Foundations of Machine Learning Book of Proofs

E. Le Pennec

January 24, 2025

In this document, you will ultimately find all the proofs of the results given in the lectures. At this time, you will either find the proof or a pointer to a book where you can find them.

Please inform me if there is a missing proof!

Contents

1. Statistical Learning: Introduction, Setting and Risk Estimation

1.1. Bayes Predictor

Claim 1.1.1

The minimizer of $\mathbb{E}\left[\ell^{0/1}(Y, f(\underline{X}))\right]$ *is given by*

$$f^\star(\underline{X}) = \begin{cases} +1 & \text{if } \mathbb{P}(Y=+1|\underline{X}) \geq \mathbb{P}(Y=-1|\underline{X}) \\ & \Leftrightarrow \mathbb{P}(Y=+1|\underline{X}) \geq 1/2 \\ -1 & \text{otherwise} \end{cases}$$

Proof. We start by noticing that

$$\arg\min_{f\in\mathcal{F}} \mathbb{E}[\ell(Y, f(\underline{X}))] = \arg\min_{f\in\mathcal{F}} \mathbb{E}_{\underline{X}}\Big[\mathbb{E}_{Y|\underline{X}}[\ell(Y, f(\underline{X}))]\Big]$$

so that we can focus on

$$\mathbb{E}_{Y|\underline{X}}[\ell(Y, f(\underline{X}))]$$

where $f(\underline{X})$ is constant.

By definition,

$$\begin{aligned} \mathbb{E}_{Y|\underline{X}}[\ell(Y, f(\underline{X}))] &= \mathbb{P}(Y=1|\underline{X})\,\ell(1, f(\underline{X})) + \mathbb{P}(Y=-1|\underline{X})\,\ell(-1, f(\underline{X})) \\ &= \begin{cases} \mathbb{P}(Y=1|\underline{X}) & \text{if } f(\underline{X}) = -1 \\ \mathbb{P}(Y=-1|\underline{X}) & \text{if } f(\underline{X}) = 1 \end{cases} \end{aligned}$$

which implies

$$f^\star(\underline{X}) = \begin{cases} +1 & \text{if } \mathbb{P}(Y=+1|\underline{X}) \geq \mathbb{P}(Y=-1|\underline{X}) \\ -1 & \text{otherwise} \end{cases}$$

.

The last element of the theorem is obtained by noticing that $\mathbb{P}(Y=+1|\underline{X}) \geq \mathbb{P}(Y=-1|\underline{X}) \Leftrightarrow \mathbb{P}(Y=+1|\underline{X}) \geq 1/2$. □

Claim 1.1.2

The minimizer of $\mathbb{E}\left[\ell^2(Y, f(\underline{X}))\right]$ *is given by*

$$f^\star(\underline{X}) = \mathbb{E}[Y|\underline{X}]$$

Proof. We start by noticing that

$$\arg\min_{f\in\mathcal{F}} \mathbb{E}[\ell(Y, f(\underline{X}))] = \arg\min_{f\in\mathcal{F}} \mathbb{E}_{\underline{X}}\Big[\mathbb{E}_{Y|\underline{X}}[\ell(Y, f(\underline{X}))]\Big]$$

so that we can focus on

$$\mathbb{E}_{Y|\underline{X}}[\ell(Y, f(\underline{X}))] = \mathbb{E}_{Y|\underline{X}}\Big[(Y - f(\underline{X}))^2\Big]$$

where $f(\underline{X})$ is constant.

Now using the definition of the conditional expectation, we obtain then

$$\begin{aligned}
\mathbb{E}_{Y|\underline{X}}[\ell(Y, f(\underline{X}))] &= \mathbb{E}_{Y|\underline{X}}\Big[(Y - f(\underline{X}))^2\Big] \\
&= \mathbb{E}_{Y|\underline{X}}\Big[(Y - \mathbb{E}[Y|\underline{X}] + \mathbb{E}[Y|\underline{X}] - f(\underline{X}))^2\Big] \\
&= \mathbb{E}_{Y|\underline{X}}\Big[(Y - \mathbb{E}[Y|\underline{X}])^2\Big] + \mathbb{E}_{Y|\underline{X}}\Big[(\mathbb{E}[Y|\underline{X}] - f(\underline{X}))^2\Big] \\
&\quad + 2\mathbb{E}_{Y|\underline{X}}[(Y - \mathbb{E}[Y|\underline{X}])(\mathbb{E}[Y|\underline{X}] - f(\underline{X}))] \\
&= \mathbb{E}_{Y|\underline{X}}\Big[(Y - \mathbb{E}[Y|\underline{X}])^2\Big] + (\mathbb{E}[Y|\underline{X}] - f(\underline{X}))^2
\end{aligned}$$

which is thus minimized by $f^\star(\underline{X}) = \mathbb{E}[Y|\underline{X}]$. □

1.2. Training Risk Optimism

Let

$$\mathcal{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(\underline{X}_i))$$

and

$$\widehat{f}_{\mathcal{S}} = \arg\min_{f\in\mathcal{S}} \mathcal{R}_n(f)$$

.

Claim 1.2.1

$$\mathcal{R}_n(\widehat{f}_{\mathcal{S}}) \leq \mathcal{R}_n(f^\star_{\mathcal{S}}) \quad \textit{and} \quad \mathbb{E}\Big[\mathcal{R}_n(\widehat{f}_{\mathcal{S}})\Big] \leq \mathcal{R}(f^\star_{\mathcal{S}})$$

Proof. The first part is nothing but the definition of $\widehat{f_{\mathcal{S}}}$ combined with the fact that $f^\star_{\mathcal{S}}$ also belongs to $\mathcal{S}$.

The second part relies on the fact that for a non-random function

$$\mathbb{E}[\mathcal{R}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(Y_i, f(\underline{X}_i))\right] = \mathbb{E}[\ell(Y, f(\underline{X}))] = \mathcal{R}(f)$$

□

1.3. Leave One Out Formula

Claim 1.3.1

For the least squares linear regression,

$$\widehat{f}^{-i}(\underline{X}_i) = \frac{\widehat{f}(\underline{X}_i) - h_{ii}Y_i}{1 - h_{ii}}$$

with h_{ii} the ith diagonal coefficient of the hat (projection) matrix.

Proof. By construction,

$$\hat{f}^{-i}(\underline{X}_i) = \underline{X}_i^{\Phi\top}\hat{\beta}^{-i} = \underline{X}_i{}^{\top}(\underline{X}^{\Phi}_{(n)-i}{}^{\top}\underline{X}^{\Phi}_{(n)-i})^{-1}\underline{X}^{\Phi}_{(n)-i}{}^{\top}\mathbb{Y}_{(n)-i}$$

Now $\underline{X}^{\Phi}_{(n)-i}{}^{\top}\underline{X}^{\Phi}_{(n)-i} = \mathbb{X}_{(n)}{}^{\Phi\top}\mathbb{X}^{\Phi}_{(n)} - \underline{X}^{\Phi}_i\underline{X}_i{}^{\top}$ and $\underline{X}^{\Phi}_{(n)-i}{}^{\top}\underline{Y}_{(n)-i} = \mathbb{X}_{(n)}{}^{\Phi\top}\mathbb{Y}_{(n)} - \underline{X}^{\Phi}_i Y_i$

Using $(M + uv^\top)^{-1} = M^{-1} - \frac{M^{-1}uv^\top M^{-1}}{1+u^\top M^{-1}v}$ with $M = \mathbb{X}^t_{(n)}\mathbb{X}_{(n)}$, $u = -v = \underline{X}_i$ yields:

$$\hat{f}^{-i}(\underline{X}_i) = \underline{X}_i^{\Phi\top}\left(M^{-1} + \frac{M^{-1}\underline{X}^{\Phi}_i\underline{X}^{\Phi\top}_i M^{-1}}{1 - \underline{X}^{\Phi\top}_i M^{-1}\underline{X}^{\Phi}_i}\right)\left(\mathbb{X}_{(n)}{}^{\Phi\top}\mathbb{Y}_{(n)} - \underline{X}^{\Phi}_i Y_i\right)$$

using $h_{ii} = \underline{X}^{\Phi\top}_i M^{-1}\underline{X}^{\Phi}_i$

$$= \hat{f}(\underline{X}_i) + \frac{h_{ii}}{1-h_{ii}}\hat{f}(\underline{X}_i) - h_{ii}Y_i - \frac{h^2_{ii}}{Y}{}_i$$

$$\hat{f}^{-i}(\underline{X}_i) = \frac{\hat{f}(\underline{X}_i) - h_{ii}Y_i}{1 - h_{ii}}$$

□

1.4. Weighted Loss and Bayes Estimator

We assume here that the loss $\ell(Y, f(\underline{X})) = C(Y)\ell^{0/1}(Y, f(\underline{X}))$ in a multi-class setting.

Claim 1.4.1

The minimizer of $\mathbb{E}[(Y, f(\underline{X}))]$ *is given by*

$$f^\star(\underline{X}) = \arg\max_k C(k)\mathbb{P}(Y = k|\underline{X})$$

Proof. As in the binary $\ell^{0/1}$ setting, we can condition with $\underline{X}$

$$\begin{aligned}\mathbb{E}_{Y|\underline{X}}[\ell(Y, f(\underline{X}))] &= \sum_k C(k)\ell^{0/1}(k, f(\underline{X}))\mathbb{P}(Y = k|\underline{X}) \\ &= \sum_{k\neq f(\underline{X})} C(k)\mathbb{P}(Y = k|\underline{X}) \\ &= -C(f(\underline{X}))\mathbb{P}\Big(Y = f(\vec{(}X))|\underline{X}\Big) + \sum kC(k)\mathbb{P}(Y = k|\underline{X})\end{aligned}$$

which is minimized by taking $f(\underline{X})$ equal to the k with the largest $C(k)\mathbb{P}(Y = k|\underline{X})$. □

2. ML Methods: Probabilistic Point of View

2.1. Classification Risk Analysis with a Probabilistic Point of View

Claim 2.1.1

If $\widehat{f} = \text{sign}(2\widehat{p}_{+1} - 1)$ *then*

$$\begin{aligned}\mathbb{E}\Big[\ell^{0,1}(Y, \widehat{f}(\underline{X}))\Big] - \mathbb{E}\Big[\ell^{0,1}(Y, f^\star(\underline{X}))\Big] &\leq \mathbb{E}\Big[\|\widehat{Y|\underline{X}} - Y|\underline{X}\|_1\Big] \\ &\leq \Big(\mathbb{E}\Big[2\,\text{KL}(Y|\underline{X}, \widehat{Y|\underline{X}}\Big]\Big)^{1/2}\end{aligned}$$

Proof. Let us denote $p_1(\underline{X}) = \mathbb{P}(Y = 1|\underline{X})$.

Step 1: Let $\tilde{f}(\underline{X}) = \text{sign}(2\tilde{p}_1(\underline{X}) - 1)$

$$\begin{aligned}\mathbb{E}\Big[\ell^{0/1}(Y, \tilde{f}(\underline{X}))\Big] &= \mathbb{E}_{\underline{X}}\Big[p_1(\underline{X})\mathbf{1}_{\tilde{f}(\underline{X})=-1} + (1 - p_1(\underline{X}))\mathbf{1}_{\tilde{f}(\underline{X})=1}\Big] \\ &= \mathbb{E}_{\underline{X}}\Big[(1 - p_1(\underline{X})) + (2p_1(\underline{X}) - 1)\mathbf{1}_{\tilde{f}(\underline{X})=-1}\Big]\end{aligned}$$

Step 2:

$$\begin{aligned}&\mathbb{E}\Big[\ell^{0/1}(Y, \tilde{f}(\underline{X}))\Big] - \mathbb{E}\Big[\ell^{0/1}(Y, \tilde{f}^\star(\underline{X}))\Big] \\ &\quad = \mathbb{E}_{\underline{X}}\Big[(2p_1(\underline{X}) - 1)(\mathbf{1}_{\tilde{f}(\underline{X})=-1} - \mathbf{1}_{f^\star(\underline{X})=-1})\Big]\end{aligned}$$

using the definition of $f^\star = \text{sign}(2p(\underline{X} - 1)$

$$= \mathbb{E}_{\underline{X}}\Big[|2p_1(\underline{X}) - 1|\mathbf{1}_{f^\star(\underline{X})\neq\tilde{f}(\underline{X})}\Big]$$

and using the fact that $f^\star(\underline{X}) \neq \tilde{f}(\underline{X})$ implies that $\widehat{p}(\underline{X})$ and $p(\underline{X})$ are not on the same side with respect to $1/2$

$$\leq 2\mathbb{E}_{\underline{X}}[|p_1(\underline{X}) - \widehat{p}_1(\underline{X})|]) = \mathbb{E}_{\underline{X}}[\|p(\underline{X}) - \widehat{p}(\underline{X})\|_1]$$

using $\|P - Q\|_1 \leq \sqrt{2\,\text{KL}(P, Q)}$ and Jensen

$$\leq \mathbb{E}_{\underline{X}}\Big[\sqrt{2\,\text{KL}(p(\underline{X}), \widehat{p}(\underline{X}))}\Big] \leq (\mathbb{E}_{\underline{X}}[2\,\text{KL}(p(\underline{X}), \widehat{p}(\underline{X}))])^{1/2}$$

□

2.2. Logistic Likelihood and Convexity

Claim 2.2.1

The maximum likelihood estimate of the logistic model is given by

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-Y_i(\underline{X}_i{}^\top \beta)}\right)$$

and the minimized function is convex in β.

Proof.

$$\begin{aligned}
&- \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{1}_{Y_i=1} \log(h(\underline{X}_i{}^\top \beta)) + \mathbf{1}_{Y_i=-1} \log(1 - h(\underline{X}_i{}^\top \beta))\right) \\
&= -\frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{1}_{Y_i=1} \log \frac{e^{\underline{X}_i{}^\top \beta}}{1 + e^{\underline{X}_i^\top \beta}} + \mathbf{1}_{Y_i=-1} \log \frac{1}{1 + e^{\underline{X}_i{}^\top \beta}}\right) \\
&= -\frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{1}_{Y_i=1} \log \frac{1}{1 + e^{-\underline{X}_i^\top \beta}} + \mathbf{1}_{Y_i=-1} \log \frac{1}{1 + e^{\underline{X}_i{}^\top \beta}}\right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-Y_i(\underline{X}_i{}^\top \beta)}\right)
\end{aligned}$$

Now let $g(\beta) = \log(1 + e^{-Y(\underline{X})^\top \beta})$, a brute force computation yields

$$\begin{aligned}
\nabla g(\beta) &= Y \frac{e^{-Y\underline{X}^\top \beta}}{1 + e^{-Y\underline{X}^\top \beta}} \underline{X} \\
\nabla^2 g(\beta) &= \frac{e^{-Y\underline{X}^\top \beta}}{1 + e^{-Y\underline{X}^\top \beta}} \frac{1}{1 + e^{-Y\underline{X}^\top \beta}} \underline{X}\underline{X}^\top
\end{aligned}$$

and thus $\nabla^2 g(\beta)$ is sdp which implies the convexity of g and hence of the likelihood of the logistic. □

3. ML Methods: Optimization Point of View

3.1. Classical Convexification

Claim 3.1.1

The following three losses

- *Logistic loss:* $\bar{\ell}(Y, f(\underline{X})) = \log_2(1 + e^{-Yf(\underline{X})})$ *(Logistic / NN)*
- *Hinge loss:* $\bar{\ell}(Y, f(\underline{X})) = (1 - Yf(\underline{X}))_+$ *(SVM)*
- *Exponential loss:* $\bar{\ell}(Y, f(\underline{X})) = e^{-Yf(\underline{X})}$ *(Boosting...)*

satisfy

$$\bar{\ell}(Y, f(\underline{X})) = l(Yf(\underline{X}))$$

with l a decreasing convex function, differentiable at 0 and such that $l'(0) < 0$.
Furthermore $\ell(Y, f(\underline{X})) \geq \ell^{0/1}(Y, f(\underline{X}))$

Proof. For the logistic loss, $l(z) = \log_2(1 + e^{-z})$. So that l is differentiable everywhere

$$l'(z) = -\frac{1}{\log(2)} \frac{e^{-z}}{1 + e^{-z}}$$
$$l''(z) = \frac{1}{\log(2)} \frac{e^{-z}}{(1 + e^{-z})^2}.$$

Thus $l'(z) < 0$ and l is decreasing with $l'(0) < 0$. Now $l''(z) > 0$ and thus l is convex.

For the hinge loss, $l(z) = \max(0, 1-z)$. This is a decreasing function, l is differentiable at 0 with $l'(0) = -1$ and l is convex as the maximum of two affine (thus convex) functions.

For the exponential loss, $l(z) = e^{-z}$. So that l is differentiable everywhere

$$l'(z) = -e^{-z}$$
$$l''(z) = e^{-z}.$$

Thus $l'(z) < 0$ and l is decreasing with $l'(0) < 0$. Now $l''(z) > 0$ and thus l is convex.

For the three losses, by construction, $l(0) = 1$ and $l(z) \geq 0$ thus $\bar{\ell}(Y, f(\underline{X})) = l(Yf(\vec{(}X))) \geq 1$ when $Yf(\vec{(}X)) \leq 0$ and $\bar{\ell}(Y, f(\underline{X})) \geq 0$ otherwise. We obtain thus that $\ell(Y, f(\underline{X})) \geq \ell^{0/1}(Y, f(\underline{X}))$.

□

3.2. Classification Risk Analysis with an Optimization Point of View

Claim 3.2.1

The minimizer of

$$\mathbb{E}\Big[\bar{\ell}(Y, f(\underline{X}))\Big] = \mathbb{E}[l(Yf(\underline{X}))]$$

is the Bayes classifier $f^\star = \text{sign}(2\eta(\underline{X}) - 1)$

Furthermore it exists a convex function Ψ *such that*

$$\Psi\left(\mathbb{E}\Big[\ell^{0/1}(Y, \text{sign}(f(\underline{X}))\Big] - \mathbb{E}\Big[\ell^{0/1}(Y, f^\star(\underline{X})\Big]\right) \leq \mathbb{E}\Big[\bar{\ell}(Y, f(\underline{X})\Big] - \mathbb{E}\Big[\bar{\ell}(Y, f^\star(\underline{X}))\Big]$$

Proof. By definition,

$$\mathbb{E}[l(Yf)|\underline{X}] = \eta(\underline{X})l(f) + (1 - \eta(\underline{X}))l(-f)$$

Let $H(f, \eta) = \eta l(f) + (1 - \eta)l(-f)$, the optimal value for $\tilde{f}$ satisfies

$$\delta H(\tilde{f}, \eta) = \eta \delta l(\tilde{f}) - (1 - \eta)\delta l(-\tilde{f}) \ni 0.$$

With a slight abuse of notation, we denote by $\delta l(\tilde{f})$ and $\delta l(-\tilde{f})$ the two subgradients such that

$$\eta \delta l(\tilde{f}) - (1 - \eta)\delta l(-\tilde{f}) = 0$$

Now we discuss the sign of $\tilde{f}$:

- If $\tilde{f} > 0$, $\delta l(-\tilde{f}) < \delta l(\tilde{f})$ and thus $\eta > (1 - \eta)$, i.e. $2\eta - 1 > 0$.
- Conversely, if $\tilde{f} < 0$ then $2\eta - 1 < 0$

Thus $\text{sign}(\tilde{f}) = \text{sign}(2\eta - 1)$ i.e. the minimizer of $\mathbb{E}[l(yf)|\underline{X}]$ is such that $\text{sign}(f^\star(\underline{X})) = \text{sign}(2\eta(\underline{X}) - 1)$

We define $H(\eta) = \inf_f H(f, \eta) = \inf_f (\eta l(f) + (1 - \eta)l(-f))$. By construction, H is a concave function satisfying $H(1/2 + x) = H(1/2 - x)$.

Furthermore, one verify that if we consider the minimum over the *wrong sign classifiers*, $\inf_{f, f(2\eta-1)<0} H(f, \eta) = l(0)$.

Indeed,

$$\begin{aligned} H(f, \eta) &= \eta l(f) + (1 - \eta)l(-f) \\ &\geq \eta(l(0) + l'(0)f) + (1 - \eta)(l(0) - l'(0)f) \\ &\geq l(0) + l'(0)f(2\eta - 1) \end{aligned}$$

so that

$$\inf_{f,f(2\eta-1)<0} H(f,\eta) \geq l(0) + \inf_{f,f(2\eta-1)<0} l'(0) f(2\eta-1) = l(0)$$

Furthermore,

$$\begin{aligned}\mathbb{E}\Big[\bar{\ell}(Y, f(\underline{X})\Big] &= \mathbb{E}_{\underline{X}}[H(f, \eta(\underline{X})] \\ \mathbb{E}\Big[\bar{\ell}(Y, f^\star(\underline{X}))\Big] &= \mathbb{E}_{\underline{X}}[H(\eta(\underline{X})]\end{aligned}$$

We define then

$$\Psi(\theta) = l(0) - H\left(\frac{1+\theta}{2}\right)$$

which is thus a convex function satisfying $\Psi(0) = 0$ and $\Psi(\theta) > 0$ for $\theta > 0$.

