1 Statistical Setting

1.1 Bayes Predictor

Claim 1. The minimizer of $E \ell_{0/1}(Y, f(X))$ is given by

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

Proof. We start by noticing that

$$\arg\min_{f \in F} E \ell(Y, f(X)) = \arg\min_{f \in F} E \mathbb{Y} X \ell(Y, f(X))$$

so that we can focus on

$$E \mathbb{Y} X \ell(Y, f(X))$$

where $f(X)$ is constant.

By definition,

$$E \mathbb{Y} X \ell(Y, f(X)) = \mathbb{P}(Y = 1|X) \ell(1, f(X)) + \mathbb{P}(Y = -1|X) \ell(-1, f(X))$$

$$= \begin{cases} \mathbb{P}(Y = 1|X) & \text{if } f(X) = -1 \\ \mathbb{P}(Y = -1|X) & \text{if } f(X) = 1 \end{cases}$$

which implies

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

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

\[\square\]
Claim 2. The minimizer of $E \left[ \ell^2(Y, f(X)) \right]$ is given by

$$f^*(X) = E [Y|X]$$

Proof. We start by noticing that

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

so that we can focus on

$$E_{Y|X} [\ell(Y, f(X))] = E_{Y|X} [(Y - f(X))^2]$$

where $f(X)$ is constant.

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

$$E_{Y|X} [\ell(Y, f(X))] = E_{Y|X} [(Y - f(X))^2]$$

$$= E_{Y|X} [(Y - E [Y|X] + E [Y|X] - f(X))^2]$$

$$= E_{Y|X} [(Y - E [Y|X])^2 + 2E_{Y|X} [(Y - E [Y|X]) (E [Y|X] - f(X))] + (E [Y|X] - f(X))^2]$$

which is thus minimized by $f^*(X) = E [Y|X]$.

1.2 Training Error Optimism

Let

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i))$$

and

$$\hat{f}_S = \arg \min_{f \in S} R_n(f)$$

Claim 3.

$$R_n(\hat{f}_S) \leq R_n(f^*_S) \quad \text{and} \quad E \left[ R_n(\hat{f}_S) \right] \leq R(f^*_S)$$

Proof. The first part is nothing but the definition of $\hat{f}_S$ combined with the fact that $f^*_S$ also belongs to $S$.

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

$$E \left[ R_n \right] = E \left[ \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)) \right] = E \left[ \ell(Y, f(X)) \right] = R(f)$$

\[ \square \]
2 Cross Validation

2.1 Leave One Out Formula

Claim 4. For the least squares linear regression,

\[ \hat{f}^{-i}(X_i) = \hat{f}(X_i) - h_{ii}Y_i \]

with \( h_{ii} \) the \( i \)th diagonal coefficient of the hat (projection) matrix.

Proof. By construction,

\[ \hat{f}^{-i}(X_i) = X_i^* \beta^{-i} = X_i^T (X_{(n)-i}^* X_{(n)-i}^*)^{-1} X_{(n)-i}^* Y_{(n)-i} \]

Now \( X_{(n)-i}^* X_{(n)-i} = X_{(n)}^* X_{(n)} - X_i^* X_i^T \) and \( X_{(n)-i}^* Y_{(n)-i} = X_{(n)}^* Y_{(n)} - X_i^* Y_i \)

Using \((M + uu^T)^{-1} = M^{-1} - \frac{M^{-1} uv^T M^{-1}}{1 + u^T M^{-1} v} \) with \( M = X_{(n)}^T X_{(n)} \), \( u = -v = X_i \) yields:

\[ \hat{f}^{-i}(X_i) = X_i^* (M^{-1} + \frac{M^{-1} X_i^* X_i^T M^{-1}}{1 - X_i^* M^{-1} X_i^*}) (X_{(n)}^* Y_{(n)} - X_i^* Y_i) \]

using \( h_{ii} = X_i^* M^{-1} X_i^* \)

\[ = \hat{f}(X_i) + \frac{h_{ii}}{1 - h_{ii}} \hat{f}(X_i) - h_{ii}Y_i - h_{ii}^2 \]

\[ \hat{f}^{-i}(X_i) = \frac{\hat{f}(X_i) - h_{ii}Y_i}{1 - h_{ii}} \]

\[ \]

2.2 Weighted Loss and Bayes Estimator

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

Claim 5. The minimizer of \( E[(Y, f(X))] \) is given by

\[ f^*(X) = \arg \max_k C(k)P(Y = k|X) \]

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

\[ E_{Y|X} [\ell(Y, f(X))] = \sum_k C(k) \ell^{0/1}(k, f(X))P(Y = k|X) \]

\[ = \sum_{k \neq f(X)} C(k)P(Y = k|X) \]

\[ = -C(f(X))P(Y = f(X)|X) + \sum_k C(k)P(Y = k|X) \]

which is minimized by taking \( f(X) \) equal to the \( k \) with the largest \( C(k)P(Y = k|X) \). 

\[ \]
3 Probabilistic Point of View

3.1 Classification Risk Analysis with a Probabilistic Point of View

Claim 6. If $\hat{f} = \text{sign}(2\hat{p} + 1)$ then

$$
E \left[ l^{0,1}(Y, \hat{f}(X)) \right] - E \left[ l^{0,1}(Y, f^*(X)) \right] \\
\leq E \left[ \|Y - \hat{Y}\|_1 \right] \\
\leq \left( E \left[ 2 KL(Y, \hat{Y}) \right] \right)^{1/2}
$$

Proof. Let us denote $p_1(X) = P(Y = 1|X)$. Let $\hat{f}(X) = \text{sign}(2\hat{p} - 1)$.

Step 1: Let $\tilde{f}(X) = \text{sign}(2p_1(X) - 1)$

$$
E \left[ l^{0,1}(Y, \tilde{f}(X)) \right] = E_X \left[ p_1(X) 1_{f(X)=-1} + (1 - p_1(X)) 1_{f(X)=1} \right] \\
= E_X \left[ (1 - p_1(X)) + (2p_1(X) - 1) 1_{f(X)=-1} \right]
$$

Step 2:

$$
E \left[ l^{0,1}(Y, \hat{f}(X)) \right] - E \left[ l^{0,1}(Y, f^*(X)) \right] \\
= E_X \left[ (2p_1(X) - 1)(1_{f(X)=-1} - 1_{f^*(X)=1}) \right]
$$

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

$$
= E_X \left[ 2p_1(X) - 1 |1_{f^*(X)\neq f(X)} \right]
$$

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

$$
\leq 2E_X \left[ |p_1(X) - \hat{p}(X)| \right] = E_X \left[ |p(X) - \hat{p}(X)| \right]
$$

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

$$
\leq E_X \left[ \sqrt{2KL(p(X), \hat{p}(X))} \right] \leq \left( E_X \left[ 2KL(p(X), \hat{p}(X)) \right] \right)^{1/2}
$$

3.2 Logistic Likelihood and Convexity

Claim 7. The maximum likelihood estimate of the logistic model is given by

$$
\hat{\beta} = \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \log \left( 1 + e^{-Y_i(X_i^T \beta)} \right)
$$

and the minimized function is convex in $\beta$. 

