forall w_1 w_2, \forall v \in \partial f(w_2)$*

$$f(w_1) \geq f(w_2) + \langle w_1 - w_2, v\rangle + \frac{\lambda}{2}\|w_1 - w_2\|^2 \tag{2}$$

Proof. Firstly, we remark that for all $\alpha \in [0,1]$,

$$w_1 - w_\alpha = (1-\alpha)(w_1 - w_2) \qquad w_2 - w_\alpha = \alpha(w_2 - w_1) \tag{3}$$

We will show both ways of the equivalence

- Suppose that Condition 2 is valid, and let α be in $]0,1]$ and v be in $\partial f(w_\alpha)$. We have:

$$f(w_1) \geq f(w_\alpha) + \langle w_1 - w_\alpha, v\rangle + \frac{\lambda}{2}\|w_1 - w_\alpha\|^2$$
$$f(w_2) \geq f(w_\alpha) + \langle w_2 - w_\alpha, v\rangle + \frac{\lambda}{2}\|w_2 - w_\alpha\|^2$$

  Then, combining these two inequalities and using the equalities (3), we obtain

$$\alpha f(w_1) + (1-\alpha) f(w_2) \geq f(w_\alpha) + \alpha(1-\alpha)\langle (w_1 - w_2) + (w_2 - w_1), v\rangle + \frac{\lambda}{2}\|w_1 - w_2\|^2 \left(\alpha(1-\alpha)^2 + (1-\alpha)\alpha^2\right)$$

  Simplifying, we finally get

$$\alpha f(w_1) + (1-\alpha) f(w_2) \geq f(w_\alpha) + \frac{\lambda}{2}\alpha(1-\alpha)\|w_1 - w_2\|^2$$

  This equality is (1) and is valid for all $\alpha \in [0,1]$, so f is λ-strongly convex.

- Suppose that f is λ-strongly convex and let α be in $[0,1[$. We have:

$$\alpha f(w_1) + (1-\alpha) f(w_2) \geq f(w_\alpha) + \frac{\lambda}{2}\alpha(1-\alpha)\|w_1 - w_2\|^2$$

  Then, we subtract $f(w_2)$ from both sides, and divide by α. We obtain:

$$f(w_1) - f(w_2) \geq \frac{f(w_\alpha) - f(w_2)}{\alpha} + \frac{\lambda}{2}(1-\alpha)\|w_1 - w_2\|^2$$

  We choose $v \in \partial f(w_2)$, so $f(w_\alpha) - f(w_2) \geq \langle w_\alpha - w_2, v\rangle$, and we have:

$$f(w_1) - f(w_2) \geq \left\langle \frac{w_\alpha - w_2}{\alpha}, v \right\rangle + \frac{\lambda}{2}(1-\alpha)\|w_1 - w_2\|^2$$

  Then, we use the equalities (3) and get

$$f(w_1) - f(w_2) \geq \langle w_1 - w_2, v\rangle + \frac{\lambda}{2}(1-\alpha)\|w_1 - w_2\|^2$$

  This inequality is valid for all $\alpha \in]0,1[$, so with α approaching 0, we obtain the inequality (2), so the theorem is valid.

□

1.2 Faster rate for stochastic (sub)-gradient descent algorithm using strong convexity

Now that we have introduced strong convexity we can analyze the stochastic (sub)-gradient descent algorithm as follows.

Algorithm 1 Stochastic Subgradient Descent

Input: f λ-strongly convex, $T > 0$, $\eta_t = \frac{1}{t}$
Output: w, probably around a minimum of f

```
w_0 ← 0
for t = 1 to T do
    choose v_t s.t. E(v_t) ∈ ∂f(w_{t-1})
    w'_t ← w_{t-1} − η_t v_t
    w_t ← Π_A(w'_t) // "slap the drunk bug" back to A if it goes too far away from its home town
end for
return w̄ = (1/T) Σ_{t=1}^{T} w_t
```

Theorem. *If f is λ-strongly convex, and we have ρ such that $\mathbb{E}\|v_t\|^2 \geq \rho$, then* $\forall w^* \in \underset{w \in A}{\arg\min}\, f(w)$

$$\mathbb{E}\left(f(\bar{w})\right) - f\left(w^*\right) \leq \frac{\rho^2}{2\lambda T}\left(1 + \log(T)\right) \tag{4}$$

$$\mathbb{E}\left\|\bar{w} - w^*\right\|^2 \leq \frac{\rho^2}{\lambda^2 T}\left(1 + \log(T)\right) \tag{5}$$

Proof. Convexity and Jensen's inequalities give :

$$f(w^*) \leq \frac{1}{T}\sum_{t=1}^{T} f(w_t)$$

$$\mathbb{E}\left[f(w^*)\right] \leq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[f(w_t)\right]$$

Assume as for the results without the strong convexity assumption, that for any t the following holds:

$$\mathbb{E}f(w_t) - f(w^*) \leq \mathbb{E}\left(\frac{\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2}{2\eta_t}\right) + \frac{\eta_t}{2}\rho^2 \tag{6}$$

The sums has telescoping terms leading to :

$$\sum_{t=1}^{T}\left\{\mathbb{E}f(w_t) - f(w^*)\right\} \leq \mathbb{E}\left\{\frac{\lambda}{2}\sum_{t=t}^{T}\|w_t - w^*\|^2\left(t - (t-1)\right)\right\} + \frac{1}{2\lambda}\sum_{t=1}^{T}\frac{1}{t}\rho^2 - \frac{\lambda}{2}\|w_t - w^*\|^2$$

