Book of Proofs

E. Le Pennec

October 18, 2021

In this document, you will ultimately find all the proofs of the results given in the lecture. For the time being, 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!

1 Statistical Setting

1.1 Bayes Predictor

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

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

Proof. We start by noticing that

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

so that we can focus on

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

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

By definition,

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

which implies

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

.

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

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

Proof. We start by noticing that

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

so that we can focus on

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

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

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

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

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

1.2 Training Error 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 3.

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

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

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

□

2 Cross Validation

2.1 Leave One Out Formula

Claim 4. *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 $\underline{X}^{\Phi}_{(n)-i}{}^{\top}\underline{X}^{\Phi}_{(n)-i} = \mathbb{X}_{(n)}{}^{\Phi\top}\mathbb{X}^{\Phi}_{(n)} - \underline{X}_i^{\Phi}\underline{X}_i{}^{\top}$ and $\underline{X}^{\Phi}_{(n)-i}{}^{\top}\underline{Y}_{(n)-i} = \mathbb{X}_{(n)}{}^{\Phi\top}\mathbb{Y}_{(n)} - \underline{X}_i^{\Phi}Y_i$

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:

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

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

$$\begin{aligned}
&= \hat{f}(\underline{X}_i) + \frac{h_{ii}}{1 - h_{ii}}\hat{f}(\underline{X}_i) - h_{ii}Y_i - \frac{h_{ii}^2}{Y}{}_i \\
\hat{f}^{-i}(\underline{X}_i) &= \frac{\hat{f}(\underline{X}_i) - h_{ii}Y_i}{1 - h_{ii}}
\end{aligned}$$

□

2.2 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 multiclass setting.

Claim 5. *The minimizer of $\mathbb{E}\left[(Y, f(\underline{X}))\right]$ is given by*

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

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

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

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

3 Probabilistic Point of View

3.1 Classification Risk Analysis with a Probabilistic Point of View

Claim 6. *If* $\widehat{f} = \operatorname{sign}(2\widehat{p}_{+1} - 1)$ *then*

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

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

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

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

Step 2:

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

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

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

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 $1/2$

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

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

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

□

3.2 Logistic Likelihood and Convexity

Claim 7. *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.

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

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. ☐

4 Optimization Point of View

4.1 Classical Convexification

Claim 8. *The following three losses*

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

satisfy

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

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

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

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

$$
\begin{aligned}
l'(z) &= -e^{-z} \\
l''(z) &= e^{-z}.
\end{aligned}
$$

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 $\ell'(Y, f(\underline{X})) = l(Yf(\vec{(X)})) \geq 1$ when $Yf(\vec{(X)}) \leq 0$ and $\ell'(Y, f(\underline{X})) \geq 0$ otherwise. We obtain thus that $\ell(Y, f(\underline{X})) \geq \ell^{0/1}(Y, f(\underline{X}))$.

□

4.2 Classification Risk Analysis with an Optimization Point of View

Claim 9. *The minimizer of*

$$
\mathbb{E}\left[\ell'(Y, f(\underline{X}))\right] = \mathbb{E}\left[l(Yf(\underline{X}))\right]
$$

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

Furthermore it exists a convex function Ψ *such that*

$$
\begin{aligned}
&\Psi\left(\mathbb{E}\left[\ell^{0/1}(Y, \operatorname{sign}(f(\underline{X}))\right] - \mathbb{E}\left[\ell^{0/1}(Y, f^\star(\underline{X})\right]\right) \\
&\leq \mathbb{E}\left[\ell'(Y, f(\underline{X})\right] - \mathbb{E}\left[\ell'(Y, f^\star(\underline{X}))\right]
\end{aligned}
$$

Proof. By definition,

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

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 $\operatorname{sign}(\tilde{f}) = \operatorname{sign}(2\eta - 1)$ i.e. the minimizer of $\mathbb{E}\left[l(yf)|\underline{X}\right]$ is $f^*(\underline{X}) = \operatorname{sign}(2\eta(\underline{X}) - 1)$

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 minimimum over the *wrong sign classifiers*, $inf_{f, f(2\eta - 1) < 0} H(f, \eta) = l(0)$.