Proof.

\[- \frac{1}{n} \sum_{i=1}^{n} \left( Y_i \log \left( \frac{e^{X_i \top \beta}}{1 + e^{X_i \top \beta}} \right) + \left( 1 - Y_i \right) \log \left( \frac{1}{1 + e^{X_i \top \beta}} \right) \right) \]

\[= - \frac{1}{n} \sum_{i=1}^{n} \left( Y_i \log \left( \frac{1}{1 + e^{-X_i \top \beta}} \right) + \left( 1 - Y_i \right) \log \left( \frac{1}{1 + e^{-X_i \top \beta}} \right) \right) \]

\[= \frac{1}{n} \sum_{i=1}^{n} \log \left( 1 + e^{-Y_i (X_i \top \beta)} \right) \]

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

\[\nabla g(\beta) = Y \frac{e^{-Y X \top \beta} X}{1 + e^{-Y X \top \beta}} \]

\[\nabla^2 g(\beta) = \frac{e^{-Y X \top \beta}}{1 + e^{-Y X \top \beta}} \frac{1}{1 + e^{-Y X \top \beta}} XX \top \]

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

4 Optimization Point of View

4.1 Classical Convexification

Claim 8. The following three losses

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

satisfy

\[\ell(Y, f(X)) = l(Y f(X))\]

with \( l \) a decreasing convex function, differentiable at 0 and such that \( l'(0) < 0 \).

Furthermore \( l(Y, f(X)) \geq \ell^{(1)}(Y, f(X)) \)

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.

Thus for the three losses, by construction, \( l(0) = 1 \) and \( l(z) \geq 0 \) thus \( \ell'(Y, f(X)) = \ell(Y f(\tilde{X})) \geq 1 \) when \( Y f(\tilde{X}) \leq 0 \) and \( \ell'(Y, f(X)) \geq 0 \) otherwise. We obtain thus that \( \ell(Y, f(X)) \geq \ell_{0/1}(Y, f(X)) \).

\[\blacksquare\]

4.2 Classification Risk Analysis with an Optimization Point of View

Claim 9. The minimizer of
\[ E[\ell'(Y, f(X))] = E[\ell(Y f(X))]. \]
Furthermore it exists a convex function \( \Psi \) such that
\[
\Psi \left( E \left[ \ell_{0/1}(Y, \text{sign}(f(X))) \right] - E \left[ \ell_{0/1}(Y, f^*(X)) \right] \right) \\
\leq E \left[ \ell'(Y, f(X)) - \ell'(Y, f^*(X)) \right]
\]

Proof. By definition,
\[ E[\ell(Y f|X)] = \eta(X)\ell(f) + (1 - \eta(X))\ell(-f). \]

Let \( H(f, \eta) = \eta\ell(f) + (1 - \eta)\ell(-f) \), the optimal value for \( \tilde{f} \) satisfies
\[
\delta H(\tilde{f}, \eta) = -\eta\delta\ell(\tilde{f}) + (1 - \eta)\delta\ell(-\tilde{f}) \geq 0.
\]

With a slight abuse of notation, we denote by \( \delta\ell(\tilde{f}) \) and \( \delta\ell(-\tilde{f}) \) the two subgradients such that
\[
\eta\delta\ell(\tilde{f}) - (1 - \eta)\delta\ell(-\tilde{f}) = 0.
\]

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

- If \( \tilde{f} > 0 \), \( \delta\ell(-\tilde{f}) < \delta\ell(\tilde{f}) \) 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 \( E[\ell(y f|X)] \) is \( f^*(X) = \text{sign}(2\eta(X) - 1) \)

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

Furthermore, one verify that if we consider the minimumum over the wrong sign classifiers, \( \inf_{f, \eta < 0} H(f, \eta) = l(0) \).
Indeed,
\[
\inf_{f, f(2\eta-1) < 0} H(f, \eta) \\
= \inf_{f, f(2\eta-1) < 0} (\eta l(f) + (1 - \eta)l(-f)) \\
\geq \inf_{f, f(2\eta-1) < 0} (\eta (l(0) + l'(0)f) + (1 - \eta)(l(0) - l'(0)f)) \\
\geq l(0) + \inf_{f, f(2\eta-1) < 0} l'(0)f(2\eta - 1) = l(0)
\]

Furthermore,
\[
E [\ell'(Y, f(X))] = E_X [H(f, \eta(X))] \\
E [\ell'(Y, f^*(X))] = E_X [H(\eta(X))]
\]

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
\[
E \left[ \ell^{0/1}(Y, \text{sign}(f(X))) \right] - E \left[ \ell^{0/1}(Y, f^*(X)) \right] \\
= E_X \left[ 2\eta(X) - 1 | 1 f^*(X) \neq \text{sign}(f(X)) \right]
\]

Using Jensen inequality, we derive
\[
\Psi \left( E \left[ \ell^{0/1}(Y, \text{sign}(f(X))) \right] - E \left[ \ell^{0/1}(Y, f^*(X)) \right] \right) \\
\leq E_X \left[ \Psi \left( 2\eta(X) - 1 | 1 f^*(X) \neq \text{sign}(f(X)) \right) \right]
\]

Using \(\Psi(0) = 0\) and the symmetry of \(H\),
\[
\Psi \left( E \left[ \ell^{0/1}(Y, \text{sign}(f(X))) \right] - E \left[ \ell^{0/1}(Y, f^*(X)) \right] \right) \\
\leq E_X \left[ (l(0) - H \left( \frac{1 + 2\eta(X) - 1}{2} \right)) 1 f^*(X) \neq \text{sign}(f(X)) \right] \\
\leq E_X \left[ l(0) - H(\eta(X)) \right] 1 f^*(X) \neq \text{sign}(f(X)) \right] \\
\leq E \left[ \ell'(Y, f(X)) \right] - E \left[ \ell'(Y, f^*(X)) \right]
\]

Using the property of the wrong sign classifiers
\[
\Psi \left( E \left[ \ell^{0/1}(Y, \text{sign}(f(X))) \right] - E \left[ \ell^{0/1}(Y, f^*(X)) \right] \right) \\
\leq E \left[ (H(f, \eta(X)) - H(f^*, \eta(X))) 1 f(X) \neq \text{sign}(f(X)) \right] \\
\leq E \left[ (H(f, \eta(X)) - H(f^*, \eta(X))) \right] \\
\leq E \left[ \ell'(Y, f(X)) \right] - E \left[ \ell'(Y, f^*(X)) \right]
\]

\[\square\]
4.3 SVM, distance and norm of $\beta$

Claim 10. The distance between $X^\top \beta + \beta^{(0)} = 1$ and $X^\top \beta + \beta^{(0)} = -1$ is given by

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

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

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

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

4.4 SVM and Hinge Loss

Claim 11. The two problems

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

and

$$\min \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^{n} \max(0, 1 - Y_i (X^\top \beta + \beta^{(0)}))$$

yield the same solution for $\beta$.

Proof. We may write

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

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

Now for any $\beta$,

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

hence the result. \hfill \Box

4.5 Constrained Optimization, Lagrangian and Dual

Claim 12.

