EE 599: Mathematics of Data

University of Southern California

Lecturer: Mahdi Soltanolkotabi September 2018 parts are adapted from student scribes

---

Lecture note 2

1 RECAP

Boundedness of gradient (L-Lipschitz condition): $\|\nabla f(\boldsymbol{z})\| \leq L$.

Theorem 1 *If $\|\boldsymbol{x}_1 - \boldsymbol{x}^*\| \leq R$, $\|\nabla f(\boldsymbol{x})\| \leq L$ and $\mu = \frac{R}{L\sqrt{t}}$, then we have the following convergence result for gradient descent*

$$f\left(\frac{1}{t}\sum_{s=1}^{t}\boldsymbol{x}_s\right) - f\left(\boldsymbol{x}^*\right) \leq \frac{RL}{\sqrt{t}} \tag{1}$$

Thus, for an error $\leq \epsilon$, we require $\Theta(1/\epsilon^2)$ iterations.

2 The L-smoothness assumption

One can make a stronger statement on convergence by imposing additional assumptions on the function $f(\boldsymbol{x})$.

Definition 1 *(L-smooth functions) A function $f(\boldsymbol{x})$ is L-smooth if its gradient is Lipschitz continuous with parameter $L \geq 0$, i.e.*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\| \tag{2}$$

Note that an L-smooth function satisfies the following properties.

Lemma 2 *If f is convex and L-smooth, then*

$$g(\boldsymbol{x}) = \frac{L}{2}\boldsymbol{x}^T\boldsymbol{x} - f(\boldsymbol{X}) \quad \textit{is convex.} \tag{3}$$

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2. \tag{4}$$

$$\nabla^2 f(\boldsymbol{x}) \leq L\boldsymbol{I}. \tag{5}$$

Let us now show that the different implications of the lemma 2 are true assuming that f is convex and its hessian is continuous

1. **Implication 1:** To prove that *if f is convex and L-smooth, then* $\Leftrightarrow$ *(3)* From (2), it can be written as $\forall \boldsymbol{x}, \boldsymbol{y}$

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), (\boldsymbol{x} - \boldsymbol{y})\rangle \leq L\|\boldsymbol{x} - \boldsymbol{y}\|^2 \rightarrow \langle L(\boldsymbol{x} - \boldsymbol{y}), (\boldsymbol{x} - \boldsymbol{y}))\rangle \tag{6}$$

$$\Longrightarrow \langle L(\boldsymbol{x} - \boldsymbol{y}) - (\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y})\rangle \geq 0 \tag{7}$$

Taking the gradient of (3), we get

$$\nabla g(\boldsymbol{x}) = L\boldsymbol{x} - \nabla f(\boldsymbol{x}) \tag{8}$$

Now, (7) becomes

$$\langle \nabla g(\boldsymbol{x}) - \nabla g(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq 0 \tag{9}$$

Note that (9) is indicates the monotonicity of the gradient mapping of g and hence it implies that g is convex

2. **Implication 2:** To prove that (3) $\Leftrightarrow$ (4) If g is convex,

$$g(\boldsymbol{y}) \geq g(\boldsymbol{x}) + \langle \nabla g(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \tag{10}$$

$$\Longrightarrow \frac{L}{2}\boldsymbol{y}^T\boldsymbol{y} - f(\boldsymbol{y}) \geq \frac{L}{2}\boldsymbol{x}^T\boldsymbol{x} - f(\boldsymbol{x}) + \langle L\boldsymbol{x} - \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \tag{11}$$

$$\Longrightarrow f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \frac{L}{2}\boldsymbol{y}^T\boldsymbol{y} - \frac{L}{2}\boldsymbol{x}^T\boldsymbol{x} + \langle \nabla f(\boldsymbol{x}) - L\boldsymbol{x}, \boldsymbol{y} - \boldsymbol{x} \rangle \tag{12}$$

By expanding and grouping the terms we get,

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \tag{13}$$

which is (4)

3. **Implication 3:** To prove (4) $\Leftrightarrow$ (5)
From Taylor's theorem and when $\exists$ t $\in$ (0,1),

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^T \nabla^2 f[\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})](\boldsymbol{y} - \boldsymbol{x}) \tag{14}$$

where,

$$\frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^T \nabla^2 f[\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})](\boldsymbol{y} - \boldsymbol{x}) \leq \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \tag{15}$$

let $\boldsymbol{y} = \boldsymbol{x} + \tau v$ for some $v \in \mathbb{R}$ and $0 \leq t \leq \tau$ Therefore (15) $\Longrightarrow$

$$\frac{1}{2}\tau^2 v^T \nabla^2 f[\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})]v \leq \tau^2 \frac{L}{2}\|v\|^2 \tag{16}$$