Recall that

$$\begin{aligned}&\mathbb{E}\Big[\ell^{0/1}(Y, \text{sign}(f(\underline{X})))\Big] - \mathbb{E}\Big[\ell^{0/1}(Y, f^\star(\underline{X}))\Big] \\ &\quad = \mathbb{E}_{\underline{X}}\Big[|2\eta(\underline{X}) - 1| \mathbf{1}_{f^\star(\underline{X}) \neq \text{sign}(f(\underline{X}))}\Big]\end{aligned}$$

Using Jensen inequality, we derive

$$\begin{aligned}&\Psi\left(\mathbb{E}\Big[\ell^{0/1}(Y, \text{sign}(f(\underline{X})))\Big] - \mathbb{E}\Big[\ell^{0/1}(Y, f^\star(\underline{X}))\Big]\right) \\ &\quad \leq \mathbb{E}_{\underline{X}}\Big[\Psi\left(|2\eta(\underline{X}) - 1| \mathbf{1}_{f^\star(\underline{X}) \neq \text{sign}(f(\underline{X}))}\right)\Big]\end{aligned}$$

Using $\Psi(0) = 0$ and the symmetry of H,

$$\begin{aligned}&\Psi\left(\mathbb{E}\Big[\ell^{0/1}(Y, \text{sign}(f(\underline{X})))\Big] - \mathbb{E}\Big[\ell^{0/1}(Y, f^\star(\underline{X}))\Big]\right) \\ &\quad \leq \mathbb{E}_{\underline{X}}\left[\left(l(0) - H\left(\left(\frac{1+|2\eta(\underline{X})-1|}{2}\right)\right)\right) \mathbf{1}_{f^\star(\underline{X}) \neq \text{sign}(f(\underline{X}))}\right] \\ &\quad \leq \mathbb{E}_{\underline{X}}\Big[(l(0) - H(\eta(\underline{X}))) \mathbf{1}_{f^\star(\underline{X}) \neq \text{sign}(f(\underline{X}))}\Big] \\ &\quad \leq \mathbb{E}_{\underline{X}}\Big[(l(0) - H(\eta(\underline{X}))) \mathbf{1}_{f(\underline{X})(2\eta(\underline{X})-1)<0}\Big]\end{aligned}$$

Using the property of the wrong sign classifiers

$$\begin{aligned}&\Psi\left(\mathbb{E}\Big[\ell^{0/1}(Y, \text{sign}(f(\underline{X})))\Big] - \mathbb{E}\Big[\ell^{0/1}(Y, f^\star(\underline{X}))\Big]\right) \\ &\quad \leq \mathbb{E}_{\underline{X}}\Big[(H(f, \eta(\underline{X})) - H(f^\star, \eta(\underline{X}))) \mathbf{1}_{f(\underline{X})(2\eta(\underline{X})-1)<0}\Big] \\ &\quad \leq \mathbb{E}_{\underline{X}}[(H(f, \eta(\underline{X})) - H(f^\star, \eta(\underline{X})))] \\ &\quad \leq \mathbb{E}\Big[\bar{\ell}(Y, f(\underline{X}))\Big] - \mathbb{E}\Big[\bar{\ell}(Y, f^\star(\underline{X}))\Big]\end{aligned}$$

□

3.3. SVM, distance and norm of β

Claim 3.3.1

The distance between $\underline{X}^\top \beta + \beta^{(0)} = 1$ *and* $\underline{X}^\top \beta + \beta^{(0)} = -1$ *is given by*

$$\frac{2}{\|\beta\|}.$$

Proof. For any $\underline{X}'$, the distance between $\underline{X}'$ and the hyperplane $\underline{X}^\top \beta + \gamma = 0$ is given by

$$\frac{|\underline{X}'^\top \beta - \gamma|}{\|\beta\|}.$$

Applying this result to the hyperplane $transp\underline{X}\beta + \beta^{(0)} = 1$ and any point in the hyperplane $transp\underline{X}'\beta + \beta^{(0)} = -1$ yields the result. □

3.4. SVM and Hinge Loss

Claim 3.4.1

The two problems

$$\min \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^n s_i \quad \textit{with} \quad \begin{cases} \forall i, Y_i(\underline{X}_i^\top \beta + \beta^{(0)}) \geq 1 - s_i \\ \forall i, s_i \geq 0 \end{cases}$$

and

$$\min \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^n \underbrace{\max(0, 1 - Y_i(\underline{X}_i^\top \beta + \beta^{(0)}))}_{\textit{Hinge Loss}}$$

yields the same solution for β.

Proof. We may write

$$\begin{aligned} &\min_{\beta,s} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^n s_i \quad \text{with} \quad \begin{cases} \forall i, Y_i(\underline{X}_i^\top \beta + \beta^{(0)}) \geq 1 - s_i \\ \forall i, s_i \geq 0 \end{cases} \\ \Leftrightarrow &\min_{\beta} \min_{s} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^n s_i \quad \text{with} \quad \begin{cases} \forall i, Y_i(\underline{X}_i^\top \beta + \beta^{(0)}) \geq 1 - s_i \\ \forall i, s_i \geq 0 \end{cases} \end{aligned}$$

Now for any β,

$$\min_{s} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^n s_i \quad \text{with} \quad \begin{cases} \forall i, Y_i(\underline{X}_i^\top \beta + \beta^{(0)}) \geq 1 - s_i \\ \forall i, s_i \geq 0 \end{cases} \quad = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^n \max(0, 1 - Y_i(\underline{X}_i^\top \beta +$$

hence the result. □

3.5. Constrained Optimization, Lagrangian and Dual

Claim 3.5.1

$$\max_{\lambda\in\mathbb{R}^p,\ \mu\in(\mathbb{R}^+)^q} \mathcal{L}(x,\lambda,\mu) = \begin{cases} f(x) & \text{if } x \text{ is feasible} \\ +\infty & \text{otherwise} \end{cases}$$

$$\min_x \max_{\lambda\in\mathbb{R}^p,\ \mu\in(\mathbb{R}^+)^q} \mathcal{L}(x,\lambda,\mu) = \min_x f(x) \quad \text{with} \quad \begin{cases} h_j(x) = 0, & j = 1,\dots p \\ g_i(x) \le 0, & i = 1,\dots q \end{cases}$$

Proof. See Chapter A. □

Claim 3.5.2

$$Q(\lambda,\mu) \le f(x), \ \text{for all feasible } x$$
$$\max_{\lambda\in\mathbb{R}^p,\ \mu\in(\mathbb{R}^+)^q} Q(\lambda,\mu) \le \min_{x \text{ feasible}} f(x)$$

Proof. See Chapter A. □

3.5.1. Duality, weak, strong and Slater's condition

Claim 3.5.3

Weak duality:

$$q^\star \le p^\star$$
$$\max_{\lambda\in\mathbb{R}^p,\ \mu\in(\mathbb{R}^+)^q} \min_x \mathcal{L}(x,\lambda,\mu) \le \min_x \max_{\lambda\in\mathbb{R}^p,\ \mu\in(\mathbb{R}^+)^q} \mathcal{L}(x,\lambda,\mu)$$

Proof. See Chapter A. □

Claim 3.5.4

*If f is convex, h_j affine and g_i convex then the **Slater's condition**, it exists a feasible point such that $h_j(x) = 0$ for all j and $g_i(x) < 0$ for all i, is sufficient to imply the strong duality:*

$$\max_{\lambda\in\mathbb{R}^p,\ \mu\in(\mathbb{R}^+)^q} \min_x \mathcal{L}(x,\lambda,\mu) = \min_x \max_{\lambda\in\mathbb{R}^p,\ \mu\in(\mathbb{R}^+)^q} \mathcal{L}(x,\lambda,\mu)$$

Proof. See Chapter A. □

3.6. Karush-Kuhn-Tucker Claim

Claim 3.6.1

If f is convex, h_j affine and g_i convex, all are differentiable and strong duality holds then $x^\star$ is a solution of the primal problem if and only if the KKT condition

- *Stationarity:*

$$\nabla_x \mathcal{L}(x^\star, \lambda, \mu) = \nabla f(x^\star) + \sum_j \lambda_j \nabla h(x^\star) + \sum_i \mu_i \nabla g(x^\star) = 0$$

- *Primal admissibility:*

$$h_j(x^\star) = 0 \quad \textit{and} \quad g_i(x^\star) \leq 0$$

- *Dual admissibility:*

$$\mu_i \geq 0$$

- *Complementary slackness:*

$$\mu_i g_i(x^\star) = 0$$

holds.

Proof. See Chapter A. □

3.7. SVM, KKT and Dual

Claim 3.7.1

For the SVM, the KKT conditions are given by

- *Stationarity:*

$$\begin{aligned}
\nabla_\beta \mathcal{L}(\beta, \beta^{(0)}, s, \alpha, \mu) &= \beta - \sum \alpha_i Y_i \underline{X}_i = 0 \\
\nabla_{\beta^{(0)}} \mathcal{L}(\beta, \beta^{(0)}, s, \alpha, \mu) &= -\sum_i \alpha_i = 0 \\
\nabla_{s_i} \mathcal{L}(\beta, \beta^{(0)}, s, \alpha, \mu) &= C - \alpha_i - \mu_i = 0
\end{aligned}$$

- *Primal and dual admissibility:*

$$(1 - s_i - Y_i(\underline{X}_i{}^\top \beta + \beta^{(0)})) \leq 0, \quad s_i \geq 0, \quad \alpha_i \geq 0, \ \textit{and} \ \mu_i \geq 0$$

- *Complementary slackness:*

$$\alpha_i(1 - s_i - Y_i(\underline{X}_i{}^\top \beta + \beta^{(0)})) = 0 \quad \textit{and} \quad \mu_i s_i = 0$$

Proof. The Lagrangian of the SVM is given by

$$\mathcal{L}(\beta, \beta^{(0)}, s, \alpha, \mu) = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n} s_i + \sum_i \alpha_i(1 - s_i - Y_i(\underline{X}_i^\top \beta + \beta^{(0)})) - \sum_i \mu_i s_i.$$

We can compute the stationarity condition and obtain immediately:

$$\begin{aligned}
\nabla_\beta \mathcal{L}(\beta, \beta^{(0)}, s, \alpha, \mu) &= \beta - \sum_i \alpha_i Y_i \underline{X}_i = 0 \\
\nabla_{\beta^{(0)}} \mathcal{L}(\beta, \beta^{(0)}, s, \alpha, \mu) &= -\sum_i \alpha_i = 0 \\
\nabla_{s_i} \mathcal{L}(\beta, \beta^{(0)}, s, \alpha, \mu) &= C - \alpha_i - \mu_i = 0
\end{aligned}$$

The remaining conditions are straightforward.

□

Claim 3.7.2

The SVM problem satisfy Slater's constraints.

Proof. It suffices to verify that $\beta = 0$, $\beta^{(0)} = 0$ and $s = 2$ is a feasible vector for which the inequalities in the constraints are strict. □

Claim 3.7.3

The solution of the SVM satisfy

- $\beta^\star = \sum_i \alpha_i Y_i \underline{X}_i$ *and* $0 \leq \alpha_i \leq C$.
- *If* $\alpha_i \neq 0$, $\underline{X}_i$ *is called a support vector and either*
  - $s_i = 0$ *and* $Y_i(\underline{X}_i^\top \beta + \beta^{(0)}) = 1$ *(margin hyperplane),*
  - *or* $\alpha_i = C$ *(outliers).*
- $\beta^{(0)*} = Y_i - \underline{X}_i^\top \beta^\star$ *for any support vector with* $0 < \alpha_i < C$.

Proof. As the SVM satisfies the Slater's constraints. The optimal $\beta^\star$, $\beta^{(0)*}$, s of the primal problem and the optimal α and μ of the dual satisfy the KKT optimality condition.

The formula for $\beta^\star$ is thus a direct consequence of $\nabla_\beta \mathcal{L}(\beta, \beta^{(0)}, s, \alpha, \mu) = 0$.

If we use $\nabla_{s_i} \mathcal{L}(\beta^\star, \beta^{(0)*}, s, \alpha, \mu) = 0$, we have $\alpha_i = C - \mu_i$ which leads to $0 \leq \alpha_i \leq C$ as $\alpha_i \geq 0$ and $\mu_i \geq 0$ by the dual admissibility condition.

By the complementary slackness condition, $\alpha_i \neq 0$ implies $Y_i(\underline{X}_i^\top \beta^\star + \beta^{(0)*}) = 1 - s_i$ thus

- either $s_i = 0$ and $Y_i(\underline{X}_i^\top \beta^\star + \beta^{(0)*}) = 1$,
- or $s_i \neq 0$ which implies $c_i = 0$ and thus $\alpha_i = C$ (outliers).

For any support vector with $0 < \alpha_i < C$, $\underline{X}_i^\top \beta^\star + \beta^{(0)*} = Y_i$ hence $\beta^{(0)*} = Y_i - \underline{X}_i^\top \beta^\star$.

□

Claim 3.7.4

The dual of the SVM

$$Q(\alpha, \mu) = \min_{\beta, \beta^{(0)}, s} \mathcal{L}(\beta, \beta^{(0)}, s, \alpha, \mu)$$

is given by

- *if* $\sum_i \alpha_i Y_i \neq 0$ *or* $\exists i, \alpha_i + \mu_i \neq C$,

$$Q(\alpha, \mu) = -\infty$$

- *if* $\sum_i \alpha_i Y_i = 0$ *and* $\forall i, \alpha_i + \mu_i = C$,

$$Q(\alpha, \mu) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j Y_i Y_j \underline{X}_i^\top \underline{X}_j$$

Proof. The dual of the SVM is defined as

$$\begin{aligned} Q(\alpha, \mu) &= \min_{\beta, \beta^{(0)}, s} \mathcal{L}(\beta, \beta^{(0)}, s, \alpha, \mu) \\ &= \min_{\beta, \beta^{(0)}, s} \frac{1}{2}\|\beta\|^2 + C \sum_{i=1}^{n} s_i + \sum_i \alpha_i (1 - s_i - Y_i(\underline{X}_i^\top \beta + \beta^{(0)})) - \sum_i \mu_i s_i \\ &= \min_{\beta, \beta^{(0)}, s} \frac{1}{2}\|\beta\|^2 - \sum_i \alpha_i Y_i \underline{X}_i^\top \beta - \sum_i \alpha_i Y_i \beta^{(0)} + \sum_i (C - \alpha_i - \mu_i) s_i + \sum_i \alpha_i \end{aligned}$$

We obtain immediately that this minimum is equal to $-\infty$ as soon as $\sum_i \alpha_i Y_i \neq 0$ or $C - \alpha_i - \mu_i \neq 0$.

Assume now that $\sum_i \alpha_i Y_i = 0$ and $C - \alpha_i - \mu_i = 0$, we obtain

$$\begin{aligned} Q(\alpha, \mu) &= \min_{\beta, \beta^{(0)}, s} \frac{1}{2}\|\beta\|^2 - \sum_i \alpha_i Y_i \underline{X}_i^\top \beta + \sum_i \alpha_i \\ &= \min_{\beta} \frac{1}{2}\|\beta\|^2 - \sum_i \alpha_i Y_i \underline{X}_i^\top \beta + \sum_i \alpha_i \end{aligned}$$

The optimal β can be obtained by setting to 0 the derivative:

$$\beta - \sum_i \alpha_i Y_i \underline{X}_i^\top = 0$$

Plugging this value in the formula yields immediately

$$Q(\alpha, \mu) = -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j Y_i Y_j \underline{X}_i^\top \underline{X}_j + \sum_i \alpha_i$$

□

3.8. Mercer Representation Claim

Claim 3.8.1

For any loss ℓ and any increasing function Φ, the minimizer in β of

$$\sum_{i=1}^{n} \ell(Y_i, \underline{X}_i^\top \beta + \beta^{(0)}) + \Phi(\|\beta\|_2)$$

is a linear combination of the input points $\beta^\star = \sum_{i=1}^{n} \alpha_i' \underline{X}_i$.

Proof. Assume β is a minimizer of

$$\sum_{i=1}^{n} \ell(Y_i, \underline{X}_i^\top \beta + \beta^{(0)}) + \Phi(\|\beta\|_2)$$

and let $\beta_{\underline{X}}$ be the orthogonal projection of β on the finite dimensional space spanned by the $\underline{X}_i$. By construction $\beta - \beta_{\underline{X}}$ is orthogonal to all the $\underline{X}_i$ and thus

$$\begin{aligned} \underline{X}_i^\top \beta + \beta^{(0)} &= \underline{X}_i^\top (\beta_{\underline{X}} + \beta - \beta_{\underline{X}}) + \beta^{(0)} \\ &= \underline{X}_i^\top \beta_{\underline{X}} + \beta^{(0)} \end{aligned}$$

and thus

$$\begin{aligned} \sum_{i=1}^{n} \ell(Y_i, \underline{X}_i^\top \beta + \beta^{(0)}) + \Phi(\|\beta\|_2) &= \sum_{i=1}^{n} \ell(Y_i, \underline{X}_i^\top \beta_{\underline{X}} + \beta^{(0)}) + \Phi(\|\beta\|_2) \\ &\geq \sum_{i=1}^{n} \ell(Y_i, \underline{X}_i^\top \beta_{\underline{X}} + \beta^{(0)}) + \Phi(\|\beta_{\underline{X}}\|_2) \end{aligned}$$

where the inequality holds because $\|\beta\|^2 = \|\beta_{\underline{X}}\|^2 + \|\beta - \beta_{\underline{X}}\|^2$. The minimum is thus reached by a β in the space spanned by the $\underline{X}_i$, i.e.

$$\beta = \sum_{i=1}^{n} \alpha_i \underline{X}_i.$$

□

3.9. Moore-Aronsajn Claim

Claim 3.9.1

For any PDS kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, it exists a Hilbert space $\mathbb{H} \subset \mathbb{R}^{\mathcal{X}}$ with a scalar product $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ such that

- *it exists a mapping* $\phi : \mathcal{X} \to \mathbb{H}$ *satisfying*

$$k(\underline{X}, \underline{X}') = \langle \phi(\underline{X}), \phi(\underline{X}) \rangle_{\mathbb{H}}$$

- *the reproducing property holds, i.e. for any* $h \in \mathbb{H}$ *and any* $\underline{X} \in \mathcal{X}$

$$h(\underline{X}) = \langle h, k(\underline{X}, \cdot) \rangle_{\mathbb{H}} .$$

Proof. See Chapter D. □

3.10. Kernel Construction Machinery

Claim 3.10.1

For any function $\Psi : \mathcal{X} \to \mathbb{R}$, $k(\underline{X}, \underline{X}') = \Psi(\underline{X})\Psi(\underline{X}')$ *is PDS.*

Proof. See Chapter D. □

Claim 3.10.2

For any PDS kernels k_1 *and* k_2, *and any* $\lambda \geq 0$ $k_1 + \lambda k_2$ *and* $\lambda k_1 k_2$ *are PDS kernels.*

Proof. See Chapter D. □

Claim 3.10.3

For any sequence of PDS kernels k_n *converging pointwise to a kernel* k, k *is a PDS kernel.*

Proof. See Chapter D. □

Claim 3.10.4

For any PDS kernel k *such that* $|k| \leq r$ *and any power series* $\sum_n a_n z^n$ *with* $a_n \geq 0$ *and a convergence radius larger than* r, $\sum_n a_n k^n$ *is a PDS kernel.*

Proof. See Chapter D. □

Claim 3.10.5

For any PDS kernel k, *the renormalized kernel* $k'(\underline{X}, \underline{X}') = \dfrac{k(\underline{X}, \underline{X}')}{\sqrt{k(\underline{X}, \underline{X}) k(\underline{X}', \underline{X}')}}$ *is a PDS kernel.*

Proof. See Chapter D. □

3.11. Mercer Representation Claim

Claim 3.11.1

Let k be a PDS kernel and $\mathbb{H}$ its corresponding RKHS,
for any increasing function Φ and any function $L : \mathbb{R}^n \to \mathbb{R}$, the optimization problem

$$\operatorname*{argmin}_{h \in \mathbb{H}} L(h(\underline{X}_1), \dots, h(\underline{X}_n)) + \Phi(\|h\|)$$

admits only solutions of the form

$$\sum_{i=1}^{n} \alpha_i' k(\underline{X}_i, \cdot).$$

Proof. See Chapter D. □

3.12. SVM and VC dimension

See **mohri18** as the VC dimension will only be defined later.

4. Optimization: Gradient Descent Algorithms

Most of the results can be found in **beck17** or in **bubeck15**.

4.1. Linear Predictor, Gradient and Hessian

Claim 4.1.1

- *Gradient:*

$$\nabla F(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} \bar{\ell}(Y_i, \langle \underline{X}_i, \boldsymbol{w}\rangle)\underline{X}_i$$

with $\bar{\ell}(y, f) = \dfrac{\partial \ell(y, f)}{\partial f}$

- *Hessian matrix:*

$$\nabla^2 F(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} \bar{\ell}'(Y_i, \langle \underline{X}_i, \boldsymbol{w}\rangle)\underline{X}_i \underline{X}_i^\top$$

with $\bar{\ell}'(y, f) = \dfrac{\partial^2 \ell(y, f)}{\partial f^2}$

4.2. Exhaustive Search

Claim 4.2.1

- *If G is C-Lipschitz, evaluating G on a grid of precision $\epsilon/(\sqrt{d}C)$ is sufficient to find a ϵ-minimizer of G.*
- *Required number of evaluation:* $N_\epsilon = O\left((C\sqrt{d}/\epsilon)^d\right)$

4.3. L Smoothness

Claim 4.3.1

*If G is twice differentiable, G is L-**smooth** if and only if for all $x \in \mathbb{R}^d$,*

$$\lambda_{\max}(\nabla^2 G(x)) \leq L.$$

Proof. Fix $x, y \in \mathbb{R}^d$ and $c > 0$. Let $g(t) = \nabla G(x + tcy)$. Thus, $g'(t) = [\nabla^2 G(x + tcy)](cy)$. By the mean value theorem, there exists some constant $t_c \in [0, 1]$ such that

$$\nabla G(x + cy) - \nabla G(x) = g(1) - g(0) = g'(t_c) = [\nabla^2 G(x + t_c cy)](cy). \tag{4.1}$$

First implication
Taking the norm of both sides of eq. (4.1) and applying the smoothness condition, we obtain

$$\|[\nabla^2 G(x + t_c cy)]y\| \leq L\|y\|.$$

By taking $c \to 0$ and using the fact that $t_c \in [0, 1]$ and $G \in C^2$, we have

$$\|[\nabla^2 G(x)]y\| \leq L\|y\|.$$

Then, $\lambda_{max}(\nabla^2 G(x)) \leq L$.
Second implication
Taking the norm of both sides of eq. (4.1), we have

$$\|\nabla G(x + cy) - \nabla G(x)\|_2 = \|[\nabla^2 G(x + t_c cy)](cy)\|_2.$$

Note that, for any real-valued symmetric matrix A and any vector u,

$$\|Au\|_2^2 = u^T A^T Au = \langle A^T Au, u\rangle \leq \lambda_{max}(A)^2\|u\|^2$$

Thus,

$$\|\nabla G(x + cy) - \nabla G(x)\|_2 \leq \lambda_{max}([\nabla^2 G(x + t_c cy)])\|(cy)\|_2 \leq L\|cy\|_2.$$

□

Claim 4.3.2

F is L-smooth in the linear regression and the logistic regression cases.

4.4. Convergence of GD

Claim 4.4.1

Let $G : \mathbb{R}^d \to \mathbb{R}$ be an L-smooth convex function. Let $\boldsymbol{w}^\star$ be the minimum of f on $\mathbb{R}^d$. Then, Gradient Descent with step size $\alpha \leq 1/L$ satisfies

$$G(\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^\star) \leq \frac{\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|_2^2}{2\alpha k}.$$

Proof. This is a consequence of Corollary C.2.4. □

Claim 4.4.2

In particular, for $\alpha = 1/L$,

$$N_\epsilon = O(L\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|_2^2/(2\epsilon))$$

iterations are sufficient to get an ϵ*-approximation of the minimal value of* G.