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

$$\min_x \max_{\lambda \in \mathbb{R}^p, \mu \in (\mathbb{R}^+)^q} L(x, \lambda, \mu) = \min_x \max_{\lambda \in \mathbb{R}^p, \mu \in (\mathbb{R}^+)^q} L(x, \lambda, \mu) = \begin{cases} h_j(x) = 0, & j = 1, \ldots p \\ g_i(x) \leq 0, & i = 1, \ldots q \end{cases}$$
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
  \[
  \mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{j=1}^{p} \lambda_j h_j(x) + \sum_{i=1}^{q} \mu_i g_i(x)
  \leq f(x) = \mathcal{L}(x, 0, 0)
  \]
  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 \( \lambda_i = \kappa \text{sign}(h_i(x)) \), \( \lambda_{i'} = 0 \) for \( i' \neq i \) and \( \mu = 0 \)
    \[
    \mathcal{L}(x, \lambda, \mu) = f(x) + \kappa \text{sign}(h_i(x)) h_i(x)
    \]
    goes to \( +\infty \) when \( \kappa \) goes to \( \infty \)
  - 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 \( \kappa \) goes to \( \infty \)
  which implies \( \max_{\lambda \in \mathbb{R}^p, \mu \in (\mathbb{R}^+)^q} \mathcal{L}(x, \lambda, \mu) = +\infty \).

Claim 13.
\[
Q(\lambda, \mu) \leq f(x), \text{ for all feasible } x
\]
\[
\max_{\lambda \in \mathbb{R}^p, \mu \in (\mathbb{R}^+)^q} Q(\lambda, \mu) \leq \min_{x \text{ feasible}} f(x)
\]
Proof. The second part is a direct consequence of Claim 13.
By definition,
\[
Q(\lambda, \mu) = \min_x \mathcal{L}(x, \lambda, \mu)
\]
\[
\leq \min_{x \text{ feasible}} \mathcal{L}(x, \lambda, \mu)
\]
\[
\leq \min_{x \text{ feasible}} f(x)
\]
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:
\[
q^* \leq p^*
\]
\[
\max_{\lambda \in \mathbb{R}^p, \mu \in (\mathbb{R}^+)^q} \min_x \mathcal{L}(x, \lambda, \mu) \leq \min_{x \text{ feasible}} \max_{\lambda \in \mathbb{R}^p, \mu \in (\mathbb{R}^+)^q} \mathcal{L}(x, \lambda, \mu)
\]
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:**
  \[
  h_j(x^*) = 0 \quad \text{and} \quad g_i(x^*) \leq 0
  \]

- **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

\[
 f(x^*) = \mathcal{L}(x^*, \lambda, \mu)
 = \min_x \mathcal{L}(x^*, \lambda, \mu)
 \leq \max_{\lambda, \mu} Q(\lambda, \mu) \leq f(x^*)
\]

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

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

\[
 f(x^*) = Q(\lambda^*, \mu^*)
 = \min_x \mathcal{L}(x, \lambda^*, \mu^*)
 \leq \mathcal{L}(x^*, \lambda^*, \mu^*)
 \leq f(x^*)
\]
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:

\[
\begin{align*}
  h_j(x^*) &= 0 \quad \text{and} \quad g_i(x^*) \leq 0 \\
  \mu_i &\geq 0
\end{align*}
\]

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.

\[\square\]

4.8 SVM, KKT and Dual

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

- **Stationarity:**
  \[
  \nabla_\beta \mathcal{L}(\beta, \beta^{(0)}, s, \alpha, \mu) = \beta - \sum_i \alpha_i Y_i X^\top_i = 0 \\
  \nabla_{\beta^{(0)}} \mathcal{L}(\beta, \beta^{(0)}, s, \alpha, \mu) = - \sum_i \alpha_i = 0 \\
  \nabla_s \mathcal{L}(\beta, \beta^{(0)}, s, \alpha, \mu) = C - \alpha_i - \mu_i = 0
  \]

- **Primal and dual admissibility:**
  \[
  (1 - s_i - Y_i (X^\top_i \beta + \beta^{(0)})) \leq 0, \quad s_i \geq 0, \quad \alpha_i \geq 0, \quad \text{and} \quad \mu_i \geq 0
  \]

- **Complementary slackness:**
  \[
  \alpha_i (1 - s_i - Y_i (X^\top_i \beta + \beta^{(0)})) = 0 \quad \text{and} \quad \mu_i s_i = 0
  \]

*Proof.* The Lagrangian of the SVM is given by

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

We can compute the stationarity condition and obtain immediately:

\[
\begin{align*}
  \nabla_\beta \mathcal{L}(\beta, \beta^{(0)}, s, \alpha, \mu) &= \beta - \sum_i \alpha_i Y_i X^\top_i = 0 \\
  \nabla_{\beta^{(0)}} \mathcal{L}(\beta, \beta^{(0)}, s, \alpha, \mu) &= - \sum_i \alpha_i = 0 \\
  \nabla_s \mathcal{L}(\beta, \beta^{(0)}, s, \alpha, \mu) &= C - \alpha_i - \mu_i = 0
\end{align*}
\]

The remaining conditions are straightforward.

\[\square\]
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

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

Proof. As the SVM satisfies the Slater’s constraints. The optimal \( \beta^*, \beta^{(0)*}, s \) of the primal problem and the optimal \( \alpha \) and \( \mu \) of the dual satisfy the KKT optimality condition.

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

If we use \( \nabla_s L(\beta^*, \beta^{(0)*}, s, \alpha, \mu) = 0 \), 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 \( Y_i(X_i^\top \beta^* + \beta^{(0)*}) = 1-s_i \) thus
- either \( s_i = 0 \) and \( Y_i(X_i^\top \beta^* + \beta^{(0)*}) = 1 \),
- or \( s_i \neq 0 \) which implies \( \alpha_i = 0 \) and thus \( \alpha_i = C \) (outliers).

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

Claim 20. The dual of the SVM

\[
Q(\alpha, \mu) = \min_{\beta, \beta^{(0)\cdot}, s} L(\beta, \beta^{(0)}, s, \alpha, \mu)
\]

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 \( \forall i, \alpha_i + \mu_i = C \),
  \[
  Q(\alpha, \mu) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j Y_i X_i^\top X_j
  \]

Proof. The dual of the SVM is defined as

\[
Q(\alpha, \mu) = \min_{\beta, \beta^{(0)\cdot}, s} L(\beta, \beta^{(0)}, s, \alpha, \mu)
\]

\[
= \min_{\beta, \beta^{(0)}, s} \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^n s_i + \sum_i \alpha_i (1 - s_i - Y_i(X_i^\top \beta + \beta^{(0)})) - \sum_i \mu_i s_i
\]

\[
= \min_{\beta, \beta^{(0)}, s} \frac{1}{2} \|\beta\|^2 - \sum_i \alpha_i Y_i X_i^\top \beta - \sum_i \alpha_i Y_i \beta^{(0)} + \sum_i (C - \alpha_i - \mu_i) s_i + \sum_i \alpha_i
\]
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

$$Q(\alpha, \mu) = \min_{\beta, \beta' \in \mathbb{R}} \frac{1}{2} \|\beta\|^2 - \sum_i \alpha_i Y_i X_i^T \beta + \sum_i \alpha_i$$

The optimal $\beta$ can be obtained by setting to 0 the derivative:

$$\beta - \sum_i \alpha_i Y_i X_i^T = 0$$

Plugging this value in the formula yields immediately

$$Q(\alpha, \mu) = -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j Y_i Y_j X_i^T X_j + \sum_i \alpha_i$$

4.9 Mercer Representation Claim

Claim 21. For any loss $\ell$ and any increasing function $\Phi$, the minimizer in $\beta$ of

$$\sum_{i=1}^n \ell(Y_i, X_i^T \beta + \beta^{(0)}) + \Phi(\|\beta\|_2)$$

is a linear combination of the input points $\beta^* = \sum_{i=1}^n \alpha_i X_i$.