Then :

$$\sum_{t=1}^{T}\left\{\mathbb{E}f(w_t) - f(w^*)\right\} \leq -\frac{\lambda T}{2}\mathbb{E}\|w_{T+1} - w^*\|^2 + \frac{\rho^2}{2\lambda}\underbrace{\sum_{t=1}^{T}\frac{1}{t}}_{\leq \int_0^T \frac{dt}{t} \leq 1 + \log T}$$

$$\leq \frac{\rho^2\log(eT)}{2\lambda}$$

Let's now prove inequality (6) :

$$\forall t, \mathbb{E}f(w_t) - f(w^*) \leq \mathbb{E}\left(\frac{\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2}{2\eta_t}\right) + \frac{\eta_t}{2}\rho^2$$

First, the following technical lemma enables to bound the effect of projection on A convex. The details are given below, but are the same as in the non strongly convex case.

$$\forall w, \forall x \in A, \|w - x\|^2 - \|\Pi_A(w) - x\|^2 \geq 0$$

Let us consider $\mathbb{E}\|w_t - w^*\|$. By definition, $\|w_0 - w^*\|^2 = \|w^*\|^2$. Then, for any $t \geq 0$. Recall that w'_t is the "value" of w_t before the projection, so that $w'_t = w_{t-1} - \eta_t v_t$, and :

$$\begin{aligned}
\|w_{t-1} - w^*\|^2 - \|w_t - w^*\|^2 &= \|w_{t-1} - w^*\|^2 - \|\Pi_A(w'_t) - w^*\|^2 \\
&= \left(\|w_{t-1} - w^*\|^2 - \|w'_t - w^*\|^2\right) - \Big(\underbrace{\|\Pi_A(w'_t) - w^*)\|^2 - \|w'_t - w^*\|^2}_{\leq 0 \text{ (projection lemma)}}\Big) \\
&\geq \|w_{t-1} - w^*\|^2 - \|(w_{t-1} - w^*) - \eta_t v_t\|^2 \\
&\geq 2\eta_t (w_{t-1} - w^*)^\intercal v_t - \eta_t^2 \|v_t\|^2
\end{aligned}$$

but by independance of the v_i's :

$$\mathbb{E}_{v_1,\ldots,v_t}\left(\|w_{t-1} - w^*\|^2 - \|w'_t - w^*\|^2\right) = \mathbb{E}_{v_1,\ldots,v_{t-1}}\left\{2\eta_t (w_{t-1} - w^*)^\intercal (\mathbb{E}_{v_t} v_t) - \eta_t^2 \mathbb{E}_{v_t}(\|v_t\|^2)\right\}$$

and as $\mathbb{E}_{v_t}(v_t)$ is a subgradient of f at w_{t-1},

$$(w_{t-1} - w^*)^\intercal (\mathbb{E}_{v_t} v_t) \geq f(w_{t-1}) - f(w^*)$$

and as we assumed that $\mathbb{E}\|v_t\|^2 \leq \rho$, we finally get :

$$\mathbb{E}\left(\|w_{t-1} - w^*\|^2 - \|w'_t - w^*\|^2\right) \geq 2\eta_t \mathbb{E}\left(f(w_t) - f(w^*)\right) - \eta_t^2 \rho^2$$

which is exactly the inequality (6).

□

Example. *We will see some examples of convex sets A and projections over A.*

- *For $A := \{x : \|x\|_2 \leq r\} = B_2(r)$, we have*

$$\Pi_A(x) = \begin{cases} r\dfrac{x}{\|x\|_2} & \text{if } \|x\|_2 > r \\ \\ x & \text{otherwise} \end{cases}$$

- $A := \left\{x : \|x\|_p \leq r\right\}$
- *If A is a simplex, $A := \left\{x : \sum_{i=1}^d x_i = 1, \forall i\ x_i \geq 0\right\}$*

1.3 Fenchel duality, conjugate functions

Definition. *Let f be a function $f : \mathbb{R}^d \to \mathbb{R}$. We define $f^*(\theta) := \sup_{w \in \mathbb{R}^d} \left(\langle w, \theta\rangle - f(w)\right)$*

Example.

- $\dfrac{\|\cdot\|^2}{2}$*:* $\left(\dfrac{\|\cdot\|^2}{2}\right)^*(\theta) = \sup_{w \in \mathbb{R}^d}\left(\langle w, \theta\rangle - \dfrac{\|w\|^2}{2}\right) = \dfrac{\|\theta\|^2}{2}$ *because the* sup *is reached when $w \propto \theta$ and $\|w\| = \|\theta\|$.*

- *If we have* $f : w \mapsto \dfrac{w^\intercal Q w}{2}$, *with* Q *invertible, then* $f^*(\theta) = \dfrac{\theta^\intercal Q^{-1}\theta}{2}$ *(exercise)*

Theorem. *If* f *is closed and convex, then* $(f^*)^* = f^{**} = f$

Proposition. *Assume* f *is closed and convex. Then we have*

$$f(u) + f^*(v) = \langle u, v\rangle \quad \textit{iff} \quad v \in \partial f(u) \quad \textit{iff} \quad u \in \partial f^*(v)$$