Proof. In order to have an ϵ-minimizer, it suffices that $\frac{\|\boldsymbol{w}^{[0]}-\boldsymbol{w}^\star\|_2^2}{2\alpha k} \leq \epsilon$, i.e. $k \geq \frac{\|\boldsymbol{w}^{[0]}-\boldsymbol{w}^\star\|_2^2}{2\alpha\epsilon}$ which yields the result. □

Claim 4.4.3

If G *is convex and* L*-smooth, then for any* $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$

$$G(\boldsymbol{w}) \leq G(\boldsymbol{w}') + \nabla G(\boldsymbol{w}')^\top(\boldsymbol{w} - \boldsymbol{w}') + \frac{L}{2}\|\boldsymbol{w} - \boldsymbol{w}'\|_2^2.$$

Proof. Using the fact that

$$\begin{aligned} G(\boldsymbol{w}') &= G(\boldsymbol{w}) + \int_0^1 (\nabla G(\boldsymbol{w} + t(\boldsymbol{w}' - \boldsymbol{w})))^\top(\boldsymbol{w}' - \boldsymbol{w})dt \\ &= G(\boldsymbol{w}) + \nabla G(\boldsymbol{w})^\top(\boldsymbol{w}' - \boldsymbol{w}) \\ &\quad + \int_0^1 (\nabla G(\boldsymbol{w} + t(\boldsymbol{w}' - \boldsymbol{w})) - \nabla G(\boldsymbol{w}))^\top(\boldsymbol{w}' - \boldsymbol{w})dt, \end{aligned}$$

so that

$$\begin{aligned} &|G(\boldsymbol{w}') - G(\boldsymbol{w}) - (\nabla G(\boldsymbol{w}))^\top(\boldsymbol{w}' - \boldsymbol{w})| \\ &\quad \leq \int_0^1 |(\nabla G(\boldsymbol{w} + t(\boldsymbol{w}' - \boldsymbol{w})) - \nabla G(\boldsymbol{w}))^\top(\boldsymbol{w}' - \boldsymbol{w})dt| \\ &\quad \leq \int_0^1 \|\nabla G(\boldsymbol{w} + t(\boldsymbol{w}' - \boldsymbol{w})) - \nabla G(\boldsymbol{w})\| \, \|\boldsymbol{w}' - \boldsymbol{w}\| \, dt \\ &\quad \leq \int_0^1 Lt\|\boldsymbol{w}' - \boldsymbol{w}\|^2 dt = \frac{L}{2}\|\boldsymbol{w}' - \boldsymbol{w}\|^2. \end{aligned}$$

□

Claim 4.4.4

Let $G : \mathbb{R}^d \to \mathbb{R}$ *be an* L*-smooth,* μ *strongly convex function. Let* $\boldsymbol{w}^\star$ *be the minimum of* G *on* $\mathbb{R}^d$*. Then, Gradient Descent with step size* $\alpha \leq 1/L$ *satisfies*

$$G(\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^\star) \leq \frac{1}{2\alpha}\Big(1 - \alpha\mu\Big)^k \|G(\boldsymbol{w}^{[0]}) - G(\boldsymbol{w}^\star)\|_2^2.$$

Proof. This is a consequence of Corollary C.3.2. □

Claim 4.4.5

Let $G : \mathbb{R}^d \to \mathbb{R}$ be a convex function, C-Lipschitz in $B(\boldsymbol{w}^\star, R)$ where $\boldsymbol{w}^\star$ be the minimizer of f on $\mathbb{R}^d$. Assume that

$$\alpha^{[k]} > 0, \quad \alpha^{[k]} \to 0, \quad \sum_k \alpha^{[k]} = +\infty$$

and $\left\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\right\| \leq R$ Then, Subgradient Descent with step size $\alpha^{[k]}$ satisfies

$$\min_k G(\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^\star) \leq C \frac{R^2 + \sum_{k'=0}^k (\alpha^{[k']})^2}{2 \sum_{k'=0}^k \alpha^{[k']}}$$

Proof. This is a consequence of Theorem C.5.1. □

4.5. Proximal Descent

Claim 4.5.1

- $R(\boldsymbol{w}) = \mathbf{1}_\Omega(\boldsymbol{w})$*:* $\operatorname{prox}_\gamma R(\boldsymbol{w}') = P_\Omega(\boldsymbol{w}')$
- $R(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2$*:* $\operatorname{prox}_\gamma R(\boldsymbol{w}') = \frac{1}{1+\gamma}\boldsymbol{w}$.
- $R(\boldsymbol{w}) = \|\boldsymbol{w}\|_1$*:* $\operatorname{prox}_\gamma R(\boldsymbol{w}') = T_\gamma(\boldsymbol{w}')$ *with* $T_\gamma(\boldsymbol{w})_i = \operatorname{sign}(\boldsymbol{w}_i)\max(0, |\boldsymbol{w}_i| - \gamma)$ *(soft thresholding).*

Proof. If $R(\boldsymbol{w}) = \mathbf{1}_\Omega(\boldsymbol{w})$, then

$$\begin{aligned}
\operatorname{prox}_\gamma R(\boldsymbol{w}') &= \arg\min_{\boldsymbol{w}} \frac{1}{2\gamma}\|\boldsymbol{w} - \boldsymbol{w}'\|^2 + R(\boldsymbol{w}') \\
&= \arg\min_{\boldsymbol{w}\in\Omega} \frac{1}{2\gamma}\|\boldsymbol{w} - \boldsymbol{w}'\|^2 \\
&= P_\Omega(\boldsymbol{w}').
\end{aligned}$$

If $R(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2$ then

$$\begin{aligned}
\operatorname{prox}_\gamma R(\boldsymbol{w}') &= \arg\min_{\boldsymbol{w}} \frac{1}{2\gamma}\|\boldsymbol{w} - \boldsymbol{w}'\|^2 + R(\boldsymbol{w}') \\
&= \arg\min \frac{1}{2\gamma}\|\boldsymbol{w} - \boldsymbol{w}'\|^2 + \frac{1}{2}\|\boldsymbol{w}\|^2
\end{aligned}$$

The function minimized is smooth (and strongly convex) and its gradient is given by

$$\frac{1}{\gamma}(\boldsymbol{w} - \boldsymbol{w}') + \boldsymbol{w}$$

which is equal to 0 if and only if $\boldsymbol{w} = \frac{1}{1+\gamma}\boldsymbol{w}'$, hence the result.

If $R(\boldsymbol{w}) = \|\boldsymbol{w}\|_1$ then

$$\frac{1}{2\gamma}\|\boldsymbol{w} - \boldsymbol{w}'\|^2 + R(\boldsymbol{w}) = \sum_i^d \left(\frac{1}{2\gamma}(\boldsymbol{w}_i - \boldsymbol{w}'_i)^2 + |\boldsymbol{w}_i|\right).$$

We can analyze thus each coordinate independently. Let $f(x) = \frac{1}{2\gamma}(x - x')^2 + |x|$, this function is strongly convex and its subgradient is given by

$$\delta_f(x) = \begin{cases} \frac{1}{\gamma}(x - x') - 1 & \text{if } x < 0 \\ [\frac{1}{\gamma}(-x') - 1, \frac{1}{\gamma}(-x') + 1] & \text{if } x = 0 \\ \frac{1}{\gamma}(x - x') + 1 & \text{if } x > 0 \end{cases}$$

One verify easily that

- if $x' < -\gamma$ then $0 \in \delta_f(x)$ for $x = x' + \gamma$
- if $x' > \gamma$ then $0 \in \delta_f(x)$ for $x = x' - \gamma$
- if $-\gamma \leq x' \leq \gamma$ then $0 \in \delta_f(0)$

and thus

$$\text{prox}_{\gamma|\cdot\|}(x') = \begin{cases} x' + \gamma & \text{if } x' < -\gamma \\ 0 & \text{if } -\gamma \leq x \leq \gamma \\ x' - \gamma & \text{if } x' > \gamma \end{cases}$$

or equivalently

$$\text{prox}_{\gamma|\cdot\|}(x') = \text{sign}(x')\max(0, |x'| - \gamma)$$

□

Claim 4.5.2

- *F L-smooth and R simple:*

$$G(\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^\star) \leq \frac{\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|_2^2}{2\alpha k}.$$

and $N_\epsilon = O(L\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|_2^2/2\epsilon)$.

- *F L-smooth and μ-convex and R simple:*

$$G(\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^\star) \leq \frac{1}{2\alpha}\Big(1 - \alpha\mu\Big)^k \|G(\boldsymbol{w}^{[0]}) - G(\boldsymbol{w}^\star)\|_2^2.$$

and $N_\epsilon = O(-\log\epsilon/(\alpha\mu))$.

- *F C-Lipschitz and R is the characteristic function of a convex set:*

$$\min k' \leq kG(\boldsymbol{w}^{[k']}) - G(\boldsymbol{w}^\star) \leq C\frac{R^2 + r^2\log(k+1)}{4r\sqrt{k+1}}$$

and $N_\epsilon = O\left((C(-\log\epsilon)/\epsilon)^2\right)$.

Proof. Those are consequences of Theorem C.2.1, Theorem C.3.1 and Theorem C.5.1. □

4.6. Coordinate Descent

Claim 4.6.1

If G is continuously differentiable and strictly convex, then exact coordinate descent converges to a minimum.

Proof. The proof is quite technical and can be found in **saha13**. □

Claim 4.6.2

Assume that G is convex and smooth and that each G^i is L_i-smooth.
Consider a sequence $\{\boldsymbol{w}^{[k]}\}$ given by CGD with $\alpha^{[k]} = 1/L_{i_k}$ and coordinates $i_1, i_2, \ldots$ chosen at random: i.i.d and uniform distribution in $\{1, \ldots, d\}$. Then

$$\mathbb{E}\Big[G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^\star)\Big] \leq \frac{d}{d+k}\Big((1-\frac{1}{d})(G(\boldsymbol{w}^{[0]}) - G(\boldsymbol{w}^\star)) + \frac{1}{2}\left\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\right\|_L^2\Big),$$

with $\|\boldsymbol{w}\|_L^2 = \sum_{j=1}^d L_j \boldsymbol{w}_j^2$.

Proof. The proof is quite technical and can be found in **nesterov12**. □

4.7. Gradient Descent Acceleration

Claim 4.7.1

Assume that G is an L-smooth, convex function whose minimum is reached at $\boldsymbol{w}^\star$. Then, if $\beta^{[k]} = (k-1)/(k+2)$,

$$G(\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^\star) \leq \frac{2\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|_2^2}{\alpha(k+1)^2}.$$

Proof. See Corollary C.4.2. □

Claim 4.7.2

Assume that G is an L-smooth, μ strongly convex function whose minimum is reached at $\boldsymbol{w}^\star$. Then, if $\beta^{[k]} = \frac{1-\sqrt{\mu/L}}{1+\sqrt{\mu/L}}$,

$$G(\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^\star) \leq \frac{\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|_2^2}{\alpha}\Big(1 - \sqrt{\frac{\mu}{L}}\Big)^k.$$

Proof. The proof combines ideas of Theorem C.3.1 and Corollary C.4.2. It is left as an exercise or can be found in **beck17**. □

Claim 4.7.3

- *For any $\boldsymbol{w}^{[0]} \in \mathbb{R}^d$ and any k satisfying $1 \leq k \leq (d-1)/2$, there exists an L-smooth convex function f such that for any general first order method*

$$G(\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^\star) \geq \frac{3L\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|_2^2}{32(k+1)^2}.$$

- *For any $\boldsymbol{w}^{[0]} \in \mathbb{R}^d$ and any $k \leq (d-1)/2$, there exists an L-smooth, μ strongly convex function f such that for any general first order method*

$$G(\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^\star) \geq \frac{\mu}{2}\Big(\frac{1-\sqrt{\mu/L}}{1+\sqrt{\mu/L}}\Big)^{2k}\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|_2^2.$$

Proof. The proof is quite technical and can be found in **nesterov18**. □

4.8. Stochastic Gradient Descent

Claim 4.8.1

- *With $\alpha^{[k]} = 2R/(b\sqrt{k})$*

$$\mathbb{E}\left[G\Big(\frac{1}{k}\sum_{j=1}^{k}\boldsymbol{w}^{[j]}\Big)\right] - G(\boldsymbol{w}^\star) \leq \frac{3rb}{\sqrt{k}}$$

- *If G is μ-strongly convex then with $\alpha^{[k]} = 2/(\mu(k+1))$,*

$$\mathbb{E}\left[G\Big(\frac{2}{k(k+1)}\sum_{j=1}^{k} j\boldsymbol{w}^{[j]}\Big)\right] - G(\boldsymbol{w}^\star) \leq \frac{2b^2}{\mu(k+1)}.$$

Proof. Those are consequences of Theorem C.6.1. □

5. ML Methods: Neural Networks and Deep Learning

5.1. Universal Approximation Theorem

Claim 5.1.1

A single hidden layer neural network with a linear output unit can approximate any continuous function arbitrarily well, given enough hidden units.

Proof. This a consequence of Lemma E.2.2, Lemma E.2.3 and Lemma E.2.4 provided we assume that the input space is compact. □

5.2. NN and Bias-Variance Tradeoff

Claim 5.2.1

It is not always necessary to trade bias for variance when increasing model complexity.

Proof. Above the interpolation treshold, there are more and more *optimal* solutions. The one obtained by a (regularized) optimization algorithm is smoother and smoother when the capacity of the approximation set increases. Hence, the variance may be reduced simultaneously than the bias for high complexity models. Furthermore, Stochastic Gradient Algorithms appear to implicitly perform a regularization similar to the one of Tikhonov.

See **neal18** and **neal20** for more details.

□

6. ML Methods: Trees and Ensemble Methods

6.1. AdaBoost

Claim 6.1.1

The AdaBoost algorithm and the Exponential Stagewise Additive Modeling algoritm lead to exactly the same steps.

Proof. Denoting $f_t = \sum_{t'=1}^{t} \alpha_{t'} h_{t'}$,

$$\begin{aligned}\sum_{i=1}^{n} e^{-y_i(f_{t-1}(\underline{x}_i)+\alpha h)} &= \sum_{i=1}^{n} e^{-y_i f_{t-1}(\underline{x}_i)} e^{-\alpha y_i h(\underline{x}_i)} \\ &= \sum_{i=1}^{n} w'_{t,i} e^{-\alpha y_i h(\underline{x}_i)} \\ &= (e^{\alpha} - e^{-\alpha}) \sum_{i=1}^{n} w'_{t,i} \ell^{0/1}(y_i, h(\underline{x}_i)) \\ &\quad + e^{-\alpha} \sum_{i=1}^{n} w'_{t,i}\end{aligned}$$

The minimizer h_t in h is independent of α and is also the minimizer of

$$\sum_{i=1}^{n} w'_{t,i} \ell^{0/1}(y_i, h(\underline{x}_i))$$

The optimal α_t is then given by

$$\alpha_t = \frac{1}{2} \log \frac{1 - \epsilon'_t}{\epsilon'_t}$$

with $\epsilon'_t = (\sum_{i=1}^{n} w'_{t,i} \ell^{0/1}(y_i, h_t(\underline{x}_i)))/(\sum_{i=1}^{n} w'_{t,i})$

One verify then by recursion that

$$w_{t,i} = w'_{t,i} / (\sum_{i=1}^{n} w'_{t,i})$$

and thus the two procedures are equivalent! □

7. Unsupervised Learning: Dimension Reduction

7.1. High Dimensional Geometry

Claim 7.1.1

If $\underline{X}_1, \dots, \underline{X}_n$ in the hypercube of dimension d such that their coordinates are i.i.d then

$$d^{-1/p}\left(\max \|\underline{X}_i - \underline{X}_j\|_p - \min \|\underline{X}_i - \underline{X}_j\|_p\right) = 0 + O_P\left(\sqrt{\frac{\log n}{d}}\right)$$

$$\frac{\min \|\underline{X}_i - \underline{X}_j\|_p}{\max \|\underline{X}_i - \underline{X}_j\|_p} = 1 + O_P\left(\sqrt{\frac{\log n}{d}}\right).$$

Proof. By construction,

$$\|X_i - X_j\|_p^p = \sum_{l=1}^{d} (X_i^{(l)} - X_j^{(l)})^p$$

As the coordinates are independent and bounded in $[0,1]$ so are the differences at the power p.

Using Hoeffding inequality leads thus to

$$\mathbb{P}\Big(\Big|\|X_i - X_j\|_p^p - \mathbb{E}\Big[\|X_i - X_j\|_p^p\Big]\Big| > \epsilon\Big) \leq 2e^{\frac{-2\epsilon^2}{d}}$$

By a simple union bound,

$$\mathbb{P}\Big(\exists i, j, \quad \Big|\|X_i - X_j\|_p^p - \mathbb{E}\Big[\|X_i - X_j\|_p^p\Big]\Big| > \epsilon\Big) \leq 2n^2 e^{\frac{-2\epsilon^2}{d}}$$

or

$$\mathbb{P}\Big(\max_{i,j} \|X_i - X_j\|_p \leq (\mathbb{E}\Big[\|X_i - X_j\|_p^p\Big] + \epsilon)^{1/p} \text{ and } \min_{i,j} \|X_i - X_j\|_p \geq \max(0, \mathbb{E}\Big[\|X_i - X_j\|_p^p\Big] - \epsilon)^{1/p}\Big) \leq 1 - 2$$

We let $\epsilon = \lambda d \mathbb{E}\Big[\frac{1}{d}\|X_i - X_j\|_p^p\Big]\sqrt{\frac{\log n}{d}}$ so that denoting $E_p = \mathbb{E}\Big[\frac{1}{d}\|X_i - X_j\|_p^p\Big]$ which is independent of d

$$\mathbb{P}\left(\max_{i,j} \frac{\|X_i - X_j\|_p}{d^{1/p}} \leq E_p^{1/p}\left(1 + \lambda\sqrt{\frac{\log n}{d}}\right)^{1/p} \text{ and } \min_{i,j} \frac{\|X_i - X_j\|_p}{d^{1/p}} \geq E_p^{1/p} \max\left(0, 1 - \lambda\sqrt{\frac{\log n}{d}}\right)^{1/p}\right) \leq 1 - 2n^{2-\lambda^2 E_p^2}$$

Finally using $(1+x)^{1/p} \leq (1+x/p)$ for $x \geq 0$ and $\max(0,1-x)^{1/p} \geq 0$ if $x > 1/2$ and $\max(0,1-x)^{1/p} \geq (1-2x/p)$ si $x \leq 1/2$.

$$\mathbb{P}\left(\max_{i,j} \frac{\|X_i - X_j\|_p}{d^{1/p}} \leq E_p^{1/p}\left(1+\frac{\lambda}{p}\sqrt{\frac{\log n}{d}}\right) \min_{i,j} \frac{\|X_i - X_j\|_p}{d^{1/p}} \leq E_p^{1/p} \begin{cases} 0 & \text{if } \frac{\lambda}{p}\sqrt{\frac{\log n}{d}} > 1/2 \\ 1 - 2\frac{\lambda}{p}\sqrt{\frac{\log n}{d}} & \text{otherwise} \end{cases}\right) \leq 1 - 2n^{2-\lambda^2 E_p^2}$$

This implies immediately that

$$\mathbb{P}\left(\max_{i,j} \frac{\|X_i - X_j\|_p}{d^{1/p}} - \min_{i,j} \frac{\|X_i - X_j\|_p}{d^{1/p}} \leq E_p^{1/p} \begin{cases} 1 + \frac{\lambda}{p}\sqrt{\frac{\log n}{d}} & \text{if } \frac{\lambda}{p}\sqrt{\frac{\log n}{d}} > 1/2 \\ 3\frac{\lambda}{p}\sqrt{\frac{\log n}{d}} & \text{otherwise} \end{cases}\right) \leq 1 - 2n^{2-\lambda^2 E_p^2}$$

and thus

$$\mathbb{P}\left(\max_{i,j} \frac{\|X_i - X_j\|_p}{d^{1/p}} - \min_{i,j} \frac{\|X_i - X_j\|_p}{d^{1/p}} \leq 3E_p^{1/p}\frac{\lambda}{p}\sqrt{\frac{\log n}{d}}\right) \leq 1 - 2n^{2-\lambda^2 E_p^2}$$

which corresponds to the first result.

Along the same line, using $(1-2x)/(1+x) \geq 1-3x$ for $x < 1/3$, we deduce

$$\mathbb{P}\left(\frac{\min_{i,j}\|X_i - X_j\|_p}{\max_{i,j}\|X_i - X_j\|_p} \geq \begin{cases} 0 & \text{if } \frac{\lambda}{p}\sqrt{\frac{\log n}{d}} > 1/3 \\ 1 - 3\frac{\lambda}{p}\sqrt{\frac{\log n}{d}} & \text{otherwise} \end{cases}\right) \leq 1 - 2n^{2-\lambda^2 E_p^2}$$

and thus

$$\mathbb{P}\left(\frac{\min_{i,j}\|X_i - X_j\|_p}{\max_{i,j}\|X_i - X_j\|_p} \geq 1 - 3\frac{\lambda}{p}\sqrt{\frac{\log n}{d}}\right) \leq 1 - 2n^{2-\lambda^2 E_p^2}$$

□

7.2. PCA

Claim 7.2.1

The inertia satisfies

$$I = \frac{1}{2n^2}\sum_{i,j}\|\underline{X}_i - \underline{X}_j\|^2 = \frac{1}{n}\sum_{i=1}^{n}\|\underline{X}_i - m\|^2$$

Proof.

$$\begin{aligned}
\frac{1}{2n^2}\sum_{i,j}\|\underline{X}_i - \underline{X}_j\|^2 &= \frac{1}{2n^2}\sum_{i,j}\|\underline{X}_i - m + m\underline{X}_j\|^2 \\
&= \frac{1}{2n^2}\sum_{i,j}\left(\|\underline{X}_i - m\|^2 + \|\underline{X}_j - m\|^2 - 2\left\langle \underline{X}_i - m, \underline{X}_j + m\right\rangle\right) \\
&= \frac{1}{2n}\sum_{i}\|\underline{X}_i - m\|^2 + \frac{1}{2n}\sum_{j}\|\underline{X}_j - m\|^2 - \frac{2}{n^2}\sum_{i,j}\left\langle \underline{X}_i - m, \underline{X}_j + m\right\rangle
\end{aligned}$$

and thus as $\sum_i \underline{X}_i - m = 0$

$$= \frac{1}{n} \sum_i \|\underline{X}_i - m\|^2$$

□

Claim 7.2.2

The solution of rank d' of

$$\max_P \sum_{i,j} \frac{1}{2n^2} \|P\underline{X}_i - P\underline{X}_j\|^2 = \max_P \frac{1}{n} \sum_i \|P\underline{X}_i - m\|^2$$

is given by the projection on the d' largest eigenvalues of Σ.

Claim 7.2.3

The solution of rank d' of

$$\min_P \sum_i \frac{1}{n} \|\underline{X}_i - (P(\underline{X}_i - m) + m)\|^2 = \min_P \frac{1}{n} \sum_i \|(I - P)(\underline{X}_i - m)\|^2$$

is given by the projection on the d' largest eigenvalues of Σ.

Proof. By Pythagora's theorem,

$$\sum_i \|\underline{X}_i - m\|^2 = \sum_i \left(\|P(\underline{X}_i - m)\|^2 + \|(I - P)(\underline{X}_i - m)\|^2 \right)$$

□

Claim 7.2.4

The solution of rank d' of

$$\min_P \sum_{i,j} |(\underline{X}_i - m)^\top (\underline{X}_j - m) - (\Phi(\underline{X}_i) - m)^\top (\Phi(\underline{X}_j) - m)|^2$$

is given by the projection on the d' largest eigenvalues of Σ.

7.3. SVD

Claim 7.3.1

Any matrix $n \times d$ matrix A can de decomposed as

$$\underset{(n\times d)}{\boxed{\mathbf{A}}} = \underset{(n\times n)}{\boxed{\mathbf{U}}} \underset{(n\times d)}{\boxed{D}} \underset{(d\times d)}{\boxed{\mathbf{W}^\top}}$$

with U and W two orthonormal matrices and D a diagonal *matrix with decreasing values.*

Claim 7.3.2

The best low rank approximation or rank r is obtained by restriction of the matrices to the first r dimensions:

$$\underset{(n\times d)}{\boxed{\mathbf{A}}} \simeq \underset{(n\times r)}{\boxed{\mathbf{U_r}}} \underset{(r\times r)}{\boxed{D_{r,r}}} \underset{(r\times d)}{\boxed{\mathbf{W_r}^\top}}$$

for both the operator norm and the Frobenius norm!