Proof. Assume $\beta$ is a minimizer of

$$\sum_{i=1}^n \ell(Y_i, X_i^T \beta + \beta^{(0)}) + \Phi(\|\beta\|_2)$$

and let $\beta_X$ be the orthogonal projection of $\beta$ on the finite dimensional space spanned by the $X_i$. By construction $\beta - \beta_X$ is orthogonal to all the $X_i$ and thus

$$X_i^T \beta + \beta^{(0)} = X_i^T (\beta_X + \beta - \beta_X) + \beta^{(0)} = X_i^T \beta_X + \beta^{(0)}$$

and thus

$$\sum_{i=1}^n \ell(Y_i, X_i^T \beta + \beta^{(0)}) + \Phi(\|\beta\|_2) = \sum_{i=1}^n \ell(Y_i, X_i^T \beta_X + \beta^{(0)}) + \Phi(\|\beta\|_2) \geq \sum_{i=1}^n \ell(Y_i, X_i^T \beta_X + \beta^{(0)}) + \Phi(\|\beta_X\|_2)$$
where the inequality holds because $\|\beta\|^2 = \|\beta X\|^2 + \|\beta - \beta X\|^2$. The minimum is thus reached by a $\beta$ in the space spanned by the $X_i$, i.e.

$$\beta = \sum_{i=1}^{n} \alpha_i X_i.$$ 

\[4.10\] Mercer Kernel Claim

Claim 22. For any PDS kernel $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, it exists a Hilbert space $\mathbb{H} \subset \mathbb{R}^X$ with a scalar product $\langle \cdot , \cdot \rangle_{\mathbb{H}}$ such that

- it exists a mapping $\phi: \mathcal{X} \rightarrow \mathbb{H}$ satisfying
  $$k(\mathbf{X}, \mathbf{X}') = \langle \phi(\mathbf{X}), \phi(\mathbf{X}') \rangle_{\mathbb{H}}$$
- the reproducing property holds, i.e., for any $h \in \mathbb{H}$ and any $\mathbf{X} \in \mathcal{X}$
  $$h(\mathbf{X}) = \langle h, k(\mathbf{X}, \cdot) \rangle_{\mathbb{H}}.$$ 

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

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

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

$$f(\mathbf{X}) = \sum_{i=1}^{n} \alpha_i k(X_i, \mathbf{X})$$
$$= \sum_{i=1}^{n} \alpha_i \langle k(X_i, \cdot), k(X_i, \cdot) \rangle_{\mathcal{H}}$$
$$= \left( \sum_{i=1}^{n} \alpha_i k(X_i, \cdot), k(X_i, \cdot) \right)_{\mathcal{H}}$$
$$= \langle f, k(\mathbf{X}, \cdot) \rangle_{\mathcal{H}}.$$ 

$\mathcal{H}$ 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 $\Psi: \mathcal{X} \rightarrow \mathbb{R}$, $k(\mathbf{X}, \mathbf{X}') = \Psi(\mathbf{X})\Psi(\mathbf{X}')$ is PDS.

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

$$\sum_{i,j} u_i u_j k(\mathbf{X}_i, \mathbf{X}_j) = \sum_{i,j} u_i u_j \phi(\mathbf{X}_i) \phi(\mathbf{X}_j)$$
$$= \left( \sum_{i} u_i \phi(\mathbf{X}_i) \right)^2 \geq 0.$$ 

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 $X_i$ and $u_i$, we have

$$
\sum_{i,j} u_i u_j (k_1 + \lambda k_2)(X_i, X_j) = \sum_{i,j} u_i u_j \left( k_1(X_i, X_j) + \lambda k_2(X_i, X_j) \right)
$$

$$
= \sum_{i,j} u_i u_j k_1(X_i, X_j) + \lambda \sum_{i,j} u_i u_j k_2(X_i, X_j) \geq 0
$$

as a sum of two non-negative terms.

Now for the product

$$
\sum_{i,j} u_i u_j (\lambda k_1 k_2)(X_i, X_j) = \lambda \sum_{i,j} u_i u_j k_1(X_i, X_j) k_2(X_i, X_j)
$$

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

$$
\sum_{i,j} u_i u_j k_1(X_i, X_j) k_2(X_i, X_j)
$$

$$
= \lambda \sum_{i,j} u_i u_j \sum_k M_{i,k} M_{k,j} k_2(X_i, X_j)
$$

$$
= \lambda \sum_{i,j} \sum_k u_i M_{i,k} u_j M_{k,j} k_2(X_i, X_j) \geq 0
$$

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

Claim 24. 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.

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 $r$, $\sum_n a_n k^n$ is a PDS kernel.

Proof. This a direct consequence of the previous claim.

Claim 27. For any PDS kernel $k$, the renormalized kernel $k'(X, X') = \frac{k(X, X')}{\sqrt{k(X, X)k(X', X')}}$ is a PDS kernel.

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

$$
\sum_{i,j} u_i u_j k'(X_i, X_j) = \sum_{i,j} u_i u_j \frac{k(X_i, X_j)}{\sqrt{k(X_i, X)k(X_j, X')}}
$$

$$
= \sum_{i,j} \frac{u_i}{\sqrt{k(X_i, X')}} \frac{u_j}{\sqrt{k(X_j, X')}} k(X_i, X_j) \geq 0
$$
4.12 Mercer Representation Claim

Claim 28. Let $k$ be a PDS kernel and $\mathcal{H}$ its corresponding RKHS, for any increasing function $\Phi$ and any function $L : \mathbb{R}^n \to \mathbb{R}$, the optimization problem

$$\arg\min_{h \in \mathcal{H}} L(h(X_1), \ldots, h(X_n)) + \Phi(\|h\|)$$

admits only solutions of the form

$$\sum_{i=1}^{n} \alpha_i k(X_i, \cdot).$$

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

$$\arg\min_{h \in \mathcal{H}} L(h(X_1), \ldots, h(X_n)) + \Phi(\|h\|).$$

Let $h_X$ be the orthogonal projection of $h$ on the finite dimensional space spanned by the $k(X_i, \cdot)$. By construction, $h - h_X$ is orthogonal to all the $k(X_i, \cdot)$ and thus

$$h(X_i) = \langle h, k(X_i, \cdot) \rangle = \langle h_X + h - h_X, k(X_i, \cdot) \rangle = \langle h_X, k(X_i, \cdot) \rangle = h_X(X_i).$$

This implies that

$$L(h(X_1), \ldots, h(X_n)) + \Phi(\|\beta\|_2) = L(h(X_1), \ldots, h_X(X_n)) + \Phi(\|\beta\|_2)$$

$$\geq L(h(X_1), \ldots, h_X(X_n)) + \Phi(\|\beta_X\|_2)$$

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

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

\[ \square \]

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:

$$\nabla F(w) = \frac{1}{n} \sum_{i=1}^{n} \ell'(Y_i, (X_i, w))X_i$$

with $\ell'(y, f) = \frac{\partial \ell(y, f)}{\partial f}$
Hessian matrix:
\[ \nabla^2 F(w) = \frac{1}{n} \sum_{i=1}^{n} \ell''(Y_i, \langle X_i, w \rangle) X_i X_i^\top \]

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

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 \( \epsilon \)-minimizer of \( G \).