Indeed,

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

Furthermore,

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

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

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

Using Jensen inequality, we derive

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

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

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

Using the property of the wrong sign classifiers

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

□

4.3 SVM, distance and norm of β

Claim 10. *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. □

4.4 SVM and Hinge Loss

Claim 11. *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

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

yeilds 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,

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

hence the result. □

4.5 Constrained Optimization, Lagrangian and Dual

Claim 12.

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

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

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

□

Claim 13.

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

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 $\mathcal{L}(x, \lambda, \mu) \leq f(x)$. □

4.6 Duality, weak, strong and Slater's condition

Claim 14. *Weak duality:*

$$\begin{aligned} q^* &\leq p^* \\ \max_{\lambda \in \mathbb{R}^p,\ \mu \in (\mathbb{R}^+)^q} \min_x \mathcal{L}(x, \lambda, \mu) &\leq \min_x \max_{\lambda \in \mathbb{R}^p,\ \mu \in (\mathbb{R}^+)^q} \mathcal{L}(x, \lambda, \mu)\end{aligned}$$

Proof. This is a direct consequence of Claim 13. □

Claim 15. *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 Boyd and Vandenberghe 2004. □

4.7 Karush-Kuhn-Tucker Claim

Claim 16. *If f is convex, h_j affine and g_i convex, all are differentiable and strong duality holds then x^* is a solution of the primal problem if and only if the KKT condition*

- *Stationarity:*

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

- *Primal admissibility:*

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

- *Dual admissibility:*

$$\mu_i \geq 0$$

- *Complementary slackness:*

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

holds.

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

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

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

Let x^* is a solution of the primal problem and (λ^*, μ^*) be a solution of the dual. If the strong duality holds:

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

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^* is a minimizer of $\mathcal{L}(x, \lambda^*, \mu^*)$. We obtain thus the stationarity condition:

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

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

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

The complementary slackness condition is obtained by noticing that

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

which implies

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

hence the result.

□

4.8 SVM, KKT and Dual

Claim 17. *For the SVM, the KKT conditions are given by*

- *Stationarity:*

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

- *Primal and dual admissibility:*

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

- *Complementary slackness:*

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

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 18. *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. $\square$

Claim 19. *The solution of the SVM satisfy*

- $\beta^* = \sum_i \alpha_i Y_i \underline{X}_i$ *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).*
- $\beta^{(0)*} = Y_i - \underline{X}_i^\top \beta^*$ *for any support vector with* $0 < \alpha_i < C$.

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

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

If we use $\nabla_{s_i} \mathcal{L}(\beta^*, \beta^{(0)*}, s, \alpha, \mu) = 0$, 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 $Y_i(\underline{X}_i^\top \beta^* + \beta^{(0)*}) = 1 - s_i$ thus

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

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

Claim 20. *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

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

□

4.9 Mercer Representation Claim

Claim 21. *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 $\beta^* = \sum_{i=1}^n \alpha_i' \underline{X}_i$.

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.

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

□

4.10 Mercer Kernel Claim

Claim 22. *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. □

4.11 Kernel Construction Machinery

Claim 23. *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

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

□

Claim 24. *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.

$\square$

Claim 25. *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. $\square$

Claim 26. *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, $\displaystyle\sum_n a_n k^n$ is a PDS kernel.*

Proof. This a direct consequence of the previous claim. $\square$

Claim 27. *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,

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

$\square$

4.12 Mercer Representation Claim

Claim 28. *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

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

This implies that

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

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.

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

$\square$

4.13 SVM and VC dimension

See Mohri, Rostamizadeh, and Talwalkar 2012 as the VC dimension will only be defined later.

5 Optimization

Most of the results can be found in Bubeck 2015.