7.4. Multiple Factor Analysis

See **husson17** for instance.

Claim 7.4.1

The proposed coding corresponds to a χ^2 type distance.

7.5. Random Projection

Claim 7.5.1

If $\underline{X}'$ is obtained by a random projection on a space of dimension d' then if $\underline{X}$ lives in a space of dimension d'', and as soon as $d' \sim d'' \log(d'')$

$$\|\underline{X}_i - \underline{X}_j\|^2 \sim \frac{d}{d'}\|\underline{X}'_i - \underline{X}'_j\|^2$$

Proof. This is the Johnson-Lindenstrauss Lemma proved for instance in **shalev-shwartz14**. □

7.6. Graph Based Approach

Claim 7.6.1

To the points $\underline{X}'_i \in \mathbb{R}^{d'}$ minimizing

$$\frac{1}{n}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}\|\underline{X}'_i - \underline{X}'_j\|^2$$

it suffices to find the d' eigenvectors with smallest eigenvalues of the Laplacian of the graph $D - W$, where D is a diagonal matrix with $D_{i,i} = \sum_j w_{i,j}$.

7.7. Word Vectors

Claim 7.7.1

The supervised approach of Word2vec yields a representation similar to the one obtained by the explicit factorization of $-\log(\mathbb{P}(w,c)/(\mathbb{P}(w)\,\mathbb{P}(c)))$

Proof. See **levy14**. □

8. Unsupervised Learning: Clustering

8.1. k-means

Claim 8.1.1

The k-means algorithm converges in a finite number of steps.

8.2. EM Algorithm

8.3. GAN

9. Unsupervised Learning: Generative Modeling

9.1. Evidence Lower BOund

Claim 9.1.1

for any $P(X, X')$ and any $R(X'|X)$:

$$\log p(X) = \mathbb{E}_R[\log p(X, X') - \log r(X'|X)] - \mathrm{KL}\left(R(X'|X), P(X'|X)\right)$$

Proof. By construction,

$$\log p(X) = \mathbb{E}_R[\log p(X)]$$

and using $p(X, X') = p(X'|X)p(X)$

$$\begin{aligned}
&= \mathbb{E}_R[\log p(X, X') - \log p(X'|X)] \\
&= \mathbb{E}_R[\log p(X, X') - \log r(X'|X)] + \mathbb{E}_R[\log r(X'|X) - \log p(X'|X)] \\
&= \mathbb{E}_R[\log p(X, X') - \log r(X'|X)] - \mathrm{KL}\left(R(X'|X), P(X'|X)\right)
\end{aligned}$$

□

9.2. Reparametrization trick

Claim 9.2.1

For any collection of law P_θ If it exists a differentiable F_θ such that for any θ, $X \sim P_\theta$ implies $X \sim F_\theta(\omega)$ with ω follows a given law then for any (differentiable) G_θ

$$\nabla_\theta \mathbb{E}_{P_\theta}[G_\theta(X)] = \mathbb{E}_\omega[\nabla_\theta \left(G_\theta \circ F_\theta\right)(\omega)]$$

Proof. By construction,

$$\mathbb{E}_{P_\theta}[G_\theta(X)] = \mathbb{E}_\omega[G_\theta(F_\theta(\omega))] = \mathbb{E}_\omega[G_\theta \circ F_\theta(\omega)]$$

As the law of ω is independent of θ, we can differentiate inside the expectation to obtain:

$$\begin{aligned}
&\nabla_\theta \mathbb{E}_{P_\theta}[G_\theta(X)] \\
&= \mathbb{E}_\omega[\nabla_\theta \left(G_\theta \circ F_\theta\right)(\omega)]
\end{aligned}$$

□

9.3. MCMC and Langevin

9.4. EBM and Estimation

9.5. Diffusion

Claim 9.5.1

The reverse SDE of

$$dX(t) = \alpha(X(t), t)X(t)dt + \sqrt{2\beta(t)}dB(t)$$

is given by

$$dX(t) = \left(-2\beta(t)\nabla_X \log P(X, t) - \alpha(X(t), t)\right)\overline{dt} + \sqrt{2\beta(t)}\overline{dB}(t)$$

Proof. The original proof can be found in **anderson82**. It is quite technical so that we will only provide a heuristic derivation inspired by **thiery23**.

The forward SDE can be interpreted as meaning that in the limit for $\delta_t \ll 1$,

$$\mathbb{P}_t(X(t+\delta_t)|X(t)) \propto \exp\left(-\frac{\|X(t+\delta_t) - (X(t) + \alpha(X(t), t)\delta_t)\|^2}{4\beta(t)\delta t}\right)$$

Now, by Bayes Rules,

$$\mathbb{P}_t(X(t)|X(t+\delta_t)) \propto \mathbb{P}_t(X(t))\,\mathbb{P}_t(X(t+\delta_t)|X(t))$$

To deal with $\mathbb{P}_t(X(t))$, we can use

$$\log \mathbb{P}_t(X(t)) = \log \mathbb{P}_t(X(t+\delta_t)) + \langle \log \mathbb{P}_t(X_{t+\delta_t}), X_t - X(t+\delta_t)\rangle + o(\delta_t)$$

so that

$$\begin{aligned}\mathbb{P}_t(X(t)|X(t+\delta_t)) &\propto \exp\left(-\langle \nabla_X \log \mathbb{P}_t(X_{t+\delta_t}), X(t+\delta_t) - X(t)\rangle\right) \\ &\quad \times \exp\left(\frac{-\|X(t+\delta_t) - (X(t) + \alpha(X(t), t)\delta_t)\|^2}{4\beta(t)\delta t}\right) + o(\delta_t) \\ &\propto \exp\left(\frac{-\|X(t+\delta_t) - (X(t) + \alpha(X(t), t)\delta_t + 2\beta(X(t), t)\nabla_X \log \mathbb{P}_t(X_{t+\delta_t})\,\delta_t)\|^2}{4\beta(t)\delta t}\right) +\end{aligned}$$

Finaly using $\alpha(X(t+\delta_t), \delta_t) = \alpha(X(t), t) + o(1)$ and $\beta(t+\delta_t) = \beta(t) + o(1)$, we deduce

$$\begin{aligned}&\mathbb{P}_t(X(t)|X(t+\delta_t)) \\ &\quad \propto \exp\left(\frac{-\|X(t) - (X(t+\delta_t) - (\alpha(X(t+\delta_t), t+\delta_t) + 2\beta(X(t+\delta_t), t+\delta_t)\nabla_X \log \mathbb{P}_t(X_{t+\delta_t}))}{4\beta(X(t+\delta_t)\delta t}\right. \\ &\qquad \left. + o(\delta_t)\right.\end{aligned}$$

which corresponds to the reverse SDE. □

9.6. GAN

10. Statistical Learning: PAC-Bayesian Approach and Complexity Theory

10.1. Hoeffding and Finite Class

Claim 10.1.1

Let Z_i be a sequence of independent centered r.v. supported in $[a_i, b_i]$ then

$$\mathbb{P}\left(\sum_{i=1}^{n} Z_i \geq \epsilon\right) \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}}$$

Proof. This is Theorem F.1.1 proved using

Claim 10.1.2

$$\mathbb{E}\left[e^{\lambda \sum_{i=1}^{n} Z_i}\right] \leq e^{\frac{\lambda^2 \sum_{i=1}^{n}(b_i-a_i)^2}{8}}.$$

proved itself as Lemma F.1.2. □

Claim 10.1.3

If $\mathcal{S}$ is finite of cardinality $|\mathcal{S}|$,

$$\mathbb{P}\left(\sup_f \left(\mathcal{R}(f) - \mathcal{R}_n(f)\right) \leq \sqrt{\frac{\log|\mathcal{S}| + \log(1/\delta)}{2n}}\right) \geq 1 - \delta$$

$$\mathbb{P}\left(\sup_f |\mathcal{R}_n(f) - \mathcal{R}(f)| \leq \sqrt{\frac{\log|\mathcal{S}| + \log(1/\delta)}{2n}}\right) \geq 1 - 2\delta$$

Proof. Using Hoeffding, $\forall f \in \mathcal{S}$,

$$\mathbb{P}\left(\left(\mathcal{R}(f) - \mathcal{R}_n(f)\right) \geq \sqrt{\frac{\log|\mathcal{S}| + \log(1/\delta)}{2n}}\right) \leq e^{-2n\left(\frac{\log|\mathcal{S}| + \log(1/\delta)}{2n}\right)} = \frac{\delta}{|\mathcal{S}|}$$

We also have

$$\mathbb{P}\left((\mathcal{R}_n(f_{\mathcal{S}}^\star) - \mathcal{R}(f_{\mathcal{S}}^\star)) \geq \sqrt{\frac{\log(1/\delta)}{2n}}\right) \leq e^{-2n\left(\frac{\log(1/\delta)}{2n}\right)} = \delta$$

Using a union bound strategy

$$\begin{aligned}
&\mathbb{P}\left(\sup_f (\mathcal{R}(f) - \mathcal{R}_n(f)) \geq \sqrt{\frac{\log|\mathcal{S}| + \log(1/\delta)}{2n}}\right) \\
&\leq \sum_{f\in\mathcal{S}} \mathbb{P}\left((\mathcal{R}_n(f) - \mathcal{R}(f)) \geq \sqrt{\frac{\log|\mathcal{S}| + \log(1/\delta)}{2n}}\right) \\
&\leq \sum_{f\in\mathcal{S}} \frac{\delta}{|\mathcal{S}|} = \delta
\end{aligned}$$

Now

$$\begin{aligned}
&\mathbb{P}\left(\sup_f (\mathcal{R}(f) - \mathcal{R}_n(f)) + (\mathcal{R}_n(f_{\mathcal{S}}^\star) - \mathcal{R}(f_{\mathcal{S}}^\star)) \geq \sqrt{\frac{\log|\mathcal{S}| + \log(1/\delta)}{2n}} + \sqrt{\frac{\log(1/\delta)}{2n}}\right) \\
&\quad\leq \mathbb{P}\left(\sup_f (\mathcal{R}(f) - \mathcal{R}_n(f)) \geq \sqrt{\frac{\log|\mathcal{S}| + \log(1/\delta)}{2n}}\right) \\
&\quad\quad + \mathbb{P}\left((\mathcal{R}_n(f_{\mathcal{S}}^\star) - \mathcal{R}(f_{\mathcal{S}}^\star)) \geq \sqrt{\frac{\log(1/\delta)}{2n}}\right) \leq \delta + \delta = 2\delta
\end{aligned}$$

The second bound can be obtained directly by bounding $2\sup_f |\mathcal{R}(f) - \mathcal{R}_n(f)|$. □

10.2. McDiarmid and Rademacher Complexity

Claim 10.2.1

If g is a bounded difference function and $\underline{X}_i$ are independent random variables then

$$\mathbb{P}(g(\underline{X}_1, \ldots, \underline{X}_n) - \mathbb{E}[g(\underline{X}_1, \ldots, \underline{X}_n)] \geq \epsilon) \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}}$$

$$\mathbb{P}(\mathbb{E}[g(\underline{X}_1, \ldots, \underline{X}_n)] - g(\underline{X}_1, \ldots, \underline{X}_n) \geq \epsilon) \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}}$$

Proof. See Theorem F.2.1. □

Claim 10.2.2

Let σ_i be a sequence of i.i.d. random symmetric Bernoulli variables (Rademacher variables):

$$\mathbb{E}\left[\sup_{f\in\mathcal{S}}\left(\mathcal{R}(f)-\mathcal{R}_n(f)\right)\right]\leq 2\mathbb{E}\left[\sup_{f\in\mathcal{S}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell^{0/1}(Y_i,f(\underline{X}_i))\right]$$

Proof. See Theorem F.2.2 with $h_i(Z_i)=\ell^{0/1}(Y_i,f(\underline{X}_i))$. □

Claim 10.2.3

If B is finite and such that $\forall b\in B,\frac{1}{n}\|b\|_2^2\leq M^2$, then

$$R_n(B)=\mathbb{E}\left[\sup_{b\in B}\frac{1}{n}\sum_{i=1}^{n}\sigma_i b_i\right]\leq\sqrt{\frac{2M^2\log|B|}{n}}$$

Proof. See Theorem F.2.3. □

Claim 10.2.4

- *With probability greater than $1-2\delta$,*

$$\mathcal{R}(\widehat{f})-\mathcal{R}(f_{\mathcal{S}}^{\star})\leq\mathbb{E}\left[\sqrt{\frac{8\log|B_n(\mathcal{S})|}{n}}\right]+\sqrt{\frac{2\log(1/\delta)}{n}}$$

- *With probability greater than $1-\delta$, simultaneously $\forall f'\in\mathcal{S}$*

$$\mathcal{R}(f')\leq\mathcal{R}_n(f')+\mathbb{E}\left[\sqrt{\frac{8\log|B_n(\mathcal{S})|}{n}}\right]+\sqrt{\frac{\log(1/\delta)}{2n}}$$

Proof. The second bound is straightforward.

For the first one,

$$\begin{aligned}\mathbb{P}\left(\sup_f\left(\mathcal{R}(f)-\mathcal{R}_n(f)\right)+\mathcal{R}_n(f_{\mathcal{S}}^{\star})-\mathcal{R}(f_{\mathcal{S}}^{\star})\geq\mathbb{E}\left[\sqrt{\frac{8\log|B_n(\mathcal{S})|}{n}}\right]+\sqrt{\frac{2\log(1/\delta)}{n}}\right)\\ \leq\mathbb{P}\left(\sup_f\left(\mathcal{R}(f)-\mathcal{R}_n(f)\right)\geq\mathbb{E}\left[\sqrt{\frac{8\log|B_n(\mathcal{S})|}{n}}\right]+\sqrt{\frac{\log(1/\delta)}{2n}}\right)\\ +\mathbb{P}\left(\left(\mathcal{R}_n(f_{\mathcal{S}}^{\star})-\mathcal{R}(f_{\mathcal{S}}^{\star})\right)\geq\sqrt{\frac{\log(1/\delta)}{2n}}\right)\leq\delta+\delta=2\delta\end{aligned}$$

□

Claim 10.2.5

- *If $\mathcal{S}$ is finite then with probability greater than $1-2\delta$*

$$\mathcal{R}(\widehat{f}) - \mathcal{R}(f_{\mathcal{S}}^{\star}) \leq \sqrt{\frac{8\log|\mathcal{S}|}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}}$$

- *If $\mathcal{S}$ is finite then with probability greater than $1-\delta$, simultaneously $\forall f' \in \mathcal{S}$*

$$\mathcal{R}(f') \leq \mathcal{R}_n(f') + \sqrt{\frac{8\log|\mathcal{S}|}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

Proof. It suffices to notice that

$$|B_n(\mathcal{S})| = |\{(\ell^{0/1}(Y_i, f(\underline{X}_i)))_{i=1}^n, f \in \mathcal{S}\}| \leq |\mathcal{S}|$$

□

10.3. VC Dimension

Claim 10.3.1

If span($\mathcal{S}$) corresponds to the sign of functions in a linear space of dimension d then $d_{VC} \leq d$.

Claim 10.3.2

If the VC dimension d_{VC} of $\mathcal{S}$ is finite

$$s(\mathcal{S}, n) \leq \begin{cases} 2^n & \text{if } n \leq d_{VC} \\ \left(\frac{en}{d_{VC}}\right)^{d_{VC}} & \text{if } n > d_{VC} \end{cases}$$

Proof (adapted from **shalev-shwartz14***).*
Let $B(\mathcal{S}, C = (c_1, \cdots, c_n)) = \{(f(c_1), \cdots, f(c_n)), f \in \mathcal{S}\}$,

$$|B(\mathcal{S}, C)| = |\{(\ell^{0/1}(Y_i, f(c_i)))_{i=1}^n, Y_i \in \{-1, 1\}, f \in \mathcal{S}\}|$$

so that $s(\mathcal{S}, n) = \sup_{C, |C|=n} |B(\mathcal{S}, C)|$.

We say that $\mathcal{S}$ shatters B if and only if $|B(\mathcal{S}, B)| = 2^B$ or $B = \emptyset$.

Claim 10.3.3

$|B(\mathcal{S}, C)| \leq |\{A \subset C, \mathcal{S} \text{ shatters } A\}|$

The VC dimension d is the largest size of a shattered set so that

$$|\{A \subset C, \mathcal{S} \text{ shatters } A\}| \leq \sum_{i=0}^{\min(n,d)} \binom{n}{i}$$

If $n \leq d$, $\sum_{i=0}^{n} \binom{n}{i} = 2^n$.
If $n \geq d$,

$$\left(\frac{d}{n}\right)^d \sum_{i=0}^{d} \binom{n}{i} \leq \sum_{i=0}^{d} \binom{n}{i} \left(\frac{d}{n}\right)^i \leq \sum_{i=0}^{n} \binom{n}{i} \left(\frac{d}{n}\right)^i = \left(1 + \frac{d}{n}\right)^n \leq e^d$$

□

Proof of Claim 10.3.3. The lemma will be proved by induction on the size of C.
If $|C| = 1$ then

- either $\mathcal{S}$ shatters $\{c_1\}$ and $|B(\mathcal{S}, C)| = 2 \leq |\{\emptyset, C\}| = |\{A \subset C, \mathcal{S} \text{ shatters } A\}|$,
- or $\mathcal{S}$ does not shatter $\{c_1\}$ and $|B(\mathcal{S}, C)| = 1 \leq |\{\emptyset\}| = |\{A \subset C, \mathcal{S} \text{ shatters } A\}|$

Now assume the property is true for all C' of size smaller than $n - 1$.
Let $C = \{c_1, \ldots, c_n\}$, by definition,

$$\begin{aligned}
|B(\mathcal{S}, C)| &= |\{(f(c_1), \ldots, f(c_n)), f \in \mathcal{S}\}| \\
&= |\{(-1, f(c_2), \ldots, f(c_n)), f \in \mathcal{S}, f(c_1) = -1\}| \\
&\quad + |\{(1, f(c_2), \ldots, f(c_n)), f \in \mathcal{S}, f(c_1) = 1\}| \\
&= |\{(f(c_2), \ldots, f(c_n)), f \in \mathcal{S}\}| \\
&\quad + |\{(f(c_2), \ldots, f(c_n)), f \in \mathcal{S}, \exists f', f(c_1) = -f'(c_1), f(c_i) = f'(c_i), i \neq 1\}|
\end{aligned}$$

Now by construction and induction,

$$\begin{aligned}
|\{(f(c_2), \ldots, f(c_n)), f \in \mathcal{S}\}| &= |B(\mathcal{S}, C \setminus \{c_1\})| \\
&\leq |\{A \subset C \setminus \{c_1\}, \mathcal{S} \text{ shatters } A\}| \\
&\leq |\{A \subset C, c_1 \notin A, \mathcal{S} \text{ shatters } A\}|
\end{aligned}$$

Now let $\mathcal{S}' = \{f \in \mathcal{S}, \exists f', f(c_1) = -f'(c_1), f(c_i) = f'(c_i), i \neq 1\}$

$$\begin{aligned}
&|\{(f(c_2), \ldots, f(c_n)), f \in \mathcal{S}, \exists f', f(c_1) = -f'(c_1), f(c_i) = f'(c_i), i \neq 1\}| \\
&\quad = |B(\mathcal{S}', C \setminus \{c_1\})| \leq |\{A \subset C \setminus \{c_1\}, \mathcal{S}' \text{ shatters } A\}| \\
&\quad \leq |\{A \subset C \setminus \{c_1\}, \mathcal{S}' \text{ shatters } A \cup \{c_1\}\}| \quad \text{by definition of } \mathcal{S}' \\
&\quad \leq |\{A \subset C, c_1 \in A, \mathcal{S}' \text{ shatters } A\}| \leq |\{A \subset C, c_1 \in A, \mathcal{S} \text{ shatters } A\}|
\end{aligned}$$

Summing those two bounds yields the lemma property for C:

$$\begin{aligned}
|B(\mathcal{S}, C)| &\leq |\{A \subset C, c_1 \notin A, \mathcal{S} \text{ shatters } A\}| + |\{A \subset C, c_1 \in A, \mathcal{S} \text{ shatters } A\}| \\
&\leq |\{A \subset C, \mathcal{S} \text{ shatters } A\}|
\end{aligned}$$

□

Claim 10.3.4
$\log s(\mathcal{S}, n) \leq d_{VC} \log\left(\frac{en}{d_{VC}}\right)$ *if* $n > d_{VC}$.