**Required number of evaluation:** \( N_\epsilon = O \left( \frac{(C \sqrt{d}/\epsilon)^d}{\epsilon} \right) \)

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
\[ \nabla G(x + cy) - \nabla G(x) = g(1) - g(0) = g'(t_c) = [\nabla^2 G(x + t_c cy)](cy). \] (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 \( u \),
\[ \| Au \|_2^2 = u^T A^T A u = \langle A^T A u, u \rangle \leq \lambda_{\max}(A)^2 \| u \|^2 \]

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 $w^*$ be the minimum of $f$ on $\mathbb{R}^d$. Then, Gradient Descent with step size $\alpha \leq 1/L$ satisfies

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

Proof. This is a consequence of Lemma 7.

Claim 34. In particular, for $\alpha = 1/L$,

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

iterations are sufficient to get an $\epsilon$-approximation of the minimal value of $G$.

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

Claim 35. If $G$ is convex and $L$-smooth, then for any $w, w' \in \mathbb{R}^d$

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

Proof. Using the fact that

$$G(w') = G(w) + \int_0^1 (\nabla G(w + t(w' - w)))^\top (w' - w)dt$$

$$= G(w) + \nabla G(w)^\top (w' - w)$$

$$+ \int_0^1 (\nabla G(w + t(w' - w)) - \nabla G(w))^\top (w' - w)dt,$$

so that

$$|G(w') - G(w) - (\nabla G(w))^\top (w' - w)|$$

$$\leq \int_0^1 |(\nabla G(w + t(w' - w)) - \nabla G(w))^\top (w' - w)dt|$$

$$\leq \int_0^1 |\nabla G(w + t(w' - w)) - \nabla G(w)||w' - w||dt$$

$$\leq \int_0^1 Lt \|w' - w\|^2 dt = \frac{L}{2} \|w' - w\|^2.$$  

Claim 36. Let $G : \mathbb{R}^d \to \mathbb{R}$ be a $L$-smooth, $\mu$ strongly convex function. Let $w^*$ be the minimum of $G$ on $\mathbb{R}^d$. Then, Gradient Descent with step size $\alpha \leq 1/L$ satisfies

$$G(w^{[k]}) - G(w^*) \leq \frac{1}{2\alpha \mu} \left(1 - \alpha \mu\right)^k \|G(w^{[0]}) - G(w^*)\|^2.$$  

Proof. This is a consequence of Lemma 10.
Claim 37. Let \( G : \mathbb{R}^d \rightarrow \mathbb{R} \) be a convex function, \( C \)-Lipschitz in \( B(w^*, R) \) where \( w^* \) 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 \( \|w^{[0]} - w^*\| \leq R \). Then, Subgradient Descent with step size \( \alpha^{[k]} \) satisfies
\[
\min_k G(w^{[k]}) - G(w^*) \leq C R^2 + \frac{\sum_{k' = 0}^{k} (\alpha^{[k']})^2}{2 \sum_{k' = 0}^{k} \alpha^{[k']}^2}
\]
Proof. This is a consequence of Lemma 14.

5.5 Proximal Descent

Claim 38. \( R(w) = 1_\Omega(w) \): \( \text{prox}_\gamma R(w') = P_\Omega(w') \)

- \( R(w) = \frac{1}{2}\|w\|^2 \): \( \text{prox}_\gamma R(w') = \frac{1}{1+\gamma} w \).
- \( R(w) = \|w\|_1 \): \( \text{prox}_\gamma R(w') = T_\gamma(w') \) with \( T_\gamma(w)_i = \text{sign}(w_i) \max(0, |w_i| - \gamma) \) (soft thresholding).

Proof. If \( R(w) = 1_\Omega(w) \), then
\[
\text{prox}_\gamma R(w') = \arg \min_w \frac{1}{2\gamma} \|w - w'\|^2 + R(w')
\]
\[
= \arg \min_w \frac{1}{2\gamma} \|w - w'\|^2
\]
\[
= P_\Omega(w')
\]

If \( R(w) = \frac{1}{2}\|w\|^2 \) then
\[
\text{prox}_\gamma R(w') = \arg \min_w \frac{1}{2\gamma} \|w - w'\|^2 + R(w')
\]
\[
= \arg \min_w \frac{1}{2\gamma} \|w - w'\|^2 + \frac{1}{2} \|w\|^2
\]
The function minimized is smooth (and strongly convex) and its gradient is given by
\[
\frac{1}{\gamma} (w - w') + w
\]
which is equal to 0 iff \( w = \frac{1}{1+\gamma} w' \), hence the result.

If \( R(w) = \|w\|_1 \) then
\[
\frac{1}{2\gamma} \|w - w'\|^2 + R(w) = \sum_{i} \left( \frac{1}{2\gamma} (w_i - w_i')^2 + |w_i| \right).
\]
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

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

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

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

\[\square\]

**Claim 39.**
- \( F \) L-smooth and \( R \) simple:
  
  $$
  G(w[k]) - G(w^*) \leq \frac{||w[0] - w^*||^2}{2\alpha k}.
  $$
  
  and 
  
  $$
  N_\epsilon = O(L ||w[0] - w^*||^2 / 2\epsilon).
  $$

- \( F \) L-smooth and \( \mu \)-convex and \( R \) simple:
  
  $$
  G(w[k]) - G(w^*) \leq \frac{L}{2\alpha} \left(1 - \alpha \mu\right)^k ||G(w[0]) - G(w^*)||^2.
  $$
  
  and 
  
  $$
  N_\epsilon = O(-\log \epsilon / (\alpha \mu)).
  $$

- \( F \) C-Lipschitz and \( R \) is the characteristic function of a convex set:

  $$
  \min k' \leq kG(w[k']) - G(w^*) \leq C \frac{R^2 + r^2 \log(k + 1)}{4r \sqrt{k + 1}}
  $$
  
  and 
  
  $$
  N_\epsilon = O((C(-\log \epsilon) / \epsilon)^2).
  $$

**Proof.** Those are consequences of Lemma 4, Lemma 9 and Lemma 14. \[\square\]
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 $\{w[k]\}$ 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

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

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

5.7 Gradient Descent Acceleration

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

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

Proof. See Lemma 13

Claim 43. Assume that $G$ is a $L$-smooth, $\mu$ strongly convex function whose minimum is reached at $w^*$. Then, if $\beta[k] = \frac{1 - \sqrt{\mu/L}}{1 + \sqrt{\mu/L}}$,

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

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

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

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

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

Proof. The proof is quite technical and can be found in Nesterov 2018.
5.8 Stochastic Gradient Descent

Claim 45. • With $\alpha_k = 2R/(b\sqrt{k})$
\[
\mathbb{E} \left[ G\left( \frac{1}{k} \sum_{j=1}^{k} w[j] \right) \right] - G(w^*) \leq \frac{3rb}{\sqrt{k}}
\]

• If $G$ is $\mu$-strictly convex then with $\alpha_k = 2/(\mu(k+1))$,
\[
\mathbb{E} \left[ G\left( \frac{2}{k(k+1)} \sum_{j=1}^{k} jw[j] \right) \right] - G(w^*) \leq \frac{2b^2}{\mu(k+1)}.
\]

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
\[
w^{[k+1]} = \text{prox}_{\alpha_k R}(w^{[k]} - \alpha_k \delta_F(w^{[k]}))
\]
where $\delta_F(w^{[k]})$ is a subgradient of $F$ at $w^{[k]}$. If $F$ is differentiable then $\delta_F(w^{[k]}) = \nabla F(w^{[k]})$.