Proof. Firstly, we note that $\forall u, f(u) = f^{**}(u) = \sup_{w\in\mathbb{R}^d}(\langle w, u\rangle - f^*(w))$ so we have $\forall u \in \mathbb{R}^d, \forall w \in \mathbb{R}^d, f(u) \geq \langle w, u\rangle - f^*(w)$. Then, we consider u, v such that $v \in \partial f(u)$. The following are equivalent:

$$v \in \partial f(u) \tag{7}$$
$$\forall w \in \mathbb{R}^d, f(w) - f(u) \geq \langle w - u, v\rangle \tag{8}$$
$$\forall w \in \mathbb{R}^d, \langle v, u\rangle - f(u) \geq \langle v, w\rangle - f(w) \tag{9}$$
$$\langle v, u\rangle - f(u) \geq \sup_{w\in\mathbb{R}^d}(\langle v, w\rangle - f(w)) \tag{10}$$
$$\langle v, u\rangle - f(u) \geq f^*(v) \tag{11}$$
$$f(u) + f^*(v) = \langle u, v\rangle \tag{12}$$

Similarly, we can prove the equivalence between $f(u)+f^*(v) = \langle u, v\rangle$ and $u \in \partial f^*(v)$ by the duality between f and f^*. □

Proposition. *If* f *is strictly convex, then* f^* *is differentiable. In other words, if* f *has no affine part, then* $\forall\theta \in \mathbb{R}^d, \partial f^*(\theta)$ *is a singleton.*

Proof. Assume f strictly convex, and there exists w_1, w_2, θ such that $w_1, w_2 \in \partial f^*(\theta)$ and $w_1 \neq w_2$. By the previous proposition, we have $\theta \in \partial f(w_1) \cap \partial f(w_2)$. Then, for $i \in \{1, 2\}$, we have $f(w_i) + f^*(\theta) = \langle w_i, \theta\rangle$. So, $\forall\alpha \in [0, 1], \alpha f(w_1) + (1-\alpha)f(w_2) + f^*(\theta) = \langle w_\alpha, \theta\rangle$ By the definition of $f^*(\theta)$, we have $\forall\alpha \in [0, 1], \alpha f(w_1) + (1-\alpha)f(w_2) + f^*(\theta) \leq f^*(\theta) + f(w_\alpha)$. Now, as f is strictly convex, $\forall\alpha \in]0, 1[, \alpha f(w_1) + (1-\alpha)f(w_2) > f(w_\alpha)$, which is contradictory. □

Proposition. *If* f *is* λ*-strongly convex, then* $\forall v_1, v_2,\ \|\nabla f^*(v_1) - \nabla f^*(v_2)\| \leq \frac{1}{\lambda}\|v_1 - v_2\|$

Proof. Firstly, f is λ-strongly convex so f is strictly convex then $\nabla f^* exists$ Consider w_1, w_2, v_1, v_2, such that $v_1 \in \partial f(w_1), v_2 \in \partial f(w_2)$. We have

$$f(w_2) \geq f(w_1) + \langle v_1, w_2 - w_1\rangle + \frac{\lambda}{2}\|w_1 - w_2\|^2$$
$$f(w_1) \geq f(w_2) + \langle v_2, w_1 - w_2\rangle + \frac{\lambda}{2}\|w_1 - w_2\|^2$$

So, adding these inequalities, we obtain:

$$\langle v_1 - v_2, w_1 - w_2\rangle \geq \lambda\|w_1 - w_2\|^2$$

Then, by the Cauchy–Schwarz inequality,

$$\|w_1 - w_2\| \leq \frac{1}{\lambda}\|v_1 - v_2\|$$

Finally, as $v_1 \in \partial f(w_1)$, we have $w_1 \in \partial f^*(v_1) = \{\nabla f^*(v_1)\}$, and idem for w_2, so

$$\|\nabla f^*(v_1) - \nabla f^*(v_2)\| \leq \frac{1}{\lambda}\|v_1 - v_2\|$$

□

Theorem. *If f is λ-strongly convex then* $\forall w, \delta,\ f^*(w+\delta) \leq f^*(w) + \langle \delta, \partial f^*(w) \rangle + \frac{1}{2\lambda} \|\delta\|^2$

Proof. For any w, δ, we have:

$$\begin{aligned} f^*(w+\delta) - f^*(w) &= \int_0^1 \langle \nabla f^*(w+t\delta), \delta \rangle \, dt \\ &= \langle \nabla f^*(w), \delta \rangle + \int_0^1 \langle \nabla f^*(w+t\delta) - \nabla f^*(w), \delta \rangle \, dt \\ &\leq \langle \nabla f^*(w), \delta \rangle + \frac{\|\delta\|^2}{2\lambda} . \end{aligned}$$

□

2 Bounding errors

Theorem (Bartlett and Mendelson, 2002)**.** *Let $f(\cdot, y)$ be a Lipschitz function (i.e., the loss function L_f is bounded by $|L_f| \leq c$), then, for all $\delta \in [0,1]$, with probability greater that $1-\delta$, we have:*

$$\forall f \in \mathcal{F},\ \mathcal{L}(f) \leq \widehat{\mathcal{L}}(f) + 2L_f R_n(\mathcal{F}) + c\sqrt{\frac{2\log(2/\delta)}{n}}$$

with

$$\begin{aligned} \mathcal{L}(f) &= \mathbb{E}_{X,Y}\left(l[f(x_i), y_i]\right) \\ \widehat{\mathcal{L}}(f) &= \frac{1}{n}\sum_{i=1}^n l\left(f(x_i), y_i\right) \\ R_n(\mathcal{F}) &= \mathbb{E}\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\right) \end{aligned}$$

where R_n is the Rademacher complexity and $(\varepsilon_i)_{i=1,\cdots,n}$ are Rademacher variables, (i.e., they are i.i.d. over $\{-1,+1\}$ with equiprobability).