Proof. Straightforward... □

Claim 10.3.5

- *If* $\mathcal{S}$ *is of VC dimension* d_{VC} *then if* $n > d_{VC}$
- *With probability greater than* $1 - 2\delta$,

$$\mathcal{R}(\widehat{f}) - \mathcal{R}(f_{\mathcal{S}}^{\star}) \leq \sqrt{\frac{8 d_{VC} \log\left(\frac{en}{d_{VC}}\right)}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}}$$

- *With probability greater than* $1 - \delta$, *simultaneously* $\forall f' \in \mathcal{S}$,

$$\mathcal{R}(f') \leq \mathcal{R}_n(f') + \sqrt{\frac{8 d_{VC} \log\left(\frac{en}{d_{VC}}\right)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

10.4. Structural Risk Minimization

Claim 10.4.1

- *Let* $\pi_f > 0$ *such that* $\sum_{f \in \mathcal{S}} \pi_f = 1$
- *With probability greater than* $1 - 2\delta$,

$$\mathcal{R}(\widehat{f}) - \mathcal{R}(f_{\mathcal{S}}^{\star}) \leq \sqrt{\frac{\log(1/\pi_f)}{2n}} + \sqrt{\frac{2\log(1/\delta)}{n}}$$

- *With probability greater than* $1 - \delta$, *simultaneously* $\forall f' \in \mathcal{S}$,

$$\mathcal{R}(f') \leq \mathcal{R}_n(f') + \sqrt{\frac{\log(1/\pi_f)}{2n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

Claim 10.4.2
With probability $1 - \delta$, *simultaneously for all* $m \in \mathcal{M}$ *and all* $f \in \mathcal{S}_m$,

$$\mathcal{R}(f) \leq \mathcal{R}_n(f) + \mathbb{E}\left[\sqrt{\frac{8\log|B_n(\mathcal{S}_m)|}{n}}\right] + \sqrt{\frac{\log(1/\pi_m)}{2n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

Claim 10.4.3

If $\hat{f}$ is the SRM minimizer then with probability $1-2\delta$,

$$\mathcal{R}(\widehat{f}) \leq \inf_{m\in\mathcal{M}} \inf_{f\in\mathcal{S}_m} \left(\mathcal{R}(f) + \mathbb{E}\left[\sqrt{\frac{8\log|B_n(\mathcal{S}_m)|}{n}}\right] + \sqrt{\frac{\log(1/\pi_m)}{2n}} \right) + \sqrt{\frac{2\log(1/\delta)}{n}}$$

Proof. Let $f^\star_{\mathcal{S}_{m^\star}}$ be the minimizer over $m\in\mathcal{M}$ and $f\in\mathcal{S}_m$ of

$$\mathcal{R}(f) + \mathbb{E}\left[\sqrt{\frac{8\log|B_n(\mathcal{S}_m)|}{n}}\right] + \sqrt{\frac{\log(1/\pi_m)}{2n}}$$

The previous bound is thus equivalent to the fact that with probability greater than $1-2\delta$,

$$\mathcal{R}(\widehat{f}) \leq \mathcal{R}(f^\star_{\mathcal{S}_{m^\star}}) + \mathbb{E}\left[\sqrt{\frac{8\log|B_n(\mathcal{S}_{m^\star})|}{n}}\right] + \sqrt{\frac{\log(1/\pi_m)}{2n}} + \sqrt{\frac{2\log(1/\delta)}{n}}$$

We use then that with probability $1-\delta$

$$\begin{aligned}\mathcal{R}(\hat{f}) &\leq \mathcal{R}_n(f) + \mathbb{E}\left[\sqrt{\frac{8\log|B_n(\mathcal{S}_m)|}{n}}\right] + \sqrt{\frac{\log(1/\pi_m)}{2n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \\ &\leq \mathcal{R}_n(f^\star_{\mathcal{S}_{m^\star}}) + \mathbb{E}\left[\sqrt{\frac{8\log|B_n(\mathcal{S}_{m^\star})|}{n}}\right] + \sqrt{\frac{\log(1/\pi_m)}{2n}} + \sqrt{\frac{\log(1/\delta)}{2n}}\end{aligned}$$

Combining this with the fact that with probability $1-\delta$

$$\mathcal{R}(f^\star_{\mathcal{S}_{m^\star}}) \leq \mathcal{R}_n(f^\star_{\mathcal{S}_{m^\star}}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

yields the result. □

A. Convex Optimization: Lagrangian

A.1. Constrained Optimization, Lagrangian and Dual

Theorem A.1.1

$$\max_{\lambda\in\mathbb{R}^p,\ \mu\in(\mathbb{R}^+)^q} \mathcal{L}(x,\lambda,\mu) = \begin{cases} f(x) & \textit{if } x \textit{ is feasible} \\ +\infty & \textit{otherwise} \end{cases}$$

$$\min_x \max_{\lambda\in\mathbb{R}^p,\ \mu\in(\mathbb{R}^+)^q} \mathcal{L}(x,\lambda,\mu) = \min_x f(x) \quad \textit{with} \quad \begin{cases} h_j(x) = 0, & j = 1,\dots p \\ g_i(x) \leq 0, & i = 1,\dots q \end{cases}$$

Proof. The second part is a direct consequence of the first one.
For the first part,

- if x is feasible $h_i(x) = 0$ and $g_j(x) \leq 0$ thus

$$\begin{aligned} \mathcal{L}(x,\lambda,\mu) &= f(x) + \sum_{j=1}^{p} \lambda_j h_j(x) + \sum_{i=1}^{q} \mu_i g_i(x) \\ &\leq f(x) = \mathcal{L}(x,0,0) \end{aligned}$$

  and thus $\max_{\lambda\in\mathbb{R}^p,\ \mu\in(\mathbb{R}^+)^q} \mathcal{L}(x,\lambda,\mu) = f(x)$.

- if x is not feasible either
  - $\exists i, h_i(x) \neq 0$ and thus using $\lambda_i = \kappa\,\mathrm{sign}(h_i(x))$, $\lambda_{i'} = 0$ for $i' \neq i$ and $\mu = 0$

$$\mathcal{L}(x,\lambda,\mu) = f(x) + \kappa\,\mathrm{sign}(h_i(x))h_i(x)$$

    goes to $+\infty$ when κ goes to ∞
  - or $\exists j, g_j(x) > 0$ and thus using $\lambda = 0$, $\mu_j = \kappa$ and $\mu_{j'} = 0$ for $j' \neq j$

$$\mathcal{L}(x,\lambda,\mu) = f(x) + \kappa g_j(x)$$

    goes to $+\infty$ when κ goes to ∞

  which implies $\max_{\lambda\in\mathbb{R}^p,\ \mu\in(\mathbb{R}^+)^q} \mathcal{L}(x,\lambda,\mu) = +\infty$.

□

Theorem A.1.2

$$Q(\lambda,\mu) \leq f(x), \quad \textit{for all feasible } x$$
$$\max_{\lambda\in\mathbb{R}^p,\ \mu\in(\mathbb{R}^+)^q} Q(\lambda,\mu) \leq \min_{x\ \textit{feasible}} f(x)$$

Proof. The second part is a direct consequence of the first one.
By definition,

$$\begin{aligned} Q(\lambda,\mu) &= \min_x \mathcal{L}(x,\lambda,\mu) \\ &\leq \min_{x \text{ feasible}} \mathcal{L}(x,\lambda,\mu) \\ &\leq \min_{x \text{ feasible}} f(x) \end{aligned}$$

where we have used that for x feasible $\mathcal{L}(x,\lambda,\mu) \leq f(x)$. □

A.2. Duality, weak, strong and Slater's condition

Theorem A.2.1

Weak duality:

$$q^\star \leq p^\star$$
$$\max_{\lambda\in\mathbb{R}^p,\ \mu\in(\mathbb{R}^+)^q} \min_x \mathcal{L}(x,\lambda,\mu) \leq \min_x \max_{\lambda\in\mathbb{R}^p,\ \mu\in(\mathbb{R}^+)^q} \mathcal{L}(x,\lambda,\mu)$$

Proof. This is a direct consequence of Theorem A.1.2. □

Theorem A.2.2

*If f is convex, h_j affine and g_i convex then the **Slater's condition**, it exists a feasible point such that $h_j(x) = 0$ for all j and $g_i(x) < 0$ for all i is sufficient to imply the strong duality:*

$$\max_{\lambda\in\mathbb{R}^p,\ \mu\in(\mathbb{R}^+)^q} \min_x \mathcal{L}(x,\lambda,\mu) = \min_x \max_{\lambda\in\mathbb{R}^p,\ \mu\in(\mathbb{R}^+)^q} \mathcal{L}(x,\lambda,\mu)$$

Proof. The simplest proof can be found in **boyd04**. □

A.3. Karush-Kuhn-Tucker

Theorem A.3.1

If f is convex, h_j affine and g_i convex, all are differentiable and strong duality holds then $x^\star$ is a solution of the primal problem if and only if the KKT condition

- *Stationarity:*

$$\nabla_x \mathcal{L}(x^\star, \lambda, \mu) = \nabla f(x^\star) + \sum_j \lambda_j \nabla h(x^\star) + \sum_i \mu_i \nabla g(x^\star) = 0$$

- *Primal admissibility:*

$$h_j(x^\star) = 0 \quad \textit{and} \quad g_i(x^\star) \leq 0$$

- *Dual admissibility:*

$$\mu_i \geq 0$$

- *Complementary slackness:*

$$\mu_i g_i(x^\star) = 0$$

holds.

Proof. Assume first that all the KKT conditions are satisfied then

$$\begin{aligned} f(x^\star) &= \mathcal{L}(x^\star, \lambda, \mu) \\ &= \min_x \mathcal{L}(x^\star, \lambda, \mu) \\ \leq \max_{\lambda,\mu} Q(\lambda, \mu) &\leq f(x^\star) \end{aligned}$$

and thus $f(x^\star) = max_{\lambda,\mu} Q(\lambda, \mu) \leq \min_{x \text{ feasible}} f(x)$. Thus $x^\star$ is a minimizer of the primal problem.

Let $x^\star$ is a solution of the primal problem and $(\lambda^\star, \mu^\star)$ be a solution of the dual. If the strong duality holds:

$$\begin{aligned} f(x^\star) &= Q(\lambda^\star, \mu^\star) \\ &= \min_x \mathcal{L}(x, \lambda^\star, \mu^\star) && \leq \mathcal{L}(x^\star, \lambda^\star, \mu^\star) \\ &\leq f(x^\star) \end{aligned}$$

where we have used the property that the minimizer of a convex corresponds to a 0 of the (sub)differential. Hence all the inequalities are equalities. In particular, $x^\star$ is a minimizer of $\mathcal{L}(x, \lambda^\star, \mu^\star)$. We obtain thus the stationarity condition:

$$\nabla_x \mathcal{L}(x^\star, \lambda, \mu) = \nabla f(x^\star) + \sum_j \lambda_j \nabla h_j(x^\star) + \sum_i \mu_i \nabla g_i(x^\star) = 0$$

By construction, $x^\star$ is admissible and $\mu \geq 0$. This implies the admissibility conditions:

$$\begin{aligned} h_j(x^\star) = 0 \quad \text{and} \quad g_i(x^\star) \leq 0 \\ \mu_i \geq 0. \end{aligned}$$

The complementary slackness condition is obtained by noticing that

$$\mathcal{L}(x^\star, \lambda^\star, \mu^\star) = f(x^\star)$$

which implies

$$\sum_i \mu_i g_i(x^\star) = 0$$

hence the result. □

B. Convex Optimization: Gradient Descent

C. Gradient Descent Algorithm

Here we let $G = F + R$ with R simple.

The proximal gradient descent algorithm is given by

$$\boldsymbol{w}^{[k+1]} = \text{prox}_{\alpha^{[k]},R}\left(\boldsymbol{w}^{[k]} - \alpha^{[k]}\delta_F(\boldsymbol{w}^{[k]})\right)$$

where $\delta_F(\boldsymbol{w}^{[k]})$ is a subgradient of F at $\boldsymbol{w}^{[k]}$. If F is differentiable then $\delta_F(\boldsymbol{w}^{[k]}) = \nabla F(\boldsymbol{w}^{[k]})$.

Most of the proofs are adaptation from the ones of **beck17**.

C.1. A Key Lemma

Lemma C.1.1

For any differentiable function F and $\boldsymbol{w}$, if we let

$$\boldsymbol{w}^+ = \text{prox}_{\alpha,R}(\boldsymbol{w} - \alpha\nabla F(\boldsymbol{w}))$$

then as soon as α satisfy

$$F(\boldsymbol{w}^+) \leq F(\boldsymbol{w}) + \left\langle \nabla F(\boldsymbol{w}), \boldsymbol{w}^+ - \boldsymbol{w} \right\rangle + \frac{1}{2\alpha}\|\boldsymbol{w}^+ - \boldsymbol{w}\|^2$$

then for any z

$$G(z) - G(\boldsymbol{w}^+) \geq \frac{1}{2\alpha}\|z - \boldsymbol{w}^+\|^2 - \frac{1}{2\alpha}\|z - \boldsymbol{w}\|^2 + F(z) - F(\boldsymbol{w}) - \langle \nabla F(\boldsymbol{w}), z - \boldsymbol{w}\rangle .$$

Proof. We introduce the function

$$\phi(x) = F(\boldsymbol{w}) + \langle \nabla F(\boldsymbol{w}), x - \boldsymbol{w}\rangle + R(x) + \frac{1}{2\alpha}\|x - \boldsymbol{w}\|^2$$

By construction,

$$\phi(x) = R(x) + \frac{1}{2\alpha}\|x - \boldsymbol{w} - \alpha F(\boldsymbol{w})\|^2 + F(\boldsymbol{w}) - \alpha\|\nabla F(\boldsymbol{w})\|^2$$

and thus $\boldsymbol{w}^+ = \text{prox}_{\alpha,R}(\boldsymbol{w} - \alpha\nabla F(\boldsymbol{w}))$ is the minimizer of the $1/\alpha$ strongly convex function ϕ. This implies that for any z,

$$\phi(z) - \phi(\boldsymbol{w}^+) \geq \frac{1}{2\alpha}\|z - \boldsymbol{w}^+\|^2$$

Now

$$\phi(\boldsymbol{w}^+) = F(\boldsymbol{w}) + \left\langle \nabla F(\boldsymbol{w}), \boldsymbol{w}^+ - \boldsymbol{w} \right\rangle + R(\boldsymbol{w}^+) + \frac{1}{2\alpha}\|\boldsymbol{w}^+ - \boldsymbol{w}\|^2$$

and thus using the assumption on α

$$\phi(\boldsymbol{w}^+) \geq F(\boldsymbol{w}^+) + R(\boldsymbol{w}^+) = G(\boldsymbol{w}^+)$$

while

$$\phi(z) = F(\boldsymbol{w}) + \langle \nabla F(\boldsymbol{w}), z - \boldsymbol{w} \rangle + R(z) + \frac{1}{2\alpha}\|z - \boldsymbol{w}\|^2$$

adding and subtracting $F(z)$ yields

$$\phi(z) = G(z) + \frac{1}{2\alpha}\|z - \boldsymbol{w}\|^2 + F(\boldsymbol{w}) - F(z) + \langle \nabla F(\boldsymbol{w}), z - \boldsymbol{w} \rangle$$

and thus

$$G(z) + \frac{1}{2\alpha}\|z - \boldsymbol{w}\|^2 + F(\boldsymbol{w}) - F(z) + \langle \nabla F(\boldsymbol{w}), z - \boldsymbol{w} \rangle - G(\boldsymbol{w}^+) \geq \frac{1}{2\alpha}\|z - \boldsymbol{w}^+\|^2$$

which is equivalent to the inequality in the lemma. □

Corollary C.1.2

For any convex function F and $\boldsymbol{w}$, if we let

$$\boldsymbol{w}^+ = \mathrm{prox}_{\alpha,R}(\boldsymbol{w} - \alpha \nabla F(\boldsymbol{w}))$$

then as soon as α satisfy

$$F(\boldsymbol{w}^+) \leq F(\boldsymbol{w}) + \left\langle \nabla F(\boldsymbol{w}), \boldsymbol{w}^+ - \boldsymbol{w} \right\rangle + \frac{1}{2\alpha}\|\boldsymbol{w}^+ - \boldsymbol{w}\|^2$$

then for any z

$$G(z) - G(\boldsymbol{w}^+) \geq \frac{1}{2\alpha}\|z - \boldsymbol{w}^+\|^2 - \frac{1}{2\alpha}(1 - \alpha\mu)\|z - \boldsymbol{w}\|^2$$

where $\mu > 0$ if F is μ strongly convex and $\mu = 0$ otherwise. Furthermore, $\alpha\mu \leq 1$.

Proof. This is an immediate consequence of the previous lemma as

$$F(z) - F(\boldsymbol{w}) - \langle \nabla F(\boldsymbol{w}), z - \boldsymbol{w} \rangle \geq \frac{\mu}{2}\|z - \boldsymbol{w}\|^2$$

which yields the bounds.

Furthermore, as

$$F(\boldsymbol{w}^+) \geq F(\boldsymbol{w}) + \left\langle \nabla F(\boldsymbol{w}), \boldsymbol{w}^+ - \boldsymbol{w} \right\rangle + \frac{\mu}{2}\|\boldsymbol{w}^+ - \boldsymbol{w}\|^2$$

we deduce $\mu \leq \frac{1}{\alpha}$ and thus $\alpha\mu \leq 1$. □

Corollary C.1.3

If F is convex, and we use the Gradient Descent algorithm with $\alpha^{[k]}$ such that

$$F(\boldsymbol{w}^{[k+1]}) \leq F(\boldsymbol{w}^{[k]}) + \left\langle \nabla F(\boldsymbol{w}^{[k]}), \boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]} \right\rangle + \frac{1}{2\alpha^{[k]}} \|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]}\|^2$$

then

$$G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^{[k]}) \leq -\frac{1}{2\alpha^{[k]}} \|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]}\|^2$$
$$G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^{\star}) \leq \frac{1}{2\alpha^{[k]}} (1 - \alpha^{[k]}\mu) \|\boldsymbol{w}^{[k]} - \boldsymbol{w}^{\star}\|^2 - \frac{1}{2\alpha^{[k]}} \|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{\star}\|^2$$

where $\mu > 0$ if F is μ strongly convex and $\mu = 0$ otherwise. Furthermore, $\alpha^{[k]}\mu \leq 1$.

Proof. As

$$\boldsymbol{w}^{[k+1]} = \text{prox}_{\alpha,R}(\boldsymbol{w}^{[k]} - \alpha \nabla F(\boldsymbol{w}^{[k]}))$$

we can apply the previous lemma with $z = \boldsymbol{w}^{[k]}$ and $z = \boldsymbol{w}^{\star}$ as soon as

$$F(\boldsymbol{w}^{[k+1]}) \leq F(\boldsymbol{w}^{[k]}) + \left\langle \nabla F(\boldsymbol{w}^{[k]}), \boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]} \right\rangle + \frac{1}{2\alpha^{[k]}} \|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]}\|^2.$$

This leads to

$$G(\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^{k+1}) \geq \frac{1}{2\alpha^{[k]}} \|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]}\|^2$$

and

$$G(\boldsymbol{w}^{\star}) - G(\boldsymbol{w}^{[k+1]}) \geq \frac{1}{2\alpha^{[k]}} \|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{\star}\|^2 - \frac{1}{2\alpha^{[k]}} (1 - \alpha^{[k]}\mu) \|\boldsymbol{w}^{[k]} - \boldsymbol{w}^{\star}\|^2$$

□

C.2. Gradient Descent for L-smooth Function

Theorem C.2.1

If F is L-smooth and we use the Gradient Descent algorithm with $\alpha^{[k]}$ satisfying

$$F(\boldsymbol{w}^{[k+1]}) \leq F(\boldsymbol{w}^{[k]}) + \left\langle \nabla F(\boldsymbol{w}^{[k]}), \boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]} \right\rangle + \frac{1}{2\alpha^{[k]}} \|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]}\|^2$$

then

$$G(\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^{\star}) \leq \frac{\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^{\star}\|^2}{2k \left(\frac{1}{k} \sum_{k'=0}^{k-1} \alpha^{[k']} \right)}$$

Proof. Corollary C.1.3 yields

$$G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^{[k]}) \leq -\frac{1}{2\alpha^{[k]}}\|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]}\|^2$$
$$G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^{\star}) \leq \frac{1}{2\alpha^{[k]}}\|\boldsymbol{w}^{[k]} - \boldsymbol{w}^{\star}\|^2 - \frac{1}{2\alpha^{[k]}}\|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{\star}\|^2$$

The first inequality implies that the $G(\boldsymbol{w}^{[k]})$ are decreasing. For the second one, we multiply first the inequality by $\alpha^{[k]}$ and sum them over k

$$\sum_{k'=0}^{k-1} \alpha^{[k]} \left(G(\boldsymbol{w}^{[k'+1]}) - G(\boldsymbol{w}^{\star})\right) \leq \frac{1}{2}\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^{\star}\|^2 - \frac{1}{2}\|\boldsymbol{w}^{[k]} - \boldsymbol{w}^{\star}\|^2$$

and thus as $G(\boldsymbol{w}^{[k]})$ are decreasing

$$\sum_{k'=0}^{k-1} \alpha_k G(\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^{\star}) \leq \frac{1}{2}\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^{\star}\|^2$$

which implies

$$G(\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^{\star}) \leq \frac{1}{2k\left(\frac{1}{k}\sum_{k'=0}^{k-1}\alpha^{[k]}\right)}\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^{\star}\|^2$$

□

Lemma C.2.2

if F is L smooth then if $\alpha^{[k]} \leq \frac{1}{L}$ then

$$F(\boldsymbol{w}^{[k+1]}) \leq F(\boldsymbol{w}^{[k]}) + \left\langle \nabla F(\boldsymbol{w}^{[k]}), \boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]} \right\rangle + \frac{1}{2\alpha^{[k]}}\|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]}\|^2$$

Proof. if F is L-smooth then

$$F(\boldsymbol{w}^{[k+1]}) \leq F(\boldsymbol{w}^{[k]}) + \left\langle \nabla F(\boldsymbol{w}^{[k]}), \boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]} \right\rangle + \frac{L}{2}\|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]}\|^2$$

and thus

$$\leq F(\boldsymbol{w}^{[k]}) + \left\langle \nabla F(\boldsymbol{w}^{[k]}), \boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]} \right\rangle + \frac{1}{2\alpha^{[k]}}\|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]}\|^2$$

□

Lemma C.2.3

In the backtracking algorithm, at each step

$$F(\boldsymbol{w}^{[k+1]}) \leq F(\boldsymbol{w}^{[k]}) + \left\langle \nabla F(\boldsymbol{w}^{[k]}), \boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]} \right\rangle + \frac{1}{2\alpha^{[k]}} \|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]}\|^2,$$

and

$$\frac{1}{k} \sum_{k'=0}^{k-1} \alpha^{[k']} \geq \frac{\beta}{L} \qquad \textit{and} \qquad \frac{1}{2\alpha^{[k]}} \prod_{k'=0}^{k} (1 - \alpha^{[k]}\mu) \leq \frac{L}{2\beta}(1 - \frac{\beta\mu}{L})^{k+1}$$

Proof. First point is satisfied by construction as $\alpha^{[k]}$ is equal to $\beta^l \alpha_0$ where l is the smallest integer such that $\beta^l \alpha_0$ satisfies

$$F(\boldsymbol{w}^{[k+1]}) \leq F(\boldsymbol{w}^{[k]}) + \left\langle \nabla F(\boldsymbol{w}^{[k]}), \boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]} \right\rangle + \frac{1}{2\beta^l \alpha_0} \|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]}\|^2,$$

Note that such an l exists as the condition is satisfied for any l such that $\beta^l \alpha_0 \leq 1/L$. In particular, one always has that $\alpha > \beta/L$. Furthermore, as $\alpha^{[k]}\mu \leq 1$ and $L\mu \leq 1$, we obtain $0 \leq 1 - \alpha^{[k]}\mu \leq 1 - \beta\mu/L$ this implies immediately

$$\frac{1}{k} \sum_{k'=0}^{k-1} \alpha^{[k']} \geq \frac{\beta}{L} \qquad \text{and} \qquad \frac{1}{2\alpha^{[k]}} \prod_{k'=0}^{k} (1 - \alpha^{[k]}\mu) \leq \frac{L}{2\beta}(1 - \frac{\beta\mu}{L})^{k+1}$$

□

Corollary C.2.4

If F is L-smooth, and we use the Gradient Descent algorithm with $\alpha^{[k]} = \alpha \leq 1/L$ then

$$G(\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^\star) \leq \frac{\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|^2}{2\alpha k}$$

Proof. We combine Theorem C.2.1 and Lemma C.2.2 to obtain

$$\begin{aligned} G(\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^\star) &\leq \frac{\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|^2}{2k\left(\frac{1}{k}\sum_{k'=0}^{k-1} \alpha\right)} \\ &\leq \frac{\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|^2}{2k\alpha} \end{aligned}$$

□

Corollary C.2.5

If F is L-smooth, and we use the Gradient Descent algorithm with $\alpha^{[k]}$ obtained by backtracking then

$$G(\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^\star) \leq \frac{\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|^2}{2k\left(\frac{1}{k}\sum_{k'=0}^{k-1}\alpha^{[k']}\right)}$$

with $\frac{1}{k}\sum_{k'=0}^{k-1}\alpha^{[k']} \geq \beta/L$.