**Lemma 1.** For any differentiable function $F$ and $w$, if we let
\[w^+ = \text{prox}_{\alpha R}(w - \alpha \nabla F(w))\]
then as soon as $\alpha$ satisfy
\[F(w^+) \leq F(w) + \langle \nabla F(w), w^+ - w \rangle + \frac{1}{2\alpha} \|w^+ - w\|^2\]
then for any $z$
\[G(z) - G(w^+) \geq \frac{1}{2\alpha} \|z - w^+\|^2 - \frac{1}{2\alpha} \|z - w\|^2 + F(z) - F(w) - \langle \nabla F(w), z - w \rangle\]

Proof. We introduce the function \[
\phi(x) = F(w) + \langle \nabla F(w), x - w \rangle + R(x) + \frac{1}{2\alpha} \|x - w\|^2 \]
By construction,
\[
\phi(x) = R(x) + \frac{1}{2\alpha} \|x - w - \alpha F(w)\|^2 + F(w) - \alpha \|\nabla F(w)\|^2
\]
and thus $w^+ = \text{prox}_{\alpha R}(w - \alpha \nabla F(w))$ is the minimizer of the $1/\alpha$ strictly convex function $\phi$. This implies that for any $z$,
\[
\phi(z) - \phi(w^+) \geq \frac{1}{2\alpha} \|z - w^+\|^2
\]
\[
\phi(w^+) = F(w) + \langle \nabla F(w), w^+ - w \rangle + R(w^+) + \frac{1}{2\alpha} \| w^+ - w \|^2
\]
and thus using the assumption on \( \alpha \)
\[
\phi(w^+) \geq F(w^+) + R(w^+) = G(w^+)
\]
while
\[
\phi(z) = F(w) + \langle \nabla F(w), z - w \rangle + R(z) + \frac{1}{2\alpha} \| z - w \|^2
\]
adding and subtracting \( F(z) \) yields
\[
\phi(z) = G(z) + \frac{1}{2\alpha} \| z - w \|^2 + F(w) - F(z) + \langle \nabla F(w), z - w \rangle
\]
and thus
\[
G(z) + \frac{1}{2\alpha} \| z - w \|^2 + F(w) - F(z) + \langle \nabla F(w), z - w \rangle - G(w^+) \geq \frac{1}{2\alpha} \| z - w^+ \|^2
\]
which is equivalent to the inequality in the lemma.

**Lemma 2.** For any convex function \( F \) and \( w \), if we let
\[
w^+ = \text{prox}_{\alpha R}(w - \alpha \nabla F(w))
\]
then as soon as \( \alpha \) satisfy
\[
F(w^+) \leq F(w) + \langle \nabla F(w), w^+ - w \rangle + \frac{1}{2\alpha} \| w^+ - w \|^2
\]
then for any \( z \)
\[
G(z) - G(w^+) \geq \frac{1}{2\alpha} \| z - w^+ \|^2 - \frac{1}{2\alpha}(1 - \alpha \mu) \| z - w \|^2
\]
where \( \mu > 0 \) if \( F \) is \( \mu \) strongly convex and \( \mu = 0 \) otherwise. Furthermore \( \alpha \mu \leq 1 \).

**Proof.** This is an immediate consequence of the previous lemma as
\[
F(z) - F(w) - \langle \nabla F(w), z - w \rangle \geq \frac{\mu}{2} \| z - w \|^2
\]
which yields the bounds.

Furthermore, as
\[
F(w^+) \geq F(w) + \langle \nabla F(w), w^+ - w \rangle + \frac{\mu}{2} \| w^+ - w \|^2
\]
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

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

then

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

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

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

Proof. As

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

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

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

This leads to

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

and

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

\[\Box\]

Lemma 4. If $F$ is $L$-smooth and we use the Gradient Descent algorithm with $\alpha^{[k]}$ satisfying

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

then

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

Proof. Lemma 3 yields

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

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

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

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

24
and thus as $G(w^{[k]})$ are decreasing

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

which implies

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

\[\square\]

**Lemma 5.** If $F$ is $L$ smooth then if $\alpha^{[k]} \leq \frac{1}{k}$ then

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

**Proof.** if $F$ is $L$-smooth then

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

and thus

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

\[\square\]

**Lemma 6.** In the backtracking algorithm, at each step

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

and

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

**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

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

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

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

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

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

Proof. We combine Lemma 4 and Lemma 5 to obtain

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

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

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

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 $\mu$ strictly convex, and we use the Gradient Descent algorithm with $\alpha^k$ satisfying

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

then

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

Proof. According to Lemma 3, we have

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

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

The second inequality implies immediately

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

so that

$$\|w^{k+1} - w^*\|^2 \leq \prod_{k'=0}^{k} (1 - \alpha^{k'}) \|w^0 - w^*\|^2.$$
Plugging this bound in the same inequality we have used yields
\[
G(w^{k+1}) - G(w^*) \leq \frac{1}{2\alpha[k]} (1 - \alpha[k] \mu) \|w[k] - w^*\|^2
\]
\[
\leq \frac{1}{2\alpha[k]} \prod_{k'=0}^{k} (1 - \alpha[k'] \mu) \|w[0] - w^*\|^2.
\]

Lemma 10. If \( F \) is \( L \)-smooth and \( \mu \) strictly convex and we use the Gradient Descent algorithm with \( \alpha[k] \) obtained by backtracking then
\[
G(w^{k+1}) - G(w^*) \leq \frac{1}{2\alpha[k]} \prod_{k'=0}^{k} (1 - \alpha[k'] \mu) \|w[0] - w^*\|^2.
\]

Proof. This is a direct consequence of Lemma 6 and Lemma 9.