This results is a variant of Theorem 7 in [BM02].

Proof. For the proof of the previous result, one needs to use the McDiarmid Inequality stated in Theorem 3 with $c_i = c/n$ for some $t > 0$ if we set $\delta = \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right)$, then we have $t = c\sqrt{\frac{\log(1/\delta)}{2n}}$ Then, we have $\sup_h \mathbb{E}_h - \widehat{\mathbb{E}}_n(h)$ satisfying McDiarmid's inequality with $c_i = c/n$ □

2.1 Complexities for linear functions

Theorem. *Given λ, X and W_*. Suppose that F is a λ-strongly-convex function. We assume for simplicity that $\inf_w F(w) = 0$. We denote*

$$\begin{aligned} \chi &= \{x : \|x\| \leq X\} \\ F_W &= \{f : x \to \langle w, x\rangle / w \in W\} \\ W &= \{w : F(w) \leq W_*^2\} \end{aligned}$$

Then the following inequality holds: $R_n(F_W) \leq XW_* \sqrt{\frac{2}{\lambda n}}$

In the proof of this theorem, we are going to use the following lemma.

Lemma. *If* F *λ-strongly-convex* Z_i *independent variables,* $i = 1, \cdots, n$ $\mathbb{E}[Z_i] = 0$ *(centered)* $\mathbb{E}[Z_i^2] \leq B_Z$ $S_i = \sum_{j=1}^{i} Z_j$ *then (super-martingale)*

$$\mathbb{E}[F^*(S_i) - i\frac{B_Z}{2\lambda}|Z_1, \cdots, Z_{i-1}] \leq F^*(S_{i-1}) - (i-1)\frac{B_Z}{2\lambda}$$

If moreover $\inf(F) = 0$*, then* $\mathbb{E}[F^*(S_i)] \leq i\frac{B_Z}{2\lambda}$.

Proof of the lemma.

$$F^*(S_{n-1} + Z_n) \leq F^*(S_{n-1}) + \langle Z_n, \nabla F^*(S_{n-1})\rangle + \frac{1}{2\lambda}\|Z_n\|^2$$

Let's denote $\mathbb{E}_n[\,\cdot\,] = \mathbb{E}[\,\cdot\,|Z_1, \cdots, Z_n]$.

$$\mathbb{E}_{n-1}[F^*(S_{n-1} + Z_n)] \leq \mathbb{E}_{n-1}[F^*(S_{n-1})] + \frac{1}{2\lambda}B_Z + \mathbb{E}_{n-1}[\langle Z_n, \nabla F^*(S_{n-1})\rangle]$$

where

$$\mathbb{E}_{n-1}[F^*(S_{n-1})] = F^*(S_{n-1})$$

and

$$\begin{aligned}\mathbb{E}_{n-1}[\langle Z_n, \nabla F^*(S_{n-1})\rangle] &= \langle \mathbb{E}_{n-1}[Z_n], \nabla F^*(S_{n-1})\rangle \\ &= \langle \mathbb{E}[Z_n], \nabla F^*(S_{n-1})\rangle \\ &= 0\end{aligned}$$

Hence we finally get

$$\begin{aligned}\mathbb{E}_{n-1}[F^*(S_{n-1} + Z_n)] &\leq F^*(S_{n-1}) + \frac{B_Z}{2\lambda} \\ \mathbb{E}_{n-1}[F^*(S_n) - i\frac{B_Z}{2\lambda}] &\leq F^*(S_{n-1}) - (i-1)\frac{B_Z}{2\lambda}\end{aligned}$$

If moreover $\inf(F) = 0$ then $F^*(0) = -\inf(F) = 0$ and recursively we have

$$\begin{aligned}\mathbb{E}[F^*(S_0)] &= \mathbb{E}[F^*(0)] \\ &= 0 \\ \mathbb{E}[F^*(S_{i+1})] &= \mathbb{E}[\mathbb{E}_i[F^*(S_{i+1})]] \\ &\leq \mathbb{E}[F^*(S_i) + \frac{B_Z}{2\lambda}] \\ \mathbb{E}[F^*(S_{i+1})] &\leq \mathbb{E}[F^*(S_i)] + \frac{B_Z}{2\lambda} \\ &\leq i\frac{B_Z}{2\lambda} + \frac{B_Z}{2\lambda} \\ &\leq (i+1)\frac{B_Z}{2\lambda}\,.\end{aligned}$$

□

Proof of the theorem. Assume $\chi = \{x_1, \dots, x_n\}$ (with $\|x_i\| \leq X$). Let $\theta = \frac{1}{n}\sum_{i=1}^n \varepsilon_i x_i$. For $w \in W$ and $\alpha > 0$, we have from the definition of F^*: $\langle w, \alpha\theta\rangle \leq F(w) + F^*(\alpha\theta)$ hence

$$\begin{aligned}\frac{1}{\alpha}\langle w, \alpha\theta\rangle =& \langle w, \theta\rangle \\ \leq& \frac{1}{\alpha}(F(w) + F^*(\alpha\theta)) \\ \leq& \frac{1}{\alpha}(W_*^2 + F^*(\alpha\theta))\end{aligned}$$

with $\langle w, \theta\rangle = \frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle w, x_i\rangle$ where $\langle w, x_i\rangle = f_w(x_i)$.