Proof. This is the result of Theorem C.2.1 and Lemma C.2.3. □

C.3. Gradient Descent for Strongly Convex Function

Theorem C.3.1

If F is L-smooth and μ strongly convex, and we use the Gradient Descent algorithm with $\alpha^{[k]}$ satisfying

$$F(\boldsymbol{w}^{[k+1]}) \leq F(\boldsymbol{w}^{[k]}) + \left\langle \nabla F(\boldsymbol{w}^{[k]}), \boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]} \right\rangle + \frac{1}{2\alpha^{[k]}}\|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]}\|^2$$

then

$$G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^\star) \leq \frac{1}{2\alpha^{[k]}}\prod_{k'=0}^{k}(1 - \alpha^{[k]}\mu)\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|^2.$$

Proof. According to Corollary C.1.3, we have

$$G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^{[k]}) \leq -\frac{1}{2\alpha^{[k]}}\|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]}\|^2$$

$$G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^\star) \leq \frac{1}{2\alpha^{[k]}}(1 - \alpha^{[k]}\mu)\|\boldsymbol{w}^{[k]} - \boldsymbol{w}^\star\|^2 - \frac{1}{2\alpha^{[k]}}\|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^\star\|^2$$

The second inequality implies immediately

$$\|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^\star\|^2 \leq (1 - \alpha^{[k]}\mu)\|\boldsymbol{w}^{[k]} - \boldsymbol{w}^\star\|^2$$

so that

$$\|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^\star\|^2 \leq \prod_{k'=0}^{k}(1 - \alpha^{[k]}\mu)\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|^2.$$

Plugging this bound in the same inequality we have used yields

$$\begin{aligned} G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^\star) &\leq \frac{1}{2\alpha^{[k]}}(1 - \alpha^{[k]}\mu)\|\boldsymbol{w}^{[k]} - \boldsymbol{w}^\star\|^2 \\ &\leq \frac{1}{2\alpha^{[k]}}\prod_{k'=0}^{k}(1 - \alpha^{[k]}\mu)\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|^2. \end{aligned}$$

□

Corollary C.3.2

If F is L-smooth and μ strongly convex, and we use the Gradient Descent algorithm with $\alpha^{[k]}$ obtained by backtracking then

$$G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^\star) \leq \frac{1}{2\alpha^{[k]}} \prod_{k'=0}^{k} (1 - \alpha^{[k]}\mu) \|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|^2.$$

with

$$\frac{1}{2\alpha^{[k]}} \prod_{k'=0}^{k} (1 - \alpha^{[k]}\mu) \leq \frac{L}{2\beta}(1 - \frac{\beta\mu}{L})^{k+1}$$

Proof. This is a direct consequence of Lemma C.2.3 and Theorem C.3.1. □

Corollary C.3.3

If F is L-smooth and μ strongly convex, and we use the Gradient Descent algorithm with $\alpha^{[k]} = \alpha \leq 1/L$ then

$$G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^\star) \leq \frac{1}{2\alpha} \prod_{k'=0}^{k} (1 - \alpha\mu) \|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|^2.$$

Proof. This is a direct consequence of Lemma C.2.2 and Theorem C.3.1. □

C.4. Accelerated Gradient Descent

Theorem C.4.1

If F is convex, and we use the Accelerated Gradient Descent algorithm with $\alpha^{[k]}$ decreasing such that

$$F(\boldsymbol{w}^{[k+1]}) \leq F(\boldsymbol{w}^{[k+1/2]}) + \left\langle \nabla F(\boldsymbol{w}^{[k+1/2]}), \boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k+1/2]} \right\rangle + \frac{1}{2\alpha^{[k]}} \|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k+1/2]}\|^2$$

then provided $\beta^{[k]} = (t^{[k-1]} - 1)/t^{[k]}$ with $t^{[k]}$ satisfying $t^{[0]} = 1$, $t^{[k]} \geq 1$ and $(t^{[k+1]})^2 - t^{[k+1]} \leq (t^{[k]})^2$ then

$$G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^\star) \leq \frac{1}{2(t^{[k]})^2\alpha^{[k]}} \|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|^2.$$

Proof. As

$$\boldsymbol{w}^{[k+1]} = \mathrm{prox}_{\alpha,R}(\boldsymbol{w}^{[k+1/2]} - \alpha\nabla F(\boldsymbol{w}^{[k+1/2]}))$$

with

$$\boldsymbol{w}^{[k+1/2]} = \boldsymbol{w}^{[k]} + \beta^{[k]}(\boldsymbol{w}^{[k]} - \boldsymbol{w}^{[k-1]})$$

we can apply Corollary C.1.2 with $\boldsymbol{w} = \boldsymbol{w}^{[k+1/2]}$ and $\boldsymbol{w}^+ = \boldsymbol{w}^{[k+1]}$. As soon as $\alpha^{[k]}$ is such that

$$F(\boldsymbol{w}^{[k+1]}) \leq F(\boldsymbol{w}^{[k+1/2]}) + \left\langle \nabla F(\boldsymbol{w}^{[k+1/2]}), \boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k+1/2]} \right\rangle + \frac{1}{2\alpha^{[k]}}\|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k+1/2]}\|^2$$

we have

$$G(z) - G(\boldsymbol{w}^{[k+1]}) \geq \frac{1}{2\alpha^{[k]}}\|z - \boldsymbol{w}^{[k+1]}\|^2 - \frac{1}{2\alpha^{[k]}}\|z - \boldsymbol{w}^{[k+1/2]}\|^2$$

Using $z = \theta^{[k]}\boldsymbol{w}^\star + (1-\theta^{[k]})\boldsymbol{w}^{[k]}$ yields

$$\begin{aligned} G(\theta^{[k]}\boldsymbol{w}^\star + (1-\theta^{[k]})\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^{[k+1]}) \geq{}& \frac{1}{2\alpha^{[k]}}\|\theta^{[k]}\boldsymbol{w}^\star + (1-\theta^{[k]})\boldsymbol{w}^{[k]} - \boldsymbol{w}^{[k+1]}\|^2 \\ &- \frac{1}{2\alpha^{[k]}}\|\theta^{[k]}\boldsymbol{w}^\star + (1-\theta^{[k]})\boldsymbol{w}^{[k]} - \boldsymbol{w}^{[k+1/2]}\|^2 \end{aligned}$$

By convexity of G,

$$\begin{aligned} G(\theta^{[k]}\boldsymbol{w}^\star + (1-\theta^{[k]})\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^{[k+1]}) &\leq \theta^{[k]}G(\boldsymbol{w}^\star) + (1-\theta^{[k]})G(\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^{[k+1]}) \\ &\leq (1-\theta^{[k]})\left(G(\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^\star)\right) - \left(G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^\star)\right) \end{aligned}$$

Now

$$\begin{aligned} \|\theta^{[k]}\boldsymbol{w}^\star + (1-\theta^{[k]})\boldsymbol{w}^{[k]} - \boldsymbol{w}^{[k+1/2]}\|^2 &= \|\theta^{[k]}\boldsymbol{w}^\star + (1-\theta^{[k]})\boldsymbol{w}^{[k]} - \boldsymbol{w}^{[k]} - \beta^{[k]}\left(\boldsymbol{w}^k - \boldsymbol{w}^{k-1}\right)\|^2 \\ &= \|\theta^{[k]}\boldsymbol{w}^\star + \beta^{[k]}\boldsymbol{w}^{[k-1]} - (\beta^{[k]} + \theta^{[k]})\boldsymbol{w}^k\|^2 \\ &= \left(\frac{\theta^{[k]}}{\theta^{[k-1]}}\right)^2 \left\|\theta^{[k-1]}\boldsymbol{w}^\star + \frac{\theta^{[k-1]}}{\theta^{[k]}}\beta^{[k]}\boldsymbol{w}^{[k-1]} - \frac{\theta^{[k-1]}}{\theta^{[k]}}\left(\beta^{[k]} + \theta^{[k]}\right)\boldsymbol{w}^{[k}\right. \end{aligned}$$

if we let $\theta^{[k]} = \beta^{[k]}\frac{\theta^{[k-1]}}{1-\theta^{[k-1]}}$, we obtain provided $0 \leq \theta^{[k]} \leq 1$

$$= \left(\frac{\theta^{[k]}}{\theta^{[k-1]}}\right)^2 \|\theta^{[k-1]}\boldsymbol{w}^\star + (1-\theta^{[k-1]})\boldsymbol{w}^{[k-1]} - \boldsymbol{w}^{[k]}\|^2$$

Combining the two previous bounds yields

$$\begin{aligned} &(1-\theta^{[k]})\alpha^{[k]}\left(G(\boldsymbol{w}^{[k]}) - G(\boldsymbol{w}^\star)\right) - \alpha^{[k]}\left(G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^\star)\right) \\ &\quad\geq \frac{1}{2}\|\theta^{[k]}\boldsymbol{w}^\star + (1-\theta^{[k]})\boldsymbol{w}^{[k]} - \boldsymbol{w}^{[k+1]}\|^2 - \frac{1}{2}\left(\frac{\theta^{[k]}}{\theta^{[k-1]}}\right)^2 \|\theta^{[k-1]}\boldsymbol{w}^\star + (1-\theta^{[k-1]})\boldsymbol{w}^{[k-1]} - \boldsymbol{w}^{[k]}\|^2 \end{aligned}$$

and equivalently

$$\begin{aligned}
&\frac{1}{(\theta^{[k]})^2}\left(\alpha^{[k]}\left(G(\boldsymbol{w}^{[k+1]})-G(\boldsymbol{w}^\star)\right)+\frac{1}{2}\|\theta^{[k]}\boldsymbol{w}^\star+(1-\theta^{[k]})\boldsymbol{w}^{[k]}-\boldsymbol{w}^{[k+1]}\|^2\right)\\
&\quad\leq\frac{1}{(\theta^{[k-1]})^2}\left(\frac{(\theta^{[k-1]})^2(1-\theta^{[k]})}{(\theta^{[k]})^2}\alpha^{[k]}\left(G(\boldsymbol{w}^{[k]})-G(\boldsymbol{w}^\star)\right)+\frac{1}{2}\|\theta^{[k-1]}\boldsymbol{w}^\star+(1-\theta^{[k-1]})\boldsymbol{w}^{[k-1]}-\boldsymbol{w}^{[k]}\|^2\right)\\
&\quad\leq\frac{1}{(\theta^{[k-1]})^2}\left(\alpha^{[k-1]}\left(G(\boldsymbol{w}^{[k]})-G(\boldsymbol{w}^\star)\right)+\frac{1}{2}\|\theta^{[k-1]}\boldsymbol{w}^\star+(1-\theta^{[k-1]})\boldsymbol{w}^{[k-1]}-\boldsymbol{w}^{[k]}\|^2\right)
\end{aligned}$$

provided

$$\frac{(\theta^{[k-1]})^2(1-\theta^{[k]})}{(\theta^{[k]})^2}\alpha^{[k]}\leq\alpha^{[k-1]}.$$

If this holds, one has

$$\begin{aligned}
&\frac{1}{(\theta^{[k]})^2}\left(\alpha^{[k]}\left(G(\boldsymbol{w}^{[k+1]})-G(\boldsymbol{w}^\star)\right)+\frac{1}{2}\|\theta^{[k]}\boldsymbol{w}^\star+(1-\theta^{[k]})\boldsymbol{w}^{[k]}-\boldsymbol{w}^{[k+1]}\|^2\right)\\
&\quad\leq\frac{1}{(\theta^{[0]})^2}\left(\alpha^{[0]}\left(G(\boldsymbol{w}^{[1]})-G(\boldsymbol{w}^\star)\right)+\frac{1}{2}\|\theta^{[0]}\boldsymbol{w}^\star+(1-\theta^{[0]})\boldsymbol{w}^{[0]}-\boldsymbol{w}^{[1]}\|^2\right)
\end{aligned}$$

Using the result obtained with Corollary C.1.2 at $k=0$ and using $\boldsymbol{w}^{[1/2]}=\boldsymbol{w}^{[0]}$, we obtain

$$\begin{aligned}
&\frac{1}{(\theta^{[k]})^2}\left(\alpha^{[k]}\left(G(\boldsymbol{w}^{[k+1]})-G(\boldsymbol{w}^\star)\right)+\frac{1}{2}\|\theta^{[k]}\boldsymbol{w}^\star+(1-\theta^{[k]})\boldsymbol{w}^{[k]}-\boldsymbol{w}^{[k+1]}\|^2\right)\\
&\quad\leq\frac{1}{(\theta^{[0]})^2}\left(\frac{1}{2}\|\boldsymbol{w}^{[0]}-\boldsymbol{w}^\star\|-\frac{1}{2}\|\boldsymbol{w}^{[1]}-\boldsymbol{w}^\star\|^2+\frac{1}{2}\|\theta^{[0]}\boldsymbol{w}^\star+(1-\theta^{[0]})\boldsymbol{w}^{[0]}-\boldsymbol{w}^{[1]}\|^2\right)
\end{aligned}$$

and thus if we assume that $\theta^{[0]}=1$

$$\begin{aligned}
&\frac{1}{(\theta^{[k]})^2}\left(\alpha^{[k]}\left(G(\boldsymbol{w}^{[k+1]})-G(\boldsymbol{w}^\star)\right)+\frac{1}{2}\|\theta^{[k]}\boldsymbol{w}^\star+(1-\theta^{[k]})\boldsymbol{w}^{[k]}-\boldsymbol{w}^{[k+1]}\|^2\right)\\
&\quad\leq\frac{1}{2}\|\boldsymbol{w}^{[0]}-\boldsymbol{w}^\star\|^2
\end{aligned}$$

We deduce thus the following bound

$$G(\boldsymbol{w}^{[k+1]})-G(\boldsymbol{w}^\star)\leq\frac{(\theta^{[k]})^2}{2\alpha^{[k]}}\|\boldsymbol{w}^{[0]}-\boldsymbol{w}^\star\|^2$$

Defining everything in terms of $t^{[k]}=1/\theta^{[k]}$ yields

$$\begin{aligned}
\beta^{[k]}&=\frac{\theta^{[k]}(1-\theta^{[k-1]})}{\theta^{[k-1]}}\\
&=\frac{t^{[k-1]}-1}{t^{[k]}}
\end{aligned}$$

we have obtained

$$G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^\star) \leq \frac{1}{2(t^{[k]})^2\alpha^{[k]}}\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|^2$$

provided $t^{[0]} = 1$,

$$t^{[k]} \geq 1$$

and

$$\left((t^{[k]})^2 - t^{[k]}\right)\alpha^{[k]} \leq \alpha^{[k-1]}(t^{[k-1]})^2.$$

As we assume that the $\alpha^{[k]}$ are decreasing, it is enough to verify that

$$(t^{[k]})^2 - t^{[k]} \leq (t^{[k-1]})^2$$

□

Corollary C.4.2

If F is convex, L-smooth and we use the Accelerated Gradient Descent algorithm with either $\alpha^{[k]} \leq 1/L$ or $\alpha^{[k]}$ obtain by the decreasing backtracking algorithm then for $\beta^{[k]} = (t^{[k-1]} - 1)/t^{[k]}$ defined with either Nesterov choice of $t^{[k]}$ or $t^{[k]} = \frac{k+k_0}{k_0}$ with $k_0 \geq 2$ then

$$G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^\star) \leq \frac{k_0}{2(k+k_0)^2\gamma L)^2}\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|^2.$$

with $\gamma = 1$ for the constant step size and $k_0 = 2$ for Nesterov's choice.

Proof. The bound

$$(t^{[k]})^2 - t^{[k]} \leq (t^{[k-1]})^2$$

is equivalent to

$$t^{[k]} \leq \frac{1 + \sqrt{1 + 4(t^{[k-1]})^2}}{2}$$

Nesterov parameters is obtained by optimizing this later bound and defining $t^{[k]} = \frac{1+\sqrt{1+4(t^{[k-1]})^2}}{2}$ starting from $t^{[0]} = 1$. Note that if $t^{[k]} \geq (k+2)/2$ then

$$\begin{aligned} t^{[k+1]} &= \frac{1 + \sqrt{1 + 4t^{[k]}}}{2} \\ &\geq \frac{1 + \sqrt{1 + (k+2)^2}}{2} \\ &\geq \frac{1 + k + 2}{2} = \frac{(k+1) + 2}{2} \end{aligned}$$

and thus this property is satisfied for any k.

One verify easily that the choice $t^{[k]} = \frac{k+k_0}{k_0}$ is suitable as $t^{[0]} = 1$ and

$$\begin{aligned}(t^{[k+1]})^2 - t^{[k+1]} - (t^{[k]})^2 &= \left(\frac{k+1+k_0}{k_0}\right)^2 - \frac{k+1+k_0}{k_0} - \left(\frac{k+k_0}{k_0}\right)^2 \\ &= \frac{1}{k_0^2}\left((k+1+k_0)^2 - k_0(k+1+k_0) - (k+k_0^2)\right) \\ &= \frac{1}{k_0^2}\left(2(k+k_0) + 1 - k_0(k+1+k_0)\right) \\ &= \frac{1}{k_0^2}\left((2-k_0)k + 1 - k_0(1+k_0)\right) \leq 0\end{aligned}$$

as soon as $k_0 \geq 2$. It leads to

$$\beta^{[k]} = \frac{t^{[k-1]} - 1}{t^{[k]}} = \frac{\frac{k-1+k_0}{k_0} - 1}{\frac{k+k_0}{k_0}} = \frac{k-1}{k+k_0}$$

□

C.5. Subgradient Descent

Theorem C.5.1

If F is convex such that the sub gradient δ_F can be bounded, $\|\delta_F\|^2 \leq B^2$, $\|\boldsymbol{w}^{[k]} - \boldsymbol{w}^\star\| \leq r^2$ then

$$\min_{0 \leq k' \leq k-1} F(\boldsymbol{w}^{[k']}) - F(\boldsymbol{w}^\star) \leq \frac{r^2 + \sum_{k'=0}^{k-1} (\alpha^{[k']})^2 B^2}{2\sum_{k'=0}^{k-1} \alpha^{[k']}}$$

$$F\left(\frac{1}{k}\sum_{k'=1}^{k} \boldsymbol{w}^{[k']}\right) - F(\boldsymbol{w}^\star) \leq \frac{r^2 + \sum_{k=0}^{k-1} (\alpha^{[k']})^2 B^2}{2k \min_{1 \leq k' \leq k} \alpha^{[k']}}$$

Proof. As R is the characteristic function of a convex set C and thus the proximal operator is a projection, one verify immediately that provided that $\boldsymbol{w}^{[k]} \in C$,

$$\begin{aligned}\|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^\star\|^2 &\leq \|\boldsymbol{w}^{[k]} - \alpha^{[k]}\delta_F(\boldsymbol{w}^{[k]}) - \boldsymbol{w}^\star\|^2 \\ &\leq \|\boldsymbol{w}^{[k]} - \boldsymbol{w}^\star\|^2 - 2\alpha^{[k]}\left\langle \delta_F(\boldsymbol{w}^{[k]}), \boldsymbol{w}^{[k]} - \boldsymbol{w}^\star \right\rangle + (\alpha^{[k]})^2 \|\delta_F(\boldsymbol{w}^{[k]})\|^2 \\ &\leq \|\boldsymbol{w}^{[k]} - \boldsymbol{w}^\star\|^2 + 2\alpha^{[k]}\left(F(\boldsymbol{w}^\star) - F(\boldsymbol{w}^{[k]})\right) + (\alpha^{[k]})^2 \|\delta_F(\boldsymbol{w}^{[k]})\|^2\end{aligned}$$

this implies

$$\alpha^{[k]}\left(F(\boldsymbol{w}^{[k]}) - F(\boldsymbol{w}^\star)\right) \leq \frac{1}{2}\left(\|\boldsymbol{w}^{[k]} - \boldsymbol{w}^\star\|^2 - \|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^\star\|^2\right) + \frac{(\alpha^{[k]})^2}{2}\|\delta_F(\boldsymbol{w}^{[k]})\|^2.$$

Summing those bounds along k yields

$$\sum_{k'=0}^{k-1} \alpha^{[k']} \left(F(\boldsymbol{w}^{[k']}) - F(\boldsymbol{w}^\star) \right) \leq \frac{1}{2} \|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|^2 + \sum_{k=0}^{k-1} \frac{(\alpha^{[k']})^2}{2} \|\delta_F(\boldsymbol{w}^{[k']})\|^2.$$

We deduce thus that

$$\sum_{k'=0}^{k-1} \alpha^{[k']} \left(\min_{0 \leq k' \leq k-1} F(\boldsymbol{w}^{[k']}) - F(\boldsymbol{w}^\star) \right) \leq \frac{1}{2} \|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|^2 + \sum_{k'=0}^{k-1} \frac{(\alpha^{[k']})^2}{2} \|\delta_F(\boldsymbol{w}^{[k']})\|^2$$

that is

$$\min_{0 \leq k' \leq k-1} F(\boldsymbol{w}^{[k']}) - F(\boldsymbol{w}^\star) \leq \frac{\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|^2 + \sum_{k'=0}^{k-1} (\alpha^{[k']})^2 \|\delta_F(\boldsymbol{w}^{[k']})\|^2}{2 \sum_{k=0}^{k-1} \alpha^{[k']}}$$

Along the same line, we have simultaneously

$$\min_{1 \leq k' \leq k} \alpha^{[k']} \sum_{k'=1}^{k} \left(F(\boldsymbol{w}^{[k']}) - F(\boldsymbol{w}^\star) \right) \leq \frac{1}{2} \|\boldsymbol{w}^{[1]} - \boldsymbol{w}^\star\|^2 + \sum_{k'=0}^{k-1} \frac{(\alpha^{[k']})^2}{2} \|\delta_F(\boldsymbol{w}^{[k']})\|^2$$

and thus

$$\frac{1}{k} \sum_{k'=1}^{k} \left(F(\boldsymbol{w}^{[k']}) - F(\boldsymbol{w}^\star) \right) \leq \frac{\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|^2 + \sum_{k'=0}^{k-1} (\alpha^{[k']})^2 \|\delta_F(\boldsymbol{w}^{[k']})\|^2}{2k \min_{1 \leq k' \leq k} \alpha^{[k']}}$$

and thus using the convexity of F

$$F\left(\frac{1}{k} \sum_{k'=1}^{k} \boldsymbol{w}^{[k']} \right) - F(\boldsymbol{w}^\star) \leq \frac{\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^\star\|^2 + \sum_{k'=0}^{k-1} (\alpha^{[k']})^2 \|\delta_F(\boldsymbol{w}^{[k']})\|^2}{2k \min_{1 \leq k' \leq k} \alpha^{[k']}}$$

If we assume that $\|\boldsymbol{w}^{[k]} - \boldsymbol{w}^\star\|^2 \leq r^2$ and $\|\delta_F(\boldsymbol{w}^{[k']})\|^2 \leq B^2$ then this yields

$$\min_{0 \leq k' \leq k-1} F(\boldsymbol{w}^{[k']}) - F(\boldsymbol{w}^\star) \leq \frac{r^2 + \sum_{k'=0}^{k-1} (\alpha^{[k']})^2 B^2}{2 \sum_{k'=0}^{k-1} \alpha^{[k']}}$$

$$F\left(\frac{1}{k} \sum_{k'=1}^{k} \boldsymbol{w}^{[k']} \right) - F(\boldsymbol{w}^\star) \leq \frac{r^2 + \sum_{k=0}^{k-1} (\alpha^{[k']})^2 B^2}{2k \min_{1 \leq k' \leq k} \alpha^{[k']}}$$