$\forall v, \boldsymbol{x}$ if $\tau \longrightarrow 0$ then $t \longrightarrow 0$
Taking $\lim_{\tau \to 0}$ we get

$$\frac{1}{2}v^T \nabla^2 f(\boldsymbol{x})v \leq \frac{L}{2}\|v\|^2 \tag{17}$$

thus we have

$$0 \leq \frac{1}{2}v^T(LI - \nabla^2 f(\boldsymbol{x}))v$$

and hence $\nabla^2 f(\boldsymbol{x}) \leq LI$

4. **Implication 4:** To prove (5) $\Leftrightarrow$ (2) From the property of L-smootheness and Taylor's theorem, we have

$$\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) = \int_0^1 \nabla^2 f[\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})](\boldsymbol{y} - \boldsymbol{x})dt \tag{18}$$

and

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 = \| \int_0^1 \nabla^2 f[\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})](\boldsymbol{y} - \boldsymbol{x})dt\|_2 \tag{19}$$

Now, using the property, $\| \int g(t)dt\| \leq \int \|g(t)\|dt$,

$$\begin{aligned}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 &\leq \int_0^1 \|\nabla^2 f[\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})]\|\|(\boldsymbol{y} - \boldsymbol{x})\|dt &(20)\\ &\leq L\int_0^1 \|\boldsymbol{x} - \boldsymbol{y}\|dt &(21)\\ &\leq L\|\boldsymbol{y} - \boldsymbol{x}\| &(22)\end{aligned}$$

$\Rightarrow$ f is convex and L-smooth

Further, Equation (4) is also known as the *quadratic upper bound* and has the following consequences:

$$\frac{1}{2L}\|\nabla f(\boldsymbol{x})\|^2 \leq f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \leq \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \tag{23}$$

The right hand side follows immediately from (4). Another consequence is the *co-coercivity* of gradient, given as

$$\forall \boldsymbol{x}, \boldsymbol{y}, \quad (\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^T(\boldsymbol{x} - \boldsymbol{y}) \geq \frac{1}{L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2 \tag{24}$$

Under L-smoothness assumption, we now have the following convergence result

Theorem 3 *If f is L-smooth and $0 \leq \mu \leq \frac{1}{L}$, then the gradient descent iterations $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \mu\nabla f(\boldsymbol{x}_t)$ admit the following convergence result*

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leq \frac{1}{2t\mu}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2. \tag{25}$$

Proof Using $\boldsymbol{y} = \boldsymbol{x} - \mu\nabla f(\boldsymbol{x})$ in the quadratic upper bound (4), we have

$$\begin{aligned}f(\boldsymbol{x} - \mu\nabla f(\boldsymbol{x})) &\leq f(\boldsymbol{x}) - \mu\left(1 - \frac{\mu L}{2}\right)\|\nabla f(\boldsymbol{x})\|^2\\ &\leq f(\boldsymbol{x}) - \frac{\mu}{2}\|\nabla f(\boldsymbol{x})\|^2 \qquad \left(\text{if } 0 \leq \mu \leq \frac{1}{L}\right)\\ &\leq f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x})^T(\boldsymbol{x} - \boldsymbol{x}^*) - \frac{\mu}{2}\|\nabla f(\boldsymbol{x})\|^2 \qquad \text{(tangent lower bound)}\\ &= f(\boldsymbol{x}^*) + \frac{1}{2\mu}\left[\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x} - \boldsymbol{x}^* - \mu\nabla f(\boldsymbol{x})\|^2\right]. &(26)\end{aligned}$$

Now let $\boldsymbol{x} = \boldsymbol{x}_s$ and $\boldsymbol{x}_{s+1} = \boldsymbol{x}_s - \mu \nabla f(\boldsymbol{x}_s)$. Then, from the equation above, we have

$$f(\boldsymbol{x}_{s+1}) - f(\boldsymbol{x}^*) \leq \frac{1}{2\mu} \left[\|\boldsymbol{x}_s - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{s+1} - \boldsymbol{x}^*\|^2 \right]. \tag{27}$$

Performing a telescopic sum on both sides from $s = 1 \dots t$, we get

$$\begin{aligned} \sum_{s=1}^{t} [f(\boldsymbol{x}_s) - f(\boldsymbol{x}^*)] &\leq \frac{1}{2\mu} \left[\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 \right] \\ &\leq \frac{1}{2\mu} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2. \end{aligned} \tag{28}$$

Since, the value of $f(\boldsymbol{x}_s)$ keeps decreasing with every iteration, we have

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leq \frac{1}{t} \sum_{s=1}^{t} [f(\boldsymbol{x}_s) - f(\boldsymbol{x}^*)] \leq \frac{1}{2t\mu} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2, \tag{29}$$

which is the desired result. ■

Note that to get $f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leq \epsilon$, we need $t = O\left(\frac{1}{\epsilon}\right)$.