Hence

$$\mathbb{E}[\sup_{w\in W} \langle w, \theta\rangle] \leq \frac{W_*^2}{\alpha} + \mathbb{E}[\sup_{w\in W} \frac{F^*(\alpha\theta)}{\theta}]$$

And from Fatou's lemma,

$$\mathbb{E}[\sup_{w\in W} \langle w, \theta\rangle] \leq \frac{W_*^2}{\alpha} + \sup_{w\in W} \mathbb{E}[\frac{F^*(\alpha\theta)}{\theta}]$$

Now, we use the previous lemma. Choose $Z_i = \frac{\alpha\varepsilon_i x_i}{n}$. The$\varepsilon_i$'s are centered variables thus $\mathbb{E}[Z_i] = \frac{\alpha}{n}\mathbb{E}[\varepsilon_i]\mathbb{E}[x_i] = 0$ and $\mathbb{E}[\|Z_i\|^2] = \frac{\alpha^2}{n^2}\mathbb{E}[\varepsilon_i^2]\mathbb{E}[\|x_i\|^2] = \frac{\alpha^2}{n^2}\mathbb{E}[\|x_i\|^2] \leq \frac{\alpha^2}{n^2}X^2$
Via the previous lemma, we get

$$\begin{aligned}\mathbb{E}[\frac{F^*(\alpha\theta)}{\alpha}] \leq& \frac{\alpha X^2}{2\lambda n} \\ \mathbb{E}[\sup_{w\in W} \langle w, \theta\rangle] \leq& \frac{W_*^2}{\alpha} + \frac{\alpha X^2}{2\lambda n}\end{aligned}$$

If $g(\alpha) = \frac{W_*^2}{\alpha} + \frac{\alpha X^2}{2\lambda n}$, then $g'(\alpha) = -\frac{W_*^2}{\alpha^2} + \frac{X^2}{2\lambda n} = 0$ if and only if $\alpha = \frac{W_*}{X}\sqrt{2\lambda n} = \alpha_0$
which leads to $\mathbb{E}[\sup_{w\in W} \langle w, \theta\rangle] \leq g(\alpha_0) = \frac{XW_*}{\sqrt{2\lambda n}} + \frac{XW_*\sqrt{2\lambda n}}{2\lambda n} = XW_*\sqrt{\frac{2}{\lambda n}}$ □

Example. *If $F = \frac{\|\cdot\|_2^2}{2}$ then F is $\frac{1}{2}$-strongly-convex and $W = \{w | \|w\|_2 \leq W_*\sqrt{2}\}$.*

3 Ridge regression (or Tikhonov regularization)

We consider here the classical regression model: $X \in \mathbb{R}^{n\times d}$, $Y \in \mathbb{R}^n$, $\theta^* \in \mathbb{R}^d$, $\varepsilon \in \mathbb{R}^n$ such that ε_i are centered and independent variables (noise) and $\mathrm{Var}(\varepsilon_i) \leq \sigma^2$. The Y is the observation vector, and X is a matrix whose columns are called features or explanatory variables. The observation model is the following:

$$Y = X\theta^* + \varepsilon.$$

One possible objective in this context is to recover θ^* in such a context.

A natural candidate is the least square estimate that solves

$$\hat{\theta}_{\mathrm{LS}} \in \arg\min_{\theta\in\mathbb{R}^d} \|Y - X\theta\|_2^2 .$$

Remark. *Note that the least square solution might not be unique.*

The ridge estimator, is a modified version, were some regularization on the Gramm matrix is done. It is the (unique!) solution of the problem:

$$\hat{\theta}_\lambda = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_2^2$$

where $\lambda > 0$

We call the prediction error the term $R = \frac{1}{n}\mathbb{E}[\|X\hat{\theta}_\lambda - X\theta^*\|_2^2]$

* $\lambda \to 0 \Rightarrow \hat{\theta}_\lambda \to \hat{\theta}_{\text{LS}} \in \arg\min_\theta \|Y - X\theta\|_2^2$ (least square estimate).
* $\lambda \to \infty \Rightarrow \hat{\theta}_\lambda \to 0$ (null estimate).

Let us denote by $\bar{\theta}_\lambda = \mathbb{E}[\hat{\theta}_\lambda]$.

Definition (Bias term)**.** *The bias is* $\bar{\theta}_\lambda - \theta^*$.

It is also called the approximation error, and it is a deterministic quantity.

Definition (Variance term)**.** *The variance term is given by the following formula:* $V_\lambda = \frac{1}{n}\mathbb{E}[\|X\hat{\theta}_\lambda - X\bar{\theta}_\lambda\|_2^2]$

We are looking for a bound on the prediction error as small as possible.

$$\begin{aligned} R =& \frac{1}{n}\mathbb{E}[\|X\hat{\theta}_\lambda - X\bar{\theta}_\lambda + X\bar{\theta}_\lambda - X\theta^*\|_2^2] \\ =& V_\lambda + \frac{1}{n}\|X\bar{\theta}_\lambda - X\theta^*\|_2^2 + \frac{1}{n}\mathbb{E}[\langle X(\hat{\theta}_\lambda - \bar{\theta}_\lambda), X(\bar{\theta}_\lambda - \theta^*)\rangle] \end{aligned}$$

But

$$\mathbb{E}[\langle X(\hat{\theta}_\lambda - \bar{\theta}_\lambda), X(\bar{\theta}_\lambda - \theta^*)\rangle] = \langle X\mathbb{E}[\hat{\theta}_\lambda - \bar{\theta}_\lambda], X(\bar{\theta}_\lambda - \theta^*)\rangle$$

and

$$\mathbb{E}[\hat{\theta}_\lambda - \bar{\theta}_\lambda] = \mathbb{E}[\hat{\theta}_\lambda] - \bar{\theta}_\lambda = 0.$$

Hence

$$R = V_\lambda + \frac{1}{n}\|X\bar{\theta}_\lambda - X\theta^*\|_2^2$$

Let's denote by $B_\lambda^2 = \frac{1}{n}\|X\bar{\theta}_\lambda - X\theta^*\|_2^2$ so that $R = V_\lambda + B_\lambda^2$. This is a bias/variance decomposition of the Mean Square Error.