Lemma 11. If \( F \) is \( L \)-smooth and \( \mu \) strictly convex and we use the Gradient Descent algorithm with \( \alpha[k] = \alpha \leq 1/L \) then
\[
G(w^{k+1}) - G(w^*) \leq \frac{1}{2\alpha} \prod_{k'=0}^{k} (1 - \alpha \mu) \|w[0] - w^*\|^2.
\]

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
\[
F(w^{k+1}) \leq F(w^{[k+1/2]}) + \langle \nabla F(w^{[k+1/2]}), w^{[k+1]} - w^{[k+1/2]} \rangle + \frac{1}{2\alpha[k]} \|w^{[k+1]} - w^{[k+1/2]}\|^2
\]
then provided \( \beta[k] = (\tau[k] - 1)/\tau[k] \) with \( \tau[k] \) satisfying \( \tau[0] = 1 \), \( \tau[k] \geq 1 \) and \( (\tau[k+1])^2 - \tau[k+1] \leq (\tau[k])^2 \) then
\[
G(w^{k+1}) - G(w^*) \leq \frac{1}{2\tau[k] \alpha[k]} \|w[0] - w^*\|^2.
\]

Proof. As
\[
w^{[k+1]} = \text{prox}_{\alpha,R}(w^{[k+1/2]} - \alpha \nabla F(w^{[k+1/2]}))
\]
with
\[
w^{[k+1/2]} = w^{[k]} + \beta[k](w^{[k]} - w^{[k-1]})
\]
we can apply Lemma 2 with \( w = w^{[k+1/2]} \) and \( w^* = w^{[k+1]} \). As soon as \( \alpha[k] \) is such that
\[
F(w^{[k+1]}) \leq F(w^{[k+1/2]}) + \langle \nabla F(w^{[k+1/2]}), w^{[k+1]} - w^{[k+1/2]} \rangle + \frac{1}{2\alpha[k]} \|w^{[k+1]} - w^{[k+1/2]}\|^2
\]
we have

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

Using \( z = \theta[k]w^* + (1 - \theta[k])w[k] \) yields

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

\[ - \frac{1}{2\alpha[k]} \| \theta[k]w^* + (1 - \theta[k])w[k] - w^{[k+1/2]} \|^2 \]

By convexity of \( G \),

\[ G(\theta[k]w^* + (1 - \theta[k])w[k]) - G(w^{[k+1]}) \leq \theta[k]G(w^*) + (1 - \theta[k])G(w[k]) - G(w^{[k+1]}) \]

\[ \leq (1 - \theta[k]) \left( G(w[k]) - G(w^*) \right) - \left( G(w^{[k+1]}) - G(w^*) \right) \]

Now

\[ \| \theta[k]w^* + (1 - \theta[k])w[k] - w^{[k+1/2]} \|^2 = \| \theta[k]w^* + (1 - \theta[k])w[k] - w[k] - \beta[k] (w^k - w^{k-1}) \|^2 \]

\[ = \| \theta[k]w^* + \beta[k]w^{[k-1]} - (\beta[k] + \theta[k])w^k \|^2 \]

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

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

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

Combining the two previous bounds yields

\[ (1 - \theta[k])\alpha[k] \left( G(w[k]) - G(w^*) \right) - \alpha[k] \left( G(w^{[k+1]}) - G(w^*) \right) \]

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

and equivalently

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

\[ \leq \frac{1}{(\theta[k-1])^2} \left( \alpha[k-1] \left( G(w[k]) - G(w^*) \right) + \frac{1}{2} \| \theta[k-1]w^* + (1 - \theta[k-1])w[k-1] - w[k] \|^2 \right) \]

provided

\[ \frac{(\theta[k-1])^2(1 - \theta[k])}{(\theta[k])^2} \alpha[k] \leq \alpha[k-1]. \]
If this holds, one has
\[
\frac{1}{(\theta[k])^2} \left( \alpha[k] \left( G(w^{k+1}) - G(w^*) \right) + \frac{1}{2} \|\theta[k] w^* + (1 - \theta[k]) w[k] - w^{k+1} \|^2 \right) \\
\leq \frac{1}{(\theta[0])^2} \left( \alpha[0] \left( G(w^1) - G(w^*) \right) + \frac{1}{2} \|\theta[0] w^* + (1 - \theta[0]) w[0] - w^1 \|^2 \right)
\]

Using the result obtained with Lemma 2 at \( k = 0 \) and using \( w^{1/2} = w^0 \), we obtain
\[
\frac{1}{(\theta[k])^2} \left( \alpha[k] \left( G(w^{k+1}) - G(w^*) \right) + \frac{1}{2} \|\theta[k] w^* + (1 - \theta[k]) w[k] - w^{k+1} \|^2 \right) \\
\leq \frac{1}{(\theta[0])^2} \left( \frac{1}{2} \|w^0 - w^*\| - \frac{1}{2} \|w^1 - w^*\|^2 + \frac{1}{2} \|\theta[0] w^* + (1 - \theta[0]) w[0] - w^1\|^2 \right)
\]

and thus if we assume that \( \theta[0] = 1 \)
\[
\frac{1}{(\theta[k])^2} \left( \alpha[k] \left( G(w^{k+1}) - G(w^*) \right) + \frac{1}{2} \|\theta[k] w^* + (1 - \theta[k]) w[k] - w^{k+1} \|^2 \right) \\
\leq \frac{1}{2} \|w^0 - w^*\|^2
\]

We deduce thus the following bound
\[
G(w^{k+1}) - G(w^*) \leq \frac{(\theta[k])^2}{2\alpha[k]} \|w^0 - w^*\|^2
\]

Defining everything in term of \( t[k] = 1/\theta[k] \) yields
\[
\beta[k] = \frac{\theta[k](1 - \theta[k-1])}{\theta[k-1]} \\
= \frac{t[k-1] - 1}{t[k]}
\]

we have obtained
\[
G(w^{k+1}) - G(w^*) \leq \frac{1}{2(t[k])^2\alpha[k]} \|w^0 - w^*\|^2
\]

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

\[ \Box \]
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 \( t^{[k]} = k + k_0 \) with \( k_0 \geq 2 \) then

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

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

\[
t^{[k+1]} = \frac{1 + \sqrt{1 + 4t^{[k]}}}{2} \\
\geq 1 + \frac{1 + (k + 2)^2}{2} \\
\geq \frac{1 + k + 2}{2} = \frac{(k + 1) + 2}{2}
\]