□

Theorem C.5.2

If F is convex such that the sub gradient δ_F can be bounded, $\|\delta_F\|^2 \leq B^2$, $\|\boldsymbol{w}^{[k]} - \boldsymbol{w}^\star\| \leq r^2$ then for $\alpha^{[k]} = \alpha_0/\sqrt{k}$ with $\alpha_0 = r/(\sqrt{2}B)$, we have

$$F\left(\frac{1}{k} \sum_{k'=1}^{k} \boldsymbol{w}^{[k']} \right) - F(\boldsymbol{w}^\star) \leq \frac{\sqrt{2} r B}{k}$$

and

$$\min_{k' \leq k} F(\boldsymbol{w}^{[k']}) - F(\boldsymbol{w}^\star) \leq \frac{\sqrt{2} r B}{k}$$

Proof. We start from the first bound obtain in the proof of the previous theorem

$$\alpha^{[k]}\left(F(\boldsymbol{w}^{[k]})-F(\boldsymbol{w}^\star)\right)\leq\frac{1}{2}\left(\|\boldsymbol{w}^{[k]}-\boldsymbol{w}^\star\|^2-\|\boldsymbol{w}^{[k+1]}-\boldsymbol{w}^\star\|^2\right)+\frac{(\alpha^{[k]})^2}{2}\|\delta_F(\boldsymbol{w}^{[k]})\|^2$$

or rather

$$F(\boldsymbol{w}^{[k]})-F(\boldsymbol{w}^\star)\leq\frac{1}{2\alpha^{[k]}}\left(\|\boldsymbol{w}^{[k]}-\boldsymbol{w}^\star\|^2-\|\boldsymbol{w}^{[k+1]}-\boldsymbol{w}^\star\|^2\right)+\frac{\alpha^{[k]}}{2}\|\delta_F(\boldsymbol{w}^{[k]})\|^2$$

We are going to use that the $\alpha^{[k]}$ are decreasing we have

$$\begin{aligned}\sum_{k'=1}^{k}\left(F(\boldsymbol{w}^{[k']})-F(\boldsymbol{w}^\star)\right)&\leq\sum_{k'=1}^{k}\left(\frac{1}{2\alpha^{[k']}}\left(\|\boldsymbol{w}^{[k']}-\boldsymbol{w}^\star\|^2-\|\boldsymbol{w}^{[k'+1]}-\boldsymbol{w}^\star\|^2\right)+\frac{\alpha^{[k']}}{2}\|\delta_F(\boldsymbol{w}^{[k']})\|^2\right)\\&\leq\frac{\|\boldsymbol{w}^{[1]}-\boldsymbol{w}^\star\|^2}{2\alpha^{[1]}}+\sum_{k'=2}^{k-1}\left(\frac{1}{\alpha^{[k']}}-\frac{1}{\alpha^{[k'-1]}}\right)\|\boldsymbol{w}^{[k']}-\boldsymbol{w}^\star\|^2+\sum_{k'=1}^{k}\frac{\alpha^{[k']}}{2}\|\delta_F(\boldsymbol{w}^{[k']})\|^2\\&\leq\frac{\|\boldsymbol{w}^{[1]}-\boldsymbol{w}^\star\|^2}{2\alpha^{[1]}}+\sum_{k'=2}^{k-1}\left(\frac{1}{2\alpha^{[k']}}-\frac{1}{2\alpha^{[k'-1]}}\right)\|\boldsymbol{w}^{[k']}-\boldsymbol{w}^\star\|^2+\sum_{k'=1}^{k}\frac{\alpha^{[k']}}{2}\|\delta_F(\boldsymbol{w}^{[k']})\|^2\end{aligned}$$

If we assume that $\|\boldsymbol{w}^{[k]}-\boldsymbol{w}^\star\|^2\leq r^2$ and $\|\delta_F(\boldsymbol{w}^{[k']})\|^2\leq B^2$ then this yields

$$\begin{aligned}\min_{0\leq k'\leq k-1}F(\boldsymbol{w}^{[k']})-F(\boldsymbol{w}^\star)&\leq\frac{r^2+\sum_{k'=0}^{k-1}(\alpha^{[k']})^2B^2}{2\sum_{k'=0}^{k-1}\alpha^{[k']}}\\F\left(\frac{1}{k}\sum_{k'=1}^{k}\boldsymbol{w}^{[k']}\right)-F(\boldsymbol{w}^\star)&\leq\frac{r^2+\sum_{k=0}^{k-1}(\alpha^{[k']})^2B^2}{2k\min_{1\leq k'\leq k}\alpha^{[k']}}\end{aligned}$$

and if the $\alpha^{[k]}$ are decreasing

$$\begin{aligned}\min_{0\leq k'\leq k-1}F(\boldsymbol{w}^{[k']})-F(\boldsymbol{w}^\star)&\leq\frac{\frac{r^2}{\alpha^{[1]}}+\sum_{k'=1}^{k}\alpha^{[k']}B^2}{2k}\\F\left(\frac{1}{k}\sum_{k'=1}^{k}\boldsymbol{w}^{[k']}\right)-F(\boldsymbol{w}^\star)&\leq\frac{\frac{r^2}{\alpha^{[1]}}+\sum_{k'=1}^{k}\alpha^{[k']}B^2}{2k}\end{aligned}$$

Plugging $\alpha^{[k]}=\alpha_0/\sqrt{k}$ and using $\sum_{k'=1}^{k}\frac{1}{\sqrt{k'}}\leq 2\sqrt{k}$ and $\sum_{k'=1}^{k}1/k'\leq\ln(k)+1$ yields

$$F\left(\frac{1}{k}\sum_{k'=1}^{k}\boldsymbol{w}^{[k']}\right)-F(\boldsymbol{w}^\star)\leq\frac{r^2}{2\alpha_0\sqrt{k}}+\frac{\alpha_0}{\sqrt{k}}B^2$$

Optimizing in α_0 yields $\alpha_0=r/(\sqrt{2}B)$ and

$$F\left(\frac{1}{k}\sum_{k'=1}^{k}\boldsymbol{w}^{[k']}\right)-F(\boldsymbol{w}^\star)\leq\frac{\sqrt{2}rB}{k}$$

□

Theorem C.5.3

If F is μ strongly convex and $\|\nabla F\|^2 \leq B^2$ then for $\alpha^{[k]} = \frac{\alpha_0}{k}$ with $\alpha_0 \geq \frac{2}{\mu}$

$$F\left(\frac{1}{k(k+1)}\sum_{k'=1}^{k} k' \boldsymbol{w}^{[k']}\right) - F(\boldsymbol{w}^\star) \leq \frac{\alpha_0 B^2}{2(k+1)}$$

and

$$\min_{k' \leq k} F(\boldsymbol{w}^{[k']}) - F(\boldsymbol{w}^\star) \leq \frac{\alpha_0 B^2}{2(k+1)}$$

Proof. Using the strong convexity of F

$$\begin{aligned}
\|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^\star\|^2 &\leq \|\boldsymbol{w}^{[k]} - \alpha^{[k]}\nabla F(\boldsymbol{w}^{[k]}) - \boldsymbol{w}^\star\|^2 \\
&\leq \|\boldsymbol{w}^{[k]} - \boldsymbol{w}^\star\|^2 - 2\alpha^{[k]}\left\langle \nabla F(\boldsymbol{w}^{[k]}), \boldsymbol{w}^{[k]} - \boldsymbol{w}^\star \right\rangle + (\alpha^{[k]})^2 \|\delta_F(\boldsymbol{w}^{[k]})\|^2 \\
&\leq \|\boldsymbol{w}^{[k]} - \boldsymbol{w}^\star\|^2 + +2\alpha^{[k]}\left(F(\boldsymbol{w}^\star) - F(\boldsymbol{w}^{[k]})\right) - \alpha^{[k]}\mu\|\boldsymbol{w}^{[k]} - \boldsymbol{w}^\star\|^2 + (\alpha^{[k]})^2\|\delta_F(\boldsymbol{w}^{[k]})\|^2
\end{aligned}$$

which implies

$$F(\boldsymbol{w}^{[k]}) - F(\boldsymbol{w}^\star) \leq \frac{1}{2\alpha^{[k]}}\left((1-\alpha^{[k]}\mu)\|\boldsymbol{w}^{[k]} - \boldsymbol{w}^\star\|^2 - \|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^\star\|^2\right) + \frac{\alpha^{[k]}}{2}\|\nabla F\|^2$$

We can now sum those inequalities

$$\begin{aligned}
\sum_{k'=1}^{k} k'\left(F(\boldsymbol{w}^{[k']}) - F(\boldsymbol{w}^\star)\right) &\leq \sum_{k'=1}^{k} \frac{k'}{2\alpha^{[k']}}\left((1-\alpha^{[k']}\mu)\|\boldsymbol{w}^{[k']} - \boldsymbol{w}^\star\|^2 - \|\boldsymbol{w}^{[k'+1]} - \boldsymbol{w}^\star\|^2\right) + \sum_{k'=1}^{k}\frac{k'\alpha^{[k']}}{2}\|\nabla \\
&\leq \frac{1-\alpha^{[1]}\mu}{2\alpha^{[1]}}\|\boldsymbol{w}^{[1]} - \boldsymbol{w}^\star\|^2 + \sum_{k'=2}^{k}\left(\frac{k'(1-\alpha^{[k']}\mu)}{2\alpha^{[k']}} - \frac{k'-1}{2\alpha^{[k'-1]}}\right)\|\boldsymbol{w}^{[k']} - \boldsymbol{w}^\star\|^2 \\
&\quad + \sum_{k'=1}^{k}\frac{k'\alpha^{[k']}}{2}\|\nabla F\|^2
\end{aligned}$$

One verify easily that for $\alpha^{[k]} = \alpha_0/k$ this yields

$$\leq \frac{1-\alpha_0\mu}{2\alpha_0}\|\boldsymbol{w}^{[1]} - \boldsymbol{w}^\star\|^2 + \sum_{k'=2}^{k}\frac{(2-\alpha_0\mu)k-1}{2\alpha_0}\|\boldsymbol{w}^{[k']} - \boldsymbol{w}^\star\|^2 + \frac{\alpha_0}{2}\sum_{k'=1}^{k}\|\nabla F\|$$

so that for any $\alpha_0 \geq \frac{2}{\mu}$

$$\begin{aligned}
&\leq \frac{1-\alpha_0\mu}{2\alpha_0}\|\boldsymbol{w}^{[1]}-\boldsymbol{w}^\star\|^2 + \frac{\alpha_0}{2}\sum_{k'=1}^{k}\|\nabla F\|^2 \\
&\leq \frac{\alpha_0}{2}\sum_{k'=1}^{k}\|\nabla F\|^2 \\
&\leq \frac{k\alpha_0 B^2}{2}
\end{aligned}$$

By convexity of F

$$\begin{aligned}
F\left(\frac{1}{k(k+1)}\sum_{k'=1}^{k}k'\boldsymbol{w}^{[k']}\right) - F(\boldsymbol{w}^\star) &\leq \frac{1}{k(k+1)}\sum k' = 1^k k'\left(F(\boldsymbol{w}^{[k']}) - F(\boldsymbol{w}^\star)\right) \\
&\leq \frac{\alpha_0 B^2}{2(k+1)}
\end{aligned}$$

Note that using

$$\min_{k'\leq k} F(\boldsymbol{w}^k) \leq \frac{1}{k(k+1)}\sum_{k'=1}^{k}k'F(\boldsymbol{w}^{[k']})$$

leads to

$$\min_{k'\leq k} F(\boldsymbol{w}^{[k']}) - F(\boldsymbol{w}^\star) \leq \frac{\alpha_0 B^2}{2(k+1)}$$

□

C.6. Stochastic Gradient Descent

Theorem C.6.1

Assume we have access to $\widehat{\delta_F}(\boldsymbol{w})$ which verify $\mathbb{E}\left[\widehat{\delta_F}(\boldsymbol{w})\right] = \delta_F(\boldsymbol{w})$ where $\delta_F(\boldsymbol{w})$ is a subgradient of F at $\boldsymbol{w}$ and $\mathbb{E}\left[\|\widehat{\delta_F}(\boldsymbol{w})\|^2|\boldsymbol{w}\right] \leq B$.

- *if F is convex and $\|\boldsymbol{w}^{[k]}-\boldsymbol{w}^\star\| \leq r^2$ then for $\alpha^{[k]} = \alpha_0/\sqrt{k}$ with $\alpha_0 = r/(\sqrt{2}B)$, we have*

$$\mathbb{E}\left[F\left(\frac{1}{k}\sum_{k'=1}^{k}\boldsymbol{w}^{[k']}\right)\right] - F(\boldsymbol{w}^\star) \leq \frac{\sqrt{2}rB}{k}$$

- *if F is μ strongly convex then for $\alpha^{[k]} = \frac{\alpha_0}{k}$ with $\alpha_0 \geq \frac{2}{\mu}$*

$$\mathbb{E}\left[F\left(\frac{1}{k(k+1)}\sum_{k'=1}^{k}k'\boldsymbol{w}^{[k']}\right)\right] - F(\boldsymbol{w}^\star) \leq \frac{\alpha_0 B^2}{2(k+1)}$$

Proof. In this stochastic setting, we have, if we let $\mu = 0$ if F is not strongly convex:

$$\begin{aligned}
\mathbb{E}\Big[\|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^\star\|^2 | \boldsymbol{w}^{[k]}\Big] &\leq \mathbb{E}\Big[\|\boldsymbol{w}^{[k]} - \alpha^{[k]}\widehat{\delta_F}(\boldsymbol{w}^{[k]}) - \boldsymbol{w}^\star\|^2 | \boldsymbol{w}^{[k]}\Big] \\
&\leq \mathbb{E}\Big[\|\boldsymbol{w}^{[k]} - \boldsymbol{w}^\star\|^2 | \boldsymbol{w}^{[k]}\Big] - 2\alpha^{[k]}\mathbb{E}\Big[\Big\langle \widehat{\delta_F}(\boldsymbol{w}^{[k]}), \boldsymbol{w}^{[k]} - \boldsymbol{w}^\star \Big\rangle | \boldsymbol{w}^{[k]}\Big] \\
&\qquad + (\alpha^{[k]})^2\mathbb{E}\Big[\|\delta_F(\boldsymbol{w}^{[k]})\|^2 | \boldsymbol{w}^{[k]}\Big] \\
&\leq \|\boldsymbol{w}^{[k]} - \boldsymbol{w}^\star\|^2 - 2\alpha^{[k]}\Big\langle \delta_F(\boldsymbol{w}^{[k]}), \boldsymbol{w}^{[k]} - \boldsymbol{w}^\star \Big\rangle + (\alpha^{[k]})^2 B^2 \\
&\leq (1 - \alpha^{[k]}\mu)\|\boldsymbol{w}^{[k]} - \boldsymbol{w}^\star\|^2 - 2\alpha^{[k]}\Big(F(\boldsymbol{w}^{[k]}) - F(\boldsymbol{w}^\star)\Big) + (\alpha^{[k]})^2 B^2
\end{aligned}$$

which implies

$$F(\boldsymbol{w}^{[k]}) - F(\boldsymbol{w}^\star) \leq \frac{1}{2\alpha^{[k]}}\Big((1 - \alpha^{[k]}\mu)\|\boldsymbol{w}^{[k]} - \boldsymbol{w}^\star\|^2 - \mathbb{E}\Big[\|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^\star\|^2 | \boldsymbol{w}^{[k]}\Big]\Big) + \frac{\alpha^{[k]}}{2}B^2$$

and thus

$$\mathbb{E}\Big[F(\boldsymbol{w}^{[k]})\Big] - F(\boldsymbol{w}^\star) \leq \frac{1}{2\alpha^{[k]}}\Big((1 - \alpha^{[k]}\mu)\mathbb{E}\Big[\|\boldsymbol{w}^{[k]} - \boldsymbol{w}^\star\|^2\Big] - \mathbb{E}\Big[\|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^\star\|^2\Big]\Big) + \frac{\alpha^{[k]}}{2}B^2$$

We can now repeat the proof of the previous lemmas to obtain the results. □

D. RKHS

D.1. Reproducing Kernel Hilbert Space

We propose a short introduction to RKHS, more details can be found in **berlinet04** for instance.

Definition D.1.1

A RKHS $\mathbb{H}$ is defined as a Hilbert space of real valued function defined on $\mathcal{X}$ in which the evaluation operator at $\underline{X}$, $\delta_{\underline{X}}$,

$$\mathcal{H} \to \mathbb{R} : f \mapsto f(\underline{X})$$

is continuous for all $\underline{X}$.

Remark D.1.2

The continuity of δ_x means that for any $\underline{X}$, it exists a constant $C_x < \infty$ such that

$$|f(\underline{X})| \leq C_{\underline{X}} \|f\|_{\mathbb{H}}$$

We can now define the kernel associated to the RKHS

Theorem D.1.3

If $\mathbb{H}$ is a RKHS then it exists a unique kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

- *for any $x \in \mathcal{X}$, $k(\underline{X}, \cdot) \in \mathbb{H}$*
- *and for any $x \in \mathcal{X}$ and any $f \in \mathbb{H}$*

$$f(\underline{X}) = \langle f, k(\underline{X}, \cdot) \rangle_{\mathbb{H}}$$

Proof. By definition, if $\mathbb{H}$ is a RKHS then δ_x is linear and continuous and thus, thanks to the Riesz theorem, it exists a unique $k(\underline{X}, \cdot) \in \mathbb{H}$ such that

$$f(\underline{X}) = \langle f, k(\underline{X}, \cdot) \rangle_{\mathbb{H}} .$$

By construction, k is thus the unique function from $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfying this. □

We define now the notion of Positive Definite Symmetric kernel:

Definition D.1.4

A kernel k is PDS if and only if

- *k is symmetric, i.e.*

$$k(\underline{X}, \underline{X}') = k(\underline{X}', \underline{X})$$

- *for any $N \in \mathbb{N}$ and any $(\underline{X}_1, \dots, \underline{X}_N) \in \mathcal{X}^N$,*

$$\boldsymbol{K} = [k(\underline{X}_i, \underline{X}_j)]_{1 \leq i,j \leq N}$$

is positive semi-definite, i.e. $\forall u \in \mathbb{R}^N$

$$u^\top \boldsymbol{K} u = \sum_{1 \leq i,j \leq N} u^{(i)} u^{(j)} k(\underline{X}_i, \underline{X}_j) \geq 0$$

or equivalently all the eigenvalues of $\boldsymbol{K}$ are non-negative.

Property D.1.5

The kernel k of a RKHS is a Positive Definite Symmetric kernel satisfying

$$\langle k(\underline{X}, \cdot), k(\underline{X}', \cdot) \rangle_{\mathbb{H}} = k(\underline{X}, \underline{X}')$$

Proof. By construction, as $k(\underline{X}, \cdot) \in \mathbb{H}$,

$$\langle k(\underline{X}, \cdot), k(\underline{X}', \cdot) \rangle_{\mathbb{H}} = k(\underline{X}, \underline{X}').$$

This implies immediately that k is symmetric.
Now for $N \in \mathbb{N}$ and any $(\underline{X}_1, \dots, \underline{X}_N) \in \mathcal{X}^N$ and any $u \in \mathbb{R}^N$:

$$\begin{aligned}
\sum_{1 \leq i,j \leq N} u^{(i)} u^{(j)} k(\underline{X}_i, \underline{X}_j) &= \sum_{1 \leq i,j \leq N} u^{(i)} u^{(j)} \left\langle k(\underline{X}_i, \cdot), k(\underline{X}_j, cdot \right\rangle_{\mathbb{H}} \\
&= \left\langle \sum_{1 \leq i \leq N} u^{(i)} k(\underline{X}_i, \cdot), \sum_{1 \leq j \leq N} u^{(j)} k(\underline{X}_j, \cdot) \right\rangle_{\mathbb{H}} \\
&= \left\| \sum_{1 \leq i \leq N} u^{(i)} k(\underline{X}_i, \cdot) \right\|_{\mathbb{H}}^2 \geq 0.
\end{aligned}$$

□

D.2. Moore-Aronsajn Theorem

Theorem D.2.1

For any PDS kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, it exists a Hilbert space $\mathbb{H} \subset \mathbb{R}^{\mathcal{X}}$ with a scalar product $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ such that

- *it exists a mapping $\phi : \mathcal{X} \to \mathbb{H}$ satisfying*

$$k(\underline{X}, \underline{X}') = \langle \phi(\underline{X}), \phi(\underline{X}) \rangle_{\mathbb{H}}$$

- *the reproducing property holds, i.e. for any $h \in \mathbb{H}$ and any $\underline{X} \in \mathcal{X}$*

$$h(\underline{X}) = \langle h, k(\underline{X}, \cdot) \rangle_{\mathbb{H}} .$$

Proof. For any x, we define $\Phi(\underline{X}) = k(\underline{X}, \cdot)$, $\Phi(\underline{X})$ is thus a function from $\mathcal{X} \to \mathbb{R}$. Now denote $\mathcal{H}$ the set of finite linear combination of $\phi(\underline{X})$. We can define a scalar product between the function by:

$$\langle \Phi(\underline{X}), \Phi(\underline{Y}) \rangle_{\mathcal{H}} = k(\underline{X}, \underline{Y}).$$

Indeed because k is a PDS kernel, all the properties of a scalar product are satisfied. Now let $f \in \mathcal{H}$, by definition $f = \sum_{i=1}^{n} \alpha_i k(\underline{X}_i, \cdot)$ and thus

$$\begin{aligned} f(\underline{X}) &= \sum_{i=1}^{n} \alpha_i k(\underline{X}_i, \underline{X}) \\ &\sum_{i=1}^{n} \alpha_i \langle k(\underline{X}_i, \cdot), k(\underline{X}, \cdot) \rangle_{\mathcal{H}} \\ &= \left\langle \sum_{i=1}^{n} \alpha_i k(\underline{X}_i, \cdot), k(\underline{X}, \cdot) \right\rangle_{\mathcal{H}} \\ &= \langle f, k(\underline{X}, \cdot) \rangle_{\mathcal{H}} . \end{aligned}$$

$\mathcal{H}$ is not a Hilbert space but only a pre-Hilbert space. It has to be completed by the Cauchy sequence process to obtain an Hilbert space $\mathbb{H}$ satisfying all the required properties. □

D.3. Kernel Construction Machinery

See **scholkopf02** for instance for more details.

Theorem D.3.1

For any function $\Psi : \mathcal{X} \to \mathbb{R}$, $k(\underline{X}, \underline{X}') = \Psi(\underline{X})\Psi(\underline{X}')$ is PDS.

Proof. k is symmetric by construction. Now for any N, and any $\underline{X}_i$ and u_i

$$\begin{aligned}\sum_{i,j} u_i u_j k(\underline{X}_i, \underline{X}_j) &= \sum_{i,j} u_i u_j \phi(\underline{X}_i)\phi(\underline{X}_j) \\ &= (\sum_i u_i \phi(\underline{X}_i))^2 \geq 0.\end{aligned}$$

□

Theorem D.3.2

For any PDS kernels k_1 and k_2, and any $\lambda \geq 0$ $k_1 + \lambda k_2$ and $\lambda k_1 k_2$ are PDS kernels.