In the following analysis, we first bound V_λ. From the definition of $\hat{\theta}_\lambda$, we know that

$$\nabla_\theta(\frac{1}{n}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_2^2)(\hat{\theta}_\lambda) = 2\lambda\hat{\theta}_\lambda + \frac{2}{n}X^T(X\hat{\theta}_\lambda - Y) = 0$$

which gives a new implicit expression for $\hat{\theta}_\lambda$, defining the Gramm matrix (or matrix of empirical correlations) $\Sigma = \frac{X^TX}{n}$:

$$(\Sigma + \lambda\,\mathrm{Id})\hat{\theta}_\lambda = \frac{1}{n}X^TY \quad \text{(Normal equation)}$$

Since Σ is non negative, $\Sigma + \lambda \operatorname{Id}$ is symmetric positive definite and in particular invertible (his is due to the regularization term added in the ridge estimator). So, the previous becomes

$$\begin{aligned}
\hat{\theta}_\lambda =& \frac{1}{n}(\Sigma + \lambda \operatorname{Id})^{-1} X^T Y \\
\hat{\theta}_\lambda =& \frac{1}{n}(\Sigma + \lambda \operatorname{Id})^{-1} X^T (X\theta^* + \varepsilon) \\
\bar{\theta}_\lambda =& (\Sigma + \lambda \operatorname{Id})^{-1} \Sigma \theta^* \\
\hat{\theta}_\lambda - \bar{\theta}_\lambda =& \frac{1}{n}(\Sigma + \lambda \operatorname{Id})^{-1} X^T \varepsilon
\end{aligned}$$

Now we can bound the variance term in the following way:

$$\begin{aligned}
V_n =& \mathbb{E}[(\hat{\theta}_\lambda - \bar{\theta}_\lambda)^T \Sigma (\hat{\theta}_\lambda - \bar{\theta}_\lambda)] \\
=& \mathbb{E}[(\frac{1}{n}(\Sigma + \lambda \operatorname{Id})^{-1} X^T \varepsilon)^T \Sigma (\frac{1}{n}(\Sigma + \lambda \operatorname{Id})^{-1} X^T \varepsilon)] \\
=& \frac{1}{n^2} \mathbb{E}[\varepsilon^T X (\Sigma + \lambda \operatorname{Id})^{-1} \Sigma (\Sigma + \lambda \operatorname{Id})^{-1} X^T \varepsilon] \\
=& \frac{1}{n^2} \mathbb{E}[\operatorname{Tr}(\varepsilon^T X (\Sigma + \lambda \operatorname{Id})^{-1} \Sigma (\Sigma + \lambda \operatorname{Id})^{-1} X^T \varepsilon)] \\
=& \frac{1}{n^2} \mathbb{E}[\operatorname{Tr}(\varepsilon \varepsilon^T X (\Sigma + \lambda \operatorname{Id})^{-1} \Sigma (\Sigma + \lambda \operatorname{Id})^{-1} X^T)] \\
=& \frac{1}{n^2} \operatorname{Tr}(\mathbb{E}[\varepsilon \varepsilon^T X (\Sigma + \lambda \operatorname{Id})^{-1} \Sigma (\Sigma + \lambda \operatorname{Id})^{-1} X^T]) \\
=& \frac{1}{n^2} \operatorname{Tr}(\mathbb{E}[\varepsilon \varepsilon^T] X (\Sigma + \lambda \operatorname{Id})^{-1} \Sigma (\Sigma + \lambda \operatorname{Id})^{-1} X^T) \\
=& \frac{1}{n^2} \operatorname{Tr}(\mathbb{E}[\varepsilon \varepsilon^T] X \Sigma (\Sigma + \lambda \operatorname{Id})^{-2} X^T).
\end{aligned}$$

Now, use that $\mathbb{E}[\varepsilon \varepsilon^T] = \operatorname{diag}(\sigma_i^2)$ with $\sigma_i^2 \leq \sigma^2$. Since $X\Sigma(\Sigma + \lambda \operatorname{Id})^{-2} X^T$ is symmetric non-negative

$$\begin{aligned}
V_n \leq& \frac{\sigma^2}{n^2} \operatorname{Tr}(X \Sigma (\Sigma + \lambda \operatorname{Id})^{-2} X^T) \\
\leq& \frac{\sigma^2}{n^2} \operatorname{Tr}(X^T X \Sigma (\Sigma + \lambda \operatorname{Id})^{-2}) \\
\leq& \frac{\sigma^2}{n} \operatorname{Tr}(\Sigma^2 (\Sigma + \lambda \operatorname{Id})^{-2})
\end{aligned}$$

We know from the spectral decomposition theorem that $\Sigma = Q^T \operatorname{diag}(\lambda_i) Q$ with $Q^T Q = \operatorname{Id}$ with $\forall i = 1, \cdots, d, \lambda_i \geq 0$, so:

$$\begin{aligned}
V_n \leq& \frac{\sigma^2}{n} \sum_{i=1}^{d} (\frac{\lambda_i}{\lambda + \lambda_i})^2 \\
\leq& \frac{\sigma^2}{n} \frac{\sum_{i=1}^{d} \lambda_i}{2\lambda} \\
\leq& \frac{\sigma^2}{n} \frac{\operatorname{Tr}(\Sigma)}{2\lambda}
\end{aligned}$$