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

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

as soon as \( k_0 \geq 2 \). It leads to

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

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

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

30
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 $w^k \in C$,
\[
\|w^{k+1} - w^*\|^2 \leq \|w^k - \alpha^k \delta_F(w^k) - w^*\|^2 \\
\leq \|w^k - w^*\|^2 - 2\alpha^k \langle \delta_F(w^k), w^k - w^* \rangle + (\alpha^k)^2 \|\delta_F(w^k)\|^2 \\
\leq \|w^k - w^*\|^2 + 2\alpha^k \left( F(w^*) - F(w^k) \right) + (\alpha^k)^2 \|\delta_F(w^k)\|^2
\]
this implies
\[
\alpha^k \left( F(w^k) - F(w^*) \right) \leq \frac{1}{2} \left( \|w^k - w^*\|^2 - \|w^{k+1} - w^*\|^2 \right) + \frac{(\alpha^k)^2}{2} \|\delta_F(w^k)\|^2.
\]
Summing these bounds along $k$ yields
\[
\sum_{k'=0}^{k-1} \alpha^{[k']} \left( F(w^{k'}) - F(w^*) \right) \leq \frac{1}{2} \|w^0 - w^*\|^2 + \sum_{k=0}^{k-1} \frac{(\alpha^{[k]})^2}{2} \|\delta_F(w^{k'})\|^2.
\]
We deduce thus that
\[
\min_{0 \leq k' \leq k-1} F(w^{k'}) - F(w^*) \leq \frac{1}{2} \|w^0 - w^*\|^2 + \sum_{k=0}^{k-1} \frac{(\alpha^{[k]})^2}{2} \|\delta_F(w^{k'})\|^2
\]
that is
\[
\min_{0 \leq k' \leq k-1} F(w^{k'}) - F(w^*) \leq \frac{1}{2} \|w^0 - w^*\|^2 + \sum_{k'=0}^{k-1} \frac{(\alpha^{[k]})^2}{2} \|\delta_F(w^{k'})\|^2
\]
Along the same line, we have simultaneously
\[
\min_{1 \leq k' \leq k} \alpha^{[k']} \sum_{k'=1}^{k} \left( F(w^{k'}) - F(w^*) \right) \leq \frac{1}{2} \|w^1 - w^*\|^2 + \sum_{k'=0}^{k-1} \frac{(\alpha^{[k']})^2}{2} \|\delta_F(w^{k'})\|^2
\]
and thus
\[
\frac{1}{k} \sum_{k'=1}^{k} \left( F(w^{k'}) - F(w^*) \right) \leq \frac{\|w^0 - w^*\|^2 + \sum_{k'=0}^{k-1} (\alpha^{[k']})^2 \|\delta_F(w^{k'})\|^2}{2 k \min_{1 \leq k' \leq k} \alpha^{[k']}}
\]
and thus using the convexity of $F$
\[
F \left( \frac{1}{k} \sum_{k'=1}^{k} w^{k'} \right) - F(w^*) \leq \frac{\|w^0 - w^*\|^2 + \sum_{k'=0}^{k-1} (\alpha^{[k']})^2 \|\delta_F(w^{k'})\|^2}{2 k \min_{1 \leq k' \leq k} \alpha^{[k']}}
\]
If we assume that $\|w^k - w^*\|^2 \leq r^2$ and $\|\delta_F(w^{k'})\|^2 \leq B^2$ then this yields
\[
\min_{0 \leq k' \leq k-1} F(w^{k'}) - F(w^*) \leq \frac{r^2 + \sum_{k'=0}^{k-1} (\alpha^{[k']})^2 B^2}{2 \sum_{k'=0}^{k-1} \alpha^{[k']}}
\]
\[
F \left( \frac{1}{k} \sum_{k'=1}^{k} w^{k'} \right) - F(w^*) \leq \frac{r^2 + \sum_{k'=0}^{k-1} (\alpha^{[k']})^2 B^2}{2 k \min_{1 \leq k' \leq k} \alpha^{[k']}}
\]
\[\square\]
Lemma 15. If $F$ is convex such that the sub gradient $\delta_F$ can be bounded, $\|\delta_F\|^2 \leq B^2$, $\|w^k\| = \alpha_0/k$ with $\alpha_0 = r/(\sqrt{2}B)$, we have

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

and

$$\min_{0 \leq k' \leq k} F(w^{[k']}) - F(w^*) \leq \frac{\sqrt{2rB}}{k}$$

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

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

or rather

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

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

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

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

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

If we assume that $\|w^{[1]} - w^*\|^2 \leq r^2$ and $\|\delta_F(w^{[k']})\|^2 \leq B^2$ then this yields

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

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

and if the $\alpha[k]$ are decreasing

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

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

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

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

32
Optimizing in $\alpha_0$ yields $\alpha_0 = r/(\sqrt{2}B)$ and

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

\[\square\]

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

$$F\left(\frac{1}{k(k+1)} \sum_{k'=1}^{k} k'w[k']\right) - F(w^*) \leq \frac{\alpha_0B^2}{2(k+1)}$$

and

$$\min_{k' \leq k} F(w[k']) - F(w^*) \leq \frac{\alpha_0B^2}{2(k+1)}$$

**Proof.** Using the strong convexity of $F$

$$\|w[k+1] - w^*\|^2 \leq \|w[k] - \alpha[k]\nabla F(w[k]) - w^*\|^2$$

$$\leq \|w[k] - w^*\|^2 - 2\alpha[k] \langle \nabla F(w[k]), w[k] - w^* \rangle + (\alpha[k])^2 \|\delta_F(w[k])\|^2$$

$$\leq \|w[k] - w^*\|^2 + 2\alpha[k] \left(F(w^*) - F(w[k])\right) - \alpha[k]\mu \|w[k] - w^*\|^2 + (\alpha[k])^2 \|\delta_F(w[k])\|^2$$

which implies

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

We can now sum those inequalities

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

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

$$+ \sum_{k'=1}^{k} \frac{k'\alpha[k']}{2}\|\nabla F\|^2$$

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

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

33
so that for any $\alpha_0 \geq \frac{\mu}{2}$

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

$$
\leq \frac{\alpha_0}{2} \sum_{k' = 1}^{k} \left\| \nabla F \right\|^2
$$

$$
\leq \frac{k\alpha_0 B^2}{2}
$$

By convexity of $F$

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

$$
\leq \frac{\alpha_0 B^2}{2(k + 1)}
$$

Note that using

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

leads to

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

\( \Box \)

**Lemma 17.** Assume we have access to $\hat{\delta}_F(w)$ which verify $\mathbb{E} \left[ \hat{\delta}_F(w) \right] = \delta_F(w)$ where $\delta_F(w)$ is a subgradient of $F$ at $w$ and $\mathbb{E} \left[ \left\| \hat{\delta}_F(w) \right\|^2 \right] \leq B$.

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

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

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

$$
\mathbb{E} \left[ F \left( \frac{1}{k(k + 1)} \sum_{k' = 1}^{k} k' w^{[k']} \right) \right] - F(w^*) \leq \frac{\alpha_0 B^2}{2(k + 1)}
$$
Proof. In this stochastic setting, we have, if we let $\mu = 0$ if $F$ is not strongly convex:

\[
\begin{align*}
E \left[ \|w^{[k+1]} - w^*\|^2 | w^{[k]} \right] & \leq E \left[ \|w^{[k]} - \alpha^{[k]} \tilde{\delta}_F(w^{[k]}) - w^*\|^2 | w^{[k]} \right] \\
& \leq E \left[ \|w^{[k]} - w^*\|^2 | w^{[k]} \right] - 2\alpha^{[k]} E \left[ \langle \tilde{\delta}_F(w^{[k]}), w^{[k]} - w^* \rangle | w^{[k]} \right] \\
& \quad + (\alpha^{[k]})^2 E \left[ \|\tilde{\delta}_F(w^{[k]})\|^2 | w^{[k]} \right] \\
& \leq \|w^{[k]} - w^*\|^2 - 2\alpha^{[k]} \langle \tilde{\delta}_F(w^{[k]}), w^{[k]} - w^* \rangle + (\alpha^{[k]})^2 B^2 \\
& \leq (1 - \alpha^{[k]} \mu) \|w^{[k]} - w^*\|^2 - 2\alpha^{[k]} \left( F(w^{[k]}) - F(w^*) \right) + (\alpha^{[k]})^2 B^2
\end{align*}
\]

which implies

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

and thus

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

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

References