Proof. The symmetry is a direct consequence of the symmetry of k_1 and k_2.

Now for any N, and any $\underline{X}_i$ and u_i, we have

$$\begin{aligned}\sum_{i,j} u_i u_j (k_1 + \lambda k_2)(\underline{X}_i, \underline{X}_j) &= \sum_{i,j} u_i u_j \left(k_1(\underline{X}_i, \underline{X}_j) + \lambda k_2(\underline{X}_i, \underline{X}_j)\right) \\ &= \sum_{i,j} u_i u_j k_1(\underline{X}_i, \underline{X}_j) + \lambda \sum_{i,j} u_i u_j k_2(\underline{X}_i, \underline{X}_j) \geq 0\end{aligned}$$

as a sum of two non negative term.

Now for the product

$$\sum_{i,j} u_i u_j (\lambda k_1 k_2)(\underline{X}_i, \underline{X}_j) = \lambda \sum_{i,j} u_i u_j k_1(\underline{X}_i, \underline{X}_j) k_2(\underline{X}_i, \underline{X}_j)$$

As k_1 is a PDS the matrix $K_1 = (k_1(\underline{X}_i, \underline{X}_j))$ is sdp and thus can be expressed as a product $K_1 = MM^t$ so that $k_1(\underline{X}_i, \underline{X}_j) = \sum_k M_{i,k} M_{k,j}$. We can plug this expression in the previous sum

$$\begin{aligned}&= \lambda \sum_{i,j} u_i u_j \sum_k M_{i,k} M_{k,j} k_2(\underline{X}_i, \underline{X}_j) \\ &= \lambda \sum_k \sum_{i,j} u_i M_{i,k} u_j M_{k,j} k_2(\underline{X}_i, \underline{X}_j) \geq 0\end{aligned}$$

as each term in the sum in k is non negative.

□

Theorem D.3.3

For any sequence of PDS kernels k_n converging pointwise to a kernel k, k is a PDS kernel.

Proof. The symmetry is preserved by the pointwise convergence as well as the positivity.

□

Theorem D.3.4

For any PDS kernel k such that $|k| \leq r$ and any power series $\sum_n a_n z^n$ with $a_n \geq 0$ and a convergence radius larger than r, $\sum_n a_n k^n$ is a PDS kernel.

Proof. This a direct consequence of the previous claim. □

Theorem D.3.5

For any PDS kernel k, the renormalized kernel $k'(\underline{X}, \underline{X}') = \dfrac{k(\underline{X}, \underline{X}')}{\sqrt{k(\underline{X}, \underline{X})k(\underline{X}', \underline{X}')}}$ is a PDS kernel.

Proof. As before, the symmetry is not an issue. For the positivity,

$$\sum_{i,j} u_i u_j k'(\underline{X}_i, \underline{X}_j) = \sum_{i,j} u_i u_j \frac{k(\underline{X}_i, \underline{X}_j)}{\sqrt{k(\underline{X}_i, \underline{X}_i)k(\underline{X}_j, \underline{X}_j)}}$$
$$\sum_{i,j} \frac{u_i}{\sqrt{k(\underline{X}_i, \underline{X}_i)}} \frac{u_j}{\sqrt{k(\underline{X}_j, \underline{X}_j)}} k(\underline{X}_i, \underline{X}_j) \geq 0$$

□

D.4. Mercer Representation

Theorem D.4.1

Let k be a PDS kernel and $\mathbb{H}$ its corresponding RKHS,
for any increasing function Φ and any function $L : \mathbb{R}^n \to \mathbb{R}$, the optimization problem

$$\operatorname*{argmin}_{h \in \mathbb{H}} L(h(\underline{X}_1), \dots, h(\underline{X}_n)) + \Phi(\|h\|)$$

admits only solutions of the form

$$\sum_{i=1}^{n} \alpha_i' k(\underline{X}_i, \cdot).$$

Proof. The proof is similar to the one for the non kernel setting. Assume h is a minimizer of

$$\operatorname*{argmin}_{h \in \mathbb{H}} L(h(\underline{X}_1), \dots, h(\underline{X}_n)) + \Phi(\|h\|).$$

Let $h_{\underline{X}}$ be the orthogonal projection of h on the finite dimensional space spanned by the $k(\underline{X}_i, \cdot)$. By construction, $h - h_{\underline{X}}$ is orthogonal to all the $k(\underline{X}_i, \cdot)$ and thus

$$h(X_i) = \langle h, k(X_i, \cdot) \rangle = \langle h_{\underline{X}} + h - h_{\underline{X}}, k(X_i, \cdot) \rangle = \langle h_{\underline{X}}, k(X_i, \cdot) \rangle = h_{\underline{X}}(X_i).$$

This implies that

$$\begin{aligned} L(h(\underline{X}_1), \ldots, h(\underline{X}_n)) + \Phi(\|\beta\|_2) &= L(h(\underline{X}_1), \ldots, h_{\underline{X}}(\underline{X}_n)) + \Phi(\|\beta\|_2) \\ &\geq L(h(\underline{X}_1), \ldots, h_{\underline{X}}(\underline{X}_n)) + \Phi(\|\beta_{\underline{X}}\|_2) \end{aligned}$$

where the inequality holds because $\|h\|^2 = \|h_{\underline{X}}\|^2 + \|h - h_{\underline{X}}\|^2$. The minimum is thus reached by a h in the space spanned by the $k(\underline{X}_i, \cdot)$, i.e.

$$\beta = \sum_{i=1}^{n} \alpha_i k(\underline{X}_i, \cdot).$$

□

E. Neural Networks

E.1. Perceptron

Theorem E.1.1

The perceptron algorithm converges in a finite number of steps under the linear separability assumption.

Proof. By linear separability, it exists $w^\star$ such that $Y_i\langle w^\star, X_i\rangle > 0$.

Let $C = \max_{X_i} \|C_i\|$ and $\rho = \min Y_i\langle w^\star, X_i\rangle > 0$.

Let w_t be the weight at step t. If $\min Y_i\langle w_t, X_i\rangle > 0$ then we are done.

Otherwise, let (X_i, Y_i) be the first example such that $Y_i(\langle w_t, X_i\rangle) \leq 0$ and let $w_{t+1} = w_t + \alpha Y_i X_i$. By construction,

$$\begin{aligned}\langle w^\star, w_{t+1}\rangle &= \langle w^\star, w_t\rangle + Y_i\langle w^\star, W_i\rangle \\ &\geq \langle w^\star, w_t\rangle + Y_i\langle w^\star, W_i\rangle \\ &\geq \langle w^\star, w_t\rangle + \rho \geq \langle w^\star, w_0\rangle + t\rho\end{aligned}$$

while

$$\begin{aligned}\|w_{t+1}\|^2 &= \|w_t\|^2 + \|X_i\|^2 + 2\langle w_t, Y_i X_i\rangle \\ &\leq \|w_t\|^2 + \|X_i\|^2 \leq \|w_t\|^2 + C^2 \leq \|w_0\|^2 + tC^2.\end{aligned}$$

Now $\langle w^\star, w_{t+1}\rangle \leq \|w^\star\|\|w_{t+1}\|$ so that we have

$$\langle w^\star, w_0\rangle + t\rho \leq \|w^\star\| \left(\|w_0\|^2 + tC^2\right)^{1/2}$$

which implies that such a t is upperbounded and hence the algorithm converges. □

E.2. Universal Approximation Theorem

We follow here the proof of **cybenko89**.

Definition E.2.1

An activation function σ is said to be discriminatory if for any signed regular Borel measure μ on $[0,1]^d$

$$\forall w, b \int \sigma(w^t x + b) d\mu(x) = 0 \implies \mu = 0$$

Lemma E.2.2

If σ is discriminatory then the set of single hidden layer neural networks is dense in the set of continuous function of $[0,1]^d$.

Proof. We first notice that the set $\mathcal{N}$ of single hidden layer neural network is stable by multiplication by a constant and addition and thus a sub-space of the set of continuous function, provided σ is continuous.

Assume now that $\mathcal{N}$ is not dense, the Hahn-Banach theorem implies that it exists a continuous linear function L defined on the set of continuous functions such that $L(f) = 0$ if $f \in \mathcal{N}$ and $L \neq 0$. By the Riesz representation theorem, this function L can be represented as

$$L(f) = \int f(x) d\mu(x)$$

with μ a signed regular Borel measure. Applying this definition to a single sigmoid, one deduces

$$\forall w, b \int \sigma(w^t x + b) d\mu(x) = 0$$

and thus $\mu = 0$ contradicting $L \neq 0$.

□

Lemma E.2.3

Any bounded continuous activation function satisfying

$$\lim_{\lambda \to +\infty} \sigma(\lambda(w^t x + b) + c) = \begin{cases} \sigma(\infty) & \text{if } w^t x + b > 0 \\ \sigma(c) & \text{if } w^t x + b = 0 \\ \sigma(-\infty) & \text{if } w^t x + b < 0 \end{cases}$$

with $\sigma(\infty)$ and $\sigma(-\infty)$ two different finite real numbers is discriminatory.

Proof. Fix $\boldsymbol{w}, b$ and c and let $\gamma(x) = \lim_{\lambda \to +\infty} \sigma(\lambda(\langle \boldsymbol{w}, x \rangle + b) + c)$, By the dominated convergence theorem, if

$$\forall \boldsymbol{w}', b" \int \sigma(\langle \boldsymbol{w}', x \rangle + b') d\mu(x) = 0$$

then

$$\int \gamma(x) d\mu(x) = 0.$$

One deduces thus that

$$\int \gamma(x) d\mu(x) = \sigma(\infty)\mu(\{y; \langle \boldsymbol{w}, y \rangle + b > 0\}) + \sigma(c)\mu(\{y; \langle \boldsymbol{w}, y \rangle + b = 0\}) + \sigma(-\infty)\mu(\{y; \langle \boldsymbol{w}, y \rangle + b < 0\}$$

As $\sigma(c)$ can be chosen arbitrarily between $\sigma(-\infty)$ and $\sigma(\infty)$ that are different, one verifies easily that this implies

$$\mu(\{y; \langle \boldsymbol{w}, y\rangle + b = 0\}) = 0$$

With b large enough, we obtain

$$\sigma(\infty)\mu\{y \in [0,1]^d = 0$$

from which we derive

$$\sigma(\infty)\left(\mu(\{y; \langle \boldsymbol{w}, y\rangle + b > 0\}) + \mu(\{y; \langle \boldsymbol{w}, y\rangle + b < 0\})\right) = 0,$$

while with b small enough, we have

$$\sigma(-\infty)\mu\{y \in [0,1]^d = 0$$

from which we derive

$$\sigma(-\infty)\left(\mu(\{y; \langle \boldsymbol{w}, y\rangle + b > 0\}) + \mu(\{y; \langle \boldsymbol{w}, y\rangle + b < 0\})\right) = 0.$$

As $\sigma(-\infty)$ and $\sigma(\infty)$ are different, this implies that

$$\mu(\{y; \langle \boldsymbol{w}, y\rangle + b > 0\}) + \mu(\{y; \langle \boldsymbol{w}, y\rangle + b < 0\}) = 0$$

and thus finally

$$\mu(\{y; \langle \boldsymbol{w}, y\rangle + b > 0\}) = \mu(\{y; \langle \boldsymbol{w}, y\rangle + b < 0\}) = 0.$$

This implies in particular that

$$\mu(\{y; b' > \langle \boldsymbol{w}, y\rangle \leq b\}) = 0.$$

Using the fact that the set of one dimensional piecewise constant function is dense in the set of continuous function, one deduces immediately that any continuous function $g(y) = f(\langle \boldsymbol{w}, y\rangle)$ verify

$$\int g(y) d\mu(y) = 0.$$

This is true in particular for any function of the Fourier basis. This implies thus the result for any function as the space spanned by the Fourier basis is dense in the set of continuous function of $[0,1]^d$ □

Lemma E.2.4

The RELU function is discriminatory.

Proof. An easy way to obtain the result is to notice that

$$\sigma(t) = RELU(t) - RELU(t-1)$$

satisfy the assumptions of Lemma E.2.3 and thus we can use the previous result (with twice the number of neurons). □

F. Concentration Inequalities

F.1. Hoeffding

Theorem F.1.1

Let Z_i be a sequence of ind. centered r.v. supported in $[a_i, b_i]$ then

$$\mathbb{P}\left(\sum_{i=1}^{n} Z_i \geq \epsilon\right) \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}}$$

Proof adapted from **shalev-shwartz14**. We rely on the following lemma

Lemma F.1.2

$$\mathbb{E}\left[e^{\lambda \sum_{i=1}^{n} Z_i}\right] \leq e^{\frac{\lambda^2 \sum_{i=1}^{n}(b_i-a_i)^2}{8}}.$$

then an optimization in λ leads to

$$\frac{\prod_{i=1}^{n} \mathbb{E}\left[e^{\lambda Z_i}\right]}{e^{\lambda\epsilon}} \leq e^{\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i-a_i)^2 - \lambda\epsilon}$$

is minimal for $\lambda = 4\epsilon/(\sum_{i=1}^{n}(b_i - a_i)^2)$ for which

$$e^{\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i-a_i)^2 - \lambda\epsilon} = e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}}.$$

□

Proof of Lemma F.1.2. Exponential moment function:

$$\begin{aligned}
\Psi_Z(\lambda) &= \log \mathbb{E}\left[e^{\lambda Z}\right] & \Psi_Z(0) &= 0 \\
\Psi'_Z(\lambda) &= \frac{\mathbb{E}\left[Ze^{\lambda Z}\right]}{\mathbb{E}[e^{\lambda Z}]} & \Psi'_Z(0) &= \frac{\mathbb{E}[Z]}{\Psi_Z(0)} = 0 \\
\Psi''_Z(\lambda) &= \frac{\mathbb{E}\left[Z^2e^{\lambda Z}\right]}{\mathbb{E}[e^{\lambda Z}]} - \frac{\left(\mathbb{E}\left[Ze^{\lambda Z}\right]\right)^2}{\left(\mathbb{E}[e^{\lambda Z}]\right)^2} \\
&= \mathbb{E}\left[Z^2e^{\lambda Z-\psi(\lambda)}\right] - \mathbb{E}\left[Ze^{\lambda Z-\psi(\lambda)}\right] \\
&= \mathbb{V}\mathrm{ar}\left[Z'\right]
\end{aligned}$$

with Z' a random variable with density $e^{\lambda Z-\psi(\lambda)}$ with respect to dZ.

Now as $Z' \in [a,b]$,

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[Z'\right] &= \mathbb{E}\left[(Z' - \mathbb{E}\left[Z'\right])^2\right] \\
&\leq \mathbb{E}\left[(Z' - (a+b)/2)^2\right] \leq (b-a)^2/4
\end{aligned}$$

As $\Psi_Z(0) = 0$, $\Psi'_Z(0) = 0$ and $\Psi''_Z(\lambda) \leq (b-a)^2/4$, Taylor formula ensures that $\exists\theta \in [0,\lambda]$ such that

$$\Psi_Z(\lambda) = \frac{1}{2}\Psi''_Z(\theta)\lambda^2 \leq \frac{(b-a)^2}{8}\lambda^2$$

This yields

$$\mathbb{E}\left[e^{\lambda Z}\right] \leq e^{\frac{(b-a)^2}{8}\lambda^2}$$

□

F.2. McDiarmid Inequality

Theorem F.2.1

If g is a bounded difference function and $\underline{X}_i$ are independent random variables then

$$\begin{aligned}
\mathbb{P}(g(\underline{X}_1,\dots,\underline{X}_n) - \mathbb{E}[g(\underline{X}_1,\dots,\underline{X}_n)] \geq \epsilon) &\leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}} \\
\mathbb{P}(\mathbb{E}[g(\underline{X}_1,\dots,\underline{X}_n)] - g(\underline{X}_1,\dots,\underline{X}_n) \geq \epsilon) &\leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}}
\end{aligned}$$

Proof. Let $g_i = \mathbb{E}[g(\underline{X}_1,\dots,\underline{X}_n)|\underline{X}_1,\dots,\underline{X}_i]$ so that $g_0 = \mathbb{E}[g(\underline{X}_1,\dots,\underline{X}_n)]$ and $g_n = g(\underline{X}_1,\dots,\underline{X}_n)$.

By construction,

$$g(\underline{X}_1, \dots, \underline{X}_n) - \mathbb{E}[g(\underline{X}_1, \dots, \underline{X}_n)] = \sum_{i=0}^{n-1} g_{n-i} - g_{n-i-1}.$$

Now

$$\begin{aligned}
&\mathbb{E}\left[e^{\lambda(g(\underline{X}_1,\dots,\underline{X}_n)-\mathbb{E}[g(\underline{X}_1,\dots,\underline{X}_n)])}\right] \\
&= \mathbb{E}\left[e^{\lambda(\sum_{i=0}^{n-1} g_{n-i}-g_{n-i-1})}\right] \\
&= \mathbb{E}\left[\prod_{i=0}^{n-1} e^{\lambda(g_{n-i}-g_{n-i-1})}\right]
\end{aligned}$$

by conditioning we have

$$= \mathbb{E}\left[\mathbb{E}\left[\prod_{i=0}^{n-1} e^{\lambda(g_{n-i}-g_{n-i-1})} | \underline{X}_2, \dots \underline{X}_n\right]\right]$$

and using the fact that, for $i \geq 1$, g_i is constant conditionally to $\underline{X}_2, \dots \underline{X}_n$

$$= \mathbb{E}\left[\mathbb{E}\left[e^{\lambda(g_1-g_0)} | \underline{X}_2, \dots \underline{X}_n\right] \prod_{i=1}^{n-1} e^{\lambda(g_{i+1}-g_i)}\right]$$

Now $g_1 - g_0$ is by construction a centered random variable bounded in absolute value by c_n and thus

$$\leq e^{\frac{\lambda^2 c_1^2}{8}} \mathbb{E}\left[\prod_{i=1}^{n-1} e^{\lambda(g_i-g_{i+1})}\right]$$

Reusing the same technique recursively, we obtain

$$\mathbb{E}\left[e^{\lambda(g(\underline{X}_1,\dots,\underline{X}_n)-\mathbb{E}[g(\underline{X}_1,\dots,\underline{X}_n)])}\right] \leq e^{-\frac{\lambda^2 \sum_{i=1}^n c_n^2}{8}}$$

Optimizing the corresponding Chernov bound in λ as in Hoeffding proof yields the result. □

Theorem F.2.2

Let $\mathcal{H}$ be a set of n-tuple of functions of $Z_i = (\underline{X}, Y_i)$ and let σ_i be a sequence of i.i.d. random symmetric Bernoulli variables (Rademacher variables)

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (h_i(Z_i) - \mathbb{E}[h_i(Z_i)])\right] \leq 2\mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h_i(Z_i)\right]$$

Proof. Let $\mathcal{H}$ be a set of n-tuple of functions of $Z_i = (\underline{X}, Y_i)$, we will prove that

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(h_i(Z_i) - \mathbb{E}[h_i(Z_i)])\right] \leq 2\mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i h_i(Z_i)\right]$$

Now, the doubling trick consists in introducing a second set of samples Z_i' with the same distribution:

$$\begin{aligned}\mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(h_i(Z_i) - \mathbb{E}[h_i(Z_i)])\right] &= \mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(h_i(Z_i) - \mathbb{E}\left[h_i(Z_i')\right])\right] \\ &= \mathbb{E}_Z\left[\sup_{h\in\mathcal{H}} \mathbb{E}_{Z'}\left[\frac{1}{n}\sum_{i=1}^{n}(h_i(Z_i) - h_i(Z_i'))\right]\right]\end{aligned}$$

We may now now upperbound this term by exchanging the sup and the expectation to obtain

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(h_i(Z_i) - \mathbb{E}[h_i(Z_i)])\right] \leq \mathbb{E}_{Z,Z'}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(h_i(Z_i) - h_i(Z_i'))\right]$$

By construction $h_i(Z_i) - h_i(Z_i')$ is a symmetric random variable and has thus the same law than $\sigma_i(h_i(Z_i) - h_i(Z_i'))$ where σ_i is a sequence of i.i.d. Rademacher variable.

Thus

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(h_i(Z_i) - \mathbb{E}[h_i(Z_i)])\right] \leq \mathbb{E}_{Z,Z',\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(h_i(Z_i) - h_i(Z_i')\right)\right]$$

Now we can split the sum in two and obtain

$$\begin{aligned}\mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}h_i(Z_i) - \mathbb{E}[h_i(Z_i)]\right] &\leq \mathbb{E}_{Z,Z',\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i h_i(Z_i)\right] \\ &\quad + \mathbb{E}_{Z,Z',\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i(-h_i(Z_i'))\right] \\ &\leq \mathbb{E}_{Z,\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i h_i(Z_i)\right] \\ &\quad + \mathbb{E}_{Z,\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(-\sigma_i)h_i(Z_i)\right]\end{aligned}$$

which yields by symmetry of σ_i

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}h_i(Z_i) - \mathbb{E}[h_i(Z_i)]\right] \leq 2\mathbb{E}_{Z,\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i h_i(Z_i)\right]$$

□

Theorem F.2.3

If B is finite and such that $\forall b \in B, \frac{1}{n}\|b\|_2^2 \leq M^2$, then

$$R_n(B) = \mathbb{E}\left[\sup_{b\in B} \frac{1}{n}\sum_{i=1}^{n} \sigma_i b_i\right] \leq \sqrt{\frac{2M^2 \log |B|}{n}}$$

Proof. By Jensen inequality

$$\begin{aligned} e^{\lambda \mathbb{E}\left[\sup_{b\in B} \frac{1}{n}\sum_{i=1}^{n} \sigma_i b_i\right]} &\leq \mathbb{E}\left[e^{\lambda \sup_{b\in B} \frac{1}{n}\sum_{i=1}^{n} \sigma_i b_i}\right] \\ &\leq \sum_{b\in B} \mathbb{E}\left[e^{\lambda \frac{1}{n}\sum_{i=1}^{n} \sigma_i b_i}\right] \end{aligned}$$

We can now use Hoeffding inequality with $c_i = 2b_i/n$ to obtain

$$\begin{aligned} e^{\lambda \mathbb{E}\left[\sup_{b\in B} \frac{1}{n}\sum_{i=1}^{n} \sigma_i b_i\right]} &\leq \sum_{b\in B} e^{\frac{4\lambda^2 \sum_{i=1}^{n} b_i^2}{8n^2}} \\ &\leq \sum_{b\in B} e^{\frac{M^2\lambda^2}{2n}} = |B| e^{\frac{M^2\lambda^2}{2n}} \end{aligned}$$

Thus

$$\mathbb{E}\left[\sup_{b\in B} \frac{1}{n}\sum_{i=1}^{n} \sigma_i b_i\right] \leq \frac{\log B}{\lambda} + \frac{M^2\lambda}{2n}$$

The optimal value is given by $\lambda = \sqrt{\frac{2n \log B}{M^2}}$ yielding

$$\mathbb{E}\left[\sup_{b\in B} \frac{1}{n}\sum_{i=1}^{n} \sigma_i b_i\right] \leq \sqrt{\frac{2 \log |B_n(\mathcal{S})|}{n}}$$

□