Now, let's look propose a bound on B_λ^2:

$$\begin{aligned}
B_\lambda^2 =& \frac{1}{n}\|X(\bar{\theta}_\lambda - \theta^*)\|_2^2 \\
=& (\bar{\theta}_\lambda - \theta^*)^T \Sigma (\bar{\theta}_\lambda - \theta^*) \\
=& \langle ((\Sigma + \lambda \operatorname{Id})^{-1}\Sigma - \operatorname{Id})\theta^*, \Sigma(\Sigma + \lambda \operatorname{Id})^{-1}\Sigma - \operatorname{Id})\theta^* \rangle \\
=& \theta^{*T}\Sigma((\Sigma + \lambda \operatorname{Id})^{-1}\Sigma - \operatorname{Id})^2\theta^* \\
=& \theta^{*T} Q^T \operatorname{diag}[\lambda_i (\frac{\lambda_i}{\lambda_i + \lambda} - 1)^2] Q\theta^* \\
=& \theta^{*T} Q^T \operatorname{diag}[\lambda_i (\frac{\lambda}{\lambda_i + \lambda})^2] Q\theta^* \\
\leq& \frac{\lambda}{2}\|Q\theta^*\|_2^2 \\
\leq& \frac{\lambda}{2}\|\theta^*\|_2^2
\end{aligned}$$

We eventually get the following control for the error term:

$$\begin{aligned}
R =& V_n + B_\lambda^2 \\
R \leq& \frac{1}{\lambda}\frac{\sigma^2}{2n}\operatorname{Tr}(\Sigma) + \frac{\lambda}{2}\|\theta^*\|_2^2
\end{aligned}$$

Since we want the error bound to be as accurate as possible, we then optimize over λ:

$$\nabla_\lambda (\frac{1}{\lambda}\frac{\sigma^2}{2n}\operatorname{Tr}(\Sigma) + \frac{\lambda}{2}\|\theta^*\|_2^2) = -\frac{1}{\lambda^2}\frac{\sigma^2}{2n}\operatorname{Tr}(\Sigma) + \frac{1}{2}\|\theta^*\|_2^2 = 0\,.$$

The bound reaches its minimal value at $\lambda_0 = \frac{\sigma}{\|\theta^*\|_2}\sqrt{\frac{\operatorname{Tr}(\Sigma)}{n}}$ so $R \leq \frac{1}{\lambda_0}\frac{\sigma^2}{2n}\operatorname{Tr}(\Sigma) + \frac{\lambda_0}{2}\|\theta^*\|_2^2 = \sigma\|\theta^*\|_2\sqrt{\frac{\operatorname{Tr}(\Sigma)}{n}}$

Remark. *In this expression, θ^* is still unknown, but appears in the bond only in terms of $\|\theta^*\|^2$, a scalar quantity that might be easier to estimate than the full vector θ^*.*

Appendix

Lemma (Hoeffding)**.** *Let X be a centered random variable, with $X \in [a, b]$ a.s. for $(a, b) \in \mathbb{R}^2$, then for all $s > 0$*

$$\mathbb{E}[\exp(sX)] \leq \exp(s^2(b-a)/8) \tag{13}$$

Proof. Use the convexity of the exp function to get for any $z \in [a, b]$:

$$\exp(sz) \leq \frac{z-a}{b-a}\exp(sb) + \frac{b-z}{b-a}\exp(sa) \tag{14}$$

Integrating and using that $\mathbb{E}[Z] = 0$ it holds that

$$\begin{aligned}
\mathbb{E}[\exp(sZ)] \leq& \frac{-a}{b-a}\exp(sb) + \frac{b}{b-a}\exp(sa) \\
\mathbb{E}[\exp(sZ)] \leq& \exp(sa)\left(\frac{b}{b-a} - \frac{a}{b-a}\exp\left(s[b-a]\right)\right)
\end{aligned}$$

Define $p = -\frac{a}{b-a}$ and $u = s(b-a)$ the previous inequality becomes

$$\mathbb{E}[\exp(sZ)] \leq \exp(-pu)\left(1 - p + p\exp(u)\right) = \exp(\varphi(u))$$

where $\varphi(u) = \log\left(1 - p + p\exp(u)\right) - up$. Note that $\varphi(u) = 0 = \varphi'(u)$ and that $\varphi''(u) \leq 1/4$ for all $u \geq 0$. Using Taylor-Lagrange to the second order, one gets that for some $y \in [0, u]$

$$\begin{aligned}\mathbb{E}[\exp(sZ)] &\leq \exp(\varphi(0) + u\varphi'(0) + u^2\varphi''(y)/2) \\ &\leq \exp(u^2/8) = \exp((b-a)^2/8)\end{aligned}$$

□

Theorem (McDiarmid's Inequality)**.** *If $(X_1, \cdots, X_k)$ are independent variables and we have f a bounded difference function, i.e.,*

$$\forall i = 1, \cdots, n, \quad \sup_{x_1, \cdots, x_n, x_i'} |f(x_1, \cdots, x_n) - f(x_1, \cdots, x_i', \cdots x_n)| \leq c_i \tag{15}$$

then $\forall t > 0$,

$$\mathbb{P}\left(f(X_1, \cdots, X_n) - \mathbb{E}f(X_1, \cdots, X_n) \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right) \tag{16}$$