2.1 Constrained problems (Projected gradient descent)

In the case of a convex constraint set Ω for the optimization problem, a similar convergence result can be obtained using the following lemma

Lemma 4 *If $\boldsymbol{x}^+ = \Pi_\Omega (\boldsymbol{x} - \mu \nabla f(\boldsymbol{x})) := \boldsymbol{x} - \mu g(\boldsymbol{x})$, where Π_Ω is the projector for Ω and $g(\boldsymbol{x}) = \frac{1}{\mu}(\boldsymbol{x} - \boldsymbol{x}^+)$, then*

$$f(\boldsymbol{x}^+) \leq f(\boldsymbol{x}) + g(\boldsymbol{x})^T (\boldsymbol{x}^+ - \boldsymbol{x}) + \frac{L}{2} \|\boldsymbol{x}^+ - \boldsymbol{x}\|^2. \tag{30}$$

Proof We know from the quadratic upper bound that

$$f(\boldsymbol{x}^+) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T (\boldsymbol{x}^+ - \boldsymbol{x}) + \frac{L}{2} \|\boldsymbol{x}^+ - \boldsymbol{x}\|^2. \tag{31}$$

Hence, all we need to show is

$$(\nabla f(\boldsymbol{x}) - g(\boldsymbol{x}))^T (\boldsymbol{x}^+ - \boldsymbol{x}) \leq 0. \tag{32}$$

Now, note that

$$\nabla f(\boldsymbol{x}) - g(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) - \frac{1}{\mu}(\boldsymbol{x} - \boldsymbol{x}^+) = \frac{1}{\mu} \left[\boldsymbol{x}^+ - (\boldsymbol{x} - \mu \nabla f(\boldsymbol{x})) \right]. \tag{33}$$

Therefore, using the fact that $(\Pi_\Omega(\boldsymbol{x}) - \boldsymbol{x})^T (\Pi_\Omega(\boldsymbol{x}) - \boldsymbol{x}^*) \leq 0$ for any convex set Ω, we get the desired result. ■

One can now use the lemma above at the appropriate step in Theorem 3 to get a convergence result for convex constrained problems.

2.2 How to pick μ in practice?

The convergence results so far require that the Lipschitz constant L be known in order to set a value for μ, which is not possible in reality. One way to deal with this problem is *Backtracking line search*, that automatically sets the value of μ in each iteration. To move along the search direction, we start with an estimate of the step size (in this case $\mu = 1$) and then iteratively shrink it (backtracking) until a decrease in the objective function is observed. The least amount of descent expected or the termination condition is given by equation 17,

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \gamma\mu\|\nabla f(\boldsymbol{x}_t)\|^2 \tag{34}$$

and a much greater descent would look like:

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \mu(1 - \frac{\mu L}{2})\|\nabla f(\boldsymbol{x}_t)\|^2 \tag{35}$$

Equation (18) follows from the quadratic upper bound in (4) which we always have at any step t The procedure is given below:

Backtracking line search

1. Pick descent direction $\boldsymbol{v} = -\nabla f(\boldsymbol{x}_t)$.
2. Start with $\mu = 1$.
3. Test Armijo condition

$$f(\boldsymbol{x}_t - \mu\nabla f(\boldsymbol{x}_t)) \leq f(\boldsymbol{x}_t) - \gamma\mu\|\nabla f(\boldsymbol{x}_t)\|^2. \tag{36}$$

   (a) If true, keep μ.
   (b) If false, update μ as $\mu \leftarrow \beta\mu$.

   Rule of thumb: $\gamma = 0.5$ and $\beta = 0.8$.

This procedure works because we change μ in each iteration to keep it as close to $\frac{1}{L}$ as possible. The following convergence result can be obtained for such a procedure

Theorem 5 *For an L-smooth f, and gradient descent iterations given by $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \mu_t\nabla f(\boldsymbol{x}_t)$, we have*

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leq \frac{1}{2t\mu_{\min}}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2, \tag{37}$$

where

$$\mu_{\min} := \min_{1\leq s\leq t} \mu_s \geq \min\left(1, \frac{\beta}{L}\right) \tag{38}$$

Proof Equation (37) follows directly from Theorem 3 due to definition of $\mu_{\min}$ as the minimum of all μ_s. Hence, we just need to show (38) which essentially means $\mu_{\min}$ is close to $\frac{1}{L}$ or alternatively stated - we want to find the value of μ at which the algorithm terminates.