5.1 Linear Predictor, Gradient and Hessian

Claim 29.
- *Gradient:*

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

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

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

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

5.2 Exhaustive Search

Claim 30.
- *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: $N_\epsilon = O\left((C\sqrt{d}/\epsilon)^d\right)$*

5.3 L Smoothness

Claim 31. *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

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

First implication
Taking the norm of both sides of (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 (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 32. *F is L-smooth in the linear regression and the logistic regression cases.*

5.4 Convergence of GD

Claim 33. *Let $G : \mathbb{R}^d \to \mathbb{R}$ be a 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 Lemma 7. □

Claim 34. *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 35. *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 36. *Let $G : \mathbb{R}^d \to \mathbb{R}$ be a 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 consequenc of Lemma 10. □

Claim 37. *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 Lemma 14 □

5.5 Proximal Descent

Claim 38.

- $R(\boldsymbol{w}) = \mathbf{1}_\Omega(\boldsymbol{w})$*:* $\operatorname{prox}_\gamma R(\boldsymbol{w}') = P_\Omega(\boldsymbol{w}')$
- $R(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2$*:* $\operatorname{prox}_\gamma R(\boldsymbol{w}') = \frac{1}{1+\gamma}\boldsymbol{w}$.
- $R(\boldsymbol{w}) = \|\boldsymbol{w}\|_1$*:* $\operatorname{prox}_\gamma R(\boldsymbol{w}') = T_\gamma(\boldsymbol{w}')$ *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 minimzed 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 iff $\boldsymbol{w} = \frac{1}{1+\gamma}\boldsymbol{w}'$, 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 analyse 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

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

or equivalently

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

□

Claim 39. • *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:*

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

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

Proof. Those are consequences of Lemma 4, Lemma 9 and Lemma 14. □

5.6 Coordinate Descent

Claim 40. *If G is continuously differentiable and strictly convex, then exact coordinate descent converges to a minimum.*

Claim 41. *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 $i_1, i_2, \dots$ chosen at random: i.i.d and uniform distribution in $\{1, \dots, d\}$. Then

$$\begin{aligned}&\mathbb{E}\left[G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^*)\right]\\ &\quad\le \frac{d}{d+k}\Big(\big(1-\frac{1}{d}\big)(G(\boldsymbol{w}^{[0]}) - G(\boldsymbol{w}^*)) + \frac{1}{2}\left\|\boldsymbol{w}^{[0]} - \boldsymbol{w}^*\right\|_L^2\Big),\end{aligned}$$

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

5.7 Gradient Descent Acceleration

Claim 42. *Assume that G is a L-smooth, convex function whose minimum is reached at $\boldsymbol{w}^\star$. Then, if $\beta^{[k]} = (k-1)/(k+2)$,*

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

Proof. See Lemma 13 □

Claim 43.

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

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

Proof. The proof combines ideas of Lemma 9 and Lemma 13. It is left as an exercise or can be found in Beck 2017. □

Claim 44.

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

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

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

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

Proof. The proof is quite technical and can be found in Nesterov 2018. □

5.8 Stochastic Gradient Descent

Claim 45. • *With* $\alpha^{[k]} = 2R/(b\sqrt{k})$

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

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

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

Proof. Those are consequences of Lemma 17. □

5.9 Lemma and more

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

The proximal gradient descent algorithm is given by

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

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

Lemma 1. *For any differentiable 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*

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

then for any z

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

.

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

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

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 substracting $F(z)$ 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. □

Lemma 2. *For any convex function F and $\boldsymbol{w}$, if we let*

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

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 $\alpha\mu \leq 1$. □

Lemma 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

$$\begin{aligned} G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^{[k]}) &\leq -\frac{1}{2\alpha^{[k]}} \|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]}\|^2 \\ G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^{*}) &\leq \frac{1}{2\alpha^{[k]}} (1 - \alpha^{[k]}\mu) \|\boldsymbol{w}^{[k]} - \boldsymbol{w}^{*}\|^2 - \frac{1}{2\alpha^{[k]}} \|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{*}\|^2 \end{aligned}$$

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

Proof. As

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

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

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

□

Lemma 4. *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

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

Proof. Lemma 3 yields

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

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

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

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

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

which implies

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

□

Lemma 5. *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 6. *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 a 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

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

□

Lemma 7. *If F is L-smooth and we use the Gradient Descent algorithm with $\alpha^{[k]} = \alpha \leq 1/L$ then*