Proof. Let us introduce for $k = 1, \cdots, n$ the quantity $H_k(X_1, \cdots, X_k) = \mathbb{E}[f(X_1, \cdots, X_n)|X_1, \cdots, X_k]$ (with the convention $H_0(X_0) = \mathbb{E}[f(X_1, \cdots, X_n)]$) and let $\Delta_k = H_k(X_1, \cdots, X_k) - H_{k-1}(X_1, \cdots, X_{k-1})$. First by construction:

$$\begin{aligned}\sum_{k=1}^n \Delta_k &= \sum_{k=1}^n \left(H_k(X_1, \cdots, X_k) - H_{k-1}(X_1, \cdots, X_{k-1})\right) \\ &= \mathbb{E}[f(X_1, \cdots, X_n)|X_1, \cdots, X_n] - \mathbb{E}[f(X_1, \cdots, X_n)] \\ &= f(X_1, \cdots, X_n) - \mathbb{E}[f(X_1, \cdots, X_n)]\end{aligned}$$

Then, introduce

$$\begin{aligned}\Phi_k(X_1, \cdots, X_k) &= \inf_x \left(H_k(X_1, \cdots, X_{k-1}, x) - H_k(X_1, \cdots, X_{k-1})\right) \\ \Psi_k(X_1, \cdots, X_k) &= \sup_y \left(H_k(X_1, \cdots, X_{k-1}, y) - H_k(X_1, \cdots, X_{k-1})\right).\end{aligned}$$

Reminding that the X_i are independent it holds that

$$\begin{aligned}H_k(X_1, \cdots, X_{k-1}, x) =& \mathbb{E}[f(X_1, \cdots, X_k)|X_1, \cdots, X_{k-1}, X_k = x] \\ =& \int f(X_1, \cdots, X_{k-1}, x, x_{k+1}, \cdots, x_n) d\mathbb{P}_{X_{k+1}}(x_{k+1}) \cdots d\mathbb{P}_{X_n}(x_n)\end{aligned}$$

Thus, introducing $G_k = \Psi_k(X_1, \cdots, X_k) - \Phi_k(X_1, \cdots, X_k)$, it holds that

$$\begin{aligned}
G_k =& \sup_y \left(H_k(X_1, \cdots, X_{k-1}, y) - H_k(X_1, \cdots, X_{k-1})\right) \\
& - \inf_x \left(H_k(X_1, \cdots, X_{k-1}, x) - H_k(X_1, \cdots, X_{k-1})\right) \\
=& \sup_{x,y} \left(H_k(X_1, \cdots, X_{k-1}, y) - H_k(X_{k-1}, \cdots, X_{k-1}, x)\right) \\
=& \sup_{x,y} \left[\int f(X_1, \cdots, X_{k-1}, y, x_{k+1}, \cdots, x_n) d\mathbb{P}_{X_{k+1}}(x_{k+1}) \cdots d\mathbb{P}_{X_n}(x_n) \right. \\
& \left. - \int f(X_1, \cdots, X_{k-1}, x, x_{k+1}, \cdots, x_n) d\mathbb{P}_{X_{k+1}}(x_{k+1}) \cdots d\mathbb{P}_{X_n}(x_n) \right] \\
=& \sup_{x,y} \int \left[f(X_1, \cdots, X_{k-1}, x, x_{k+1}, \cdots, x_n) - f(X_1, \cdots, X_{k-1}, y, x_{k+1}, \cdots, x_n) \right] d\mathbb{P}_{X_{k+1}}(x_{k+1}) \cdots d\mathbb{P}_{X_n}(x_n) \\
\leq & \int c_k d\mathbb{P}_{X_{k+1}}(x_{k+1}) \cdots d\mathbb{P}_{X_n}(x_n) = c_k.
\end{aligned}$$

where the inequality is provided by the bounded difference assumption on f given in (15). Thus, we can write that $\Psi_k(X_1, \cdots, X_k) \leq \Phi_k(X_1, \cdots, X_k) + c_k$, leading to

$$\Phi_k(X_1, \cdots, X_k) \leq \mathbb{E}[\Delta_k | X_1, \cdots, X_{k-1}] \leq \Phi_k(X_1, \cdots, X_k) + c_k \tag{17}$$

Now let us prove that for all $s > 0$

$$\mathbb{E}\left[\exp(s\Delta_k) | X_1, \cdots, X_{k-1}\right] \leq \exp(\frac{s^2 c_k^2}{8}). \tag{18}$$

First, remind that Indeed, $\mathbb{E}\left[\mathbb{E}[\Delta_k | X_1, \cdots, X_{k-1}] | X_1, \cdots, X_k\right] = 0$ Then, remind that for (condially) bounded variables one has the following bound:

$$\mathbb{E}[\exp(s\Delta)] \leq \exp(\frac{s^2 c^2}{8}), \tag{19}$$

when Δ is centered and belongs to an interval $[a, b]$ of length c (cf. Lemma)

□

References

[BM02] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. J. Mach. Learn. Res., 3(Spec. Issue Comput. Learn. Theory):463–482, 2002.

[HUL93a] J-B. Hiriart-Urruty and C. Lemaréchal. Convex analysis and minimization algorithms. I, volume 305. Springer-Verlag, Berlin, 1993.

[HUL93b] J-B. Hiriart-Urruty and C. Lemaréchal. Convex analysis and minimization algorithms. II, volume 306. Springer-Verlag, Berlin, 1993.