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

Proof. We combine Lemma 4 and Lemma 5 to obtain

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

□

Lemma 8. *If F is L-smooth and we use the Gradient Descent algorithm with $\alpha^{[k]}$ obtained by backtracking then*

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

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

Proof. This is the result of Lemma 4 and Lemma 6. □

Lemma 9. *If F is L-smooth and μ strictly 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

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

Proof. Acccording to Lemma 3, we have

$$\begin{aligned} G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^{[k]}) &\leq -\frac{1}{2\alpha^{[k]}}\|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^{[k]}\|^2 \\ G(\boldsymbol{w}^{[k+1]}) - G(\boldsymbol{w}^*) &\leq \frac{1}{2\alpha^{[k]}}(1 - \alpha^{[k]}\mu)\|\boldsymbol{w}^{[k]} - \boldsymbol{w}^*\|^2 - \frac{1}{2\alpha^{[k]}}\|\boldsymbol{w}^{[k+1]} - \boldsymbol{w}^*\|^2 \end{aligned}$$

The second inequality implies immediately

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

so that

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

Plugging this bound in the same inequality we have used yields

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

□

Lemma 10. *If F is L-smooth and μ stricly convex and we use the Gradient Descent algorithm with with $\alpha^{[k]}$ obtained by backtracking then*

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

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 6 and Lemma 9. □

Lemma 11. *If F is L-smooth and μ stricly convex and we use the Gradient Descent algorithm with $\alpha^{[k]} = \alpha \leq 1/L$ then*

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

Proof. This is a direct consequence of Lemma 5 and Lemma 9. □

Lemma 12. *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

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

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

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

By convexity of G,

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

Now

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

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

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

Combining the two previous bounds yields

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

and equivalently

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

provided

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

If this holds, one has

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

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

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

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

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

We deduce thus the following bound

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

Defining everything in term 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

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

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

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

and

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

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

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

□

Lemma 13. *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 then*

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

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

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

□

Lemma 14. *If F is convex such that the sub gradient δ_F can be bounded, $\|\delta_F\|^2 \leq B^2$, $\|\boldsymbol{w}^{[k]} - \boldsymbol{w}^*\| \leq r^2$ then*

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

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 $\boldsymbol{w}^{[k]} \in C$,

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

this implies

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

Summing those bounds along k yields

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

We deduce thus that

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

that is

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

Along the same line, we have simultaneously

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

and thus

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

and thus using the convexity of F

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

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

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

□

Lemma 15. *If F is convex such that the sub gradient δ_F can be bounded, $\|\delta_F\|^2 \leq B^2$, $\|\boldsymbol{w}^{[k]} - \boldsymbol{w}^*\| \leq r^2$ then for $\alpha^{[k]} = \alpha_0/\sqrt{k}$ with $\alpha_0 = r/(\sqrt{2}B)$, we have*

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

and

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

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

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

or rather

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

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

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

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

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

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

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

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

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

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

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

□

Lemma 16. *If F is μ strongly convex and $\|\nabla F\|^2 \leq B^2$ then for $\alpha^{[k]} = \frac{\alpha_0}{k}$ with $\alpha_0 \geq \frac{2}{\mu}$*

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

and

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

Proof. Using the strong convexity of F

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

which implies

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

We can now sum those inequalities

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

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

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

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

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

By convexity of F

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

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

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

$\square$

Lemma 17. *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 $\|\boldsymbol{w}^{[k]} - \boldsymbol{w}^*\| \leq r^2$ then for $\alpha^{[k]} = \alpha_0/\sqrt{k}$ with $\alpha_0 = r/(\sqrt{2}B)$, we have*

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

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

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

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

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

which implies

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

and thus

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

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

□

References

Beck, A. (2017). *First-Order Methods for Optimization*. SIAM.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. Cambridge University Press.

Bubeck, S. (2015). *Convex Optimization: Algorithms and Complexity*. Now Publisher.

Mohri, M., A. Rostamizadeh, and A. Talwalkar (2012). *Foundations of Machine Learning*. MIT Press.

Nesterov, Y. (2018). *Lectures on Convex Otimization, 2nd edition*. Springer.