Suppose $\mu = 1$. If the Armijo condition succeeds, we have from equation (17):

$$\|\nabla f(\boldsymbol{x}_t)\|^2 \leq \frac{1}{\gamma}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{t+1})) \tag{39}$$

then we keep μ as this already indicates some kind of bound on the gradient.

But if it fails, we stop the iteration and update μ and start over again. This means at the t^{th} iteration, μ_t/β fails. For such an iteration, we have from the Armijo condition

$$f(\boldsymbol{x}_t) - \frac{\gamma\mu_t}{\beta}\|\nabla f(\boldsymbol{x}_t)\|^2 \leq f\left(\boldsymbol{x}_t - \frac{\mu_t}{\beta}\nabla f(\boldsymbol{x}_t)\right) \tag{40}$$

$$\leq f(\boldsymbol{x}_t) - \frac{\mu_t}{\beta}\left(1 - \frac{\mu_t}{\beta}\frac{L}{2}\right)\|\nabla f(\boldsymbol{x}_t)\|^2 \tag{41}$$

where we arrive at the second expression by using the quadratic upper bound (4) for the right hand side. What we essentially mean by (23) and (24) is that, when μ_t/β fails, the Armijo condition is violated or the the direction of descent is reversed which can be seen as $f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \gamma\mu\|\nabla f(\boldsymbol{x}_t)\|^2$.

On further simplification of equation (24) we get,

$$-\frac{\gamma\mu_t}{2\beta} \leq -\frac{\mu_t}{\beta}(1 - \frac{\mu_t L}{2\beta})$$

$$\implies \mu_t \geq \frac{2(1-\gamma)\beta}{L}$$

Substituting $\gamma = 0.5$ and $\beta = 0.8$ we get,

$$\mu_t \geq \frac{\beta}{L} \approx \frac{0.8}{L}$$

which means in each iteration we have a guarantee that $\mu_{\min} = \min\left(1, \frac{\beta}{L}\right)$. ∎

Note that for an L-smooth function, the Armijo condition cannot be satisfied if $\mu \geq \frac{1}{L}$. Hence, backtracking line search ensures that we not only satisfy it at each iteration, but also pick the largest μ that does it.

2.3 Conditional gradient descent

For constrained optimization problems of the form $\min_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x})$, there exists another method:

Franke-Wolfe algorithm

$$\boldsymbol{y}_t = \operatorname*{argmin}_{\boldsymbol{y}\in\Omega} \nabla f(\boldsymbol{x}_t)^T\boldsymbol{y} \tag{42}$$

$$\boldsymbol{x}_{t+1} = (1-\mu_t)\boldsymbol{x}_t + \mu_t\boldsymbol{y}_t \tag{43}$$

$$\mu_t = \frac{2}{t+1} \tag{44}$$

Note that $\boldsymbol{y}_t$ is the best descent direction. The method above has the advantage that it is projection-free, norm-free and has lower space complexity.

The motivation for this algorithm is that: usual gradient methods involve a projection onto the constrained set Ω in each iteration which can be very expensive. Whereas the Frank-Wolfe method only requires the solution to the linear approximation of the function over the constrained set in each iteration. For instance, consider the local linear approximation of $f(\boldsymbol{x})$,

$$\hat{f}^{lin} = f(\boldsymbol{x_t}) + \nabla f(\boldsymbol{x_t})^T(\boldsymbol{y} - \boldsymbol{x_t}) \tag{45}$$

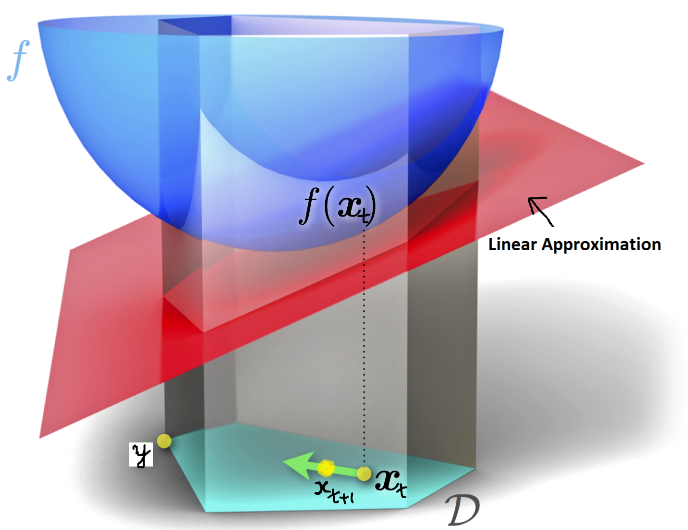
Trying to minimize it over y, would just send the linear approximation to $-\infty$ and cannot project on the constrained set for the same reason as it is not well defined. Therefore Frank-Wolfe minimizes this over the whole constrained set.

$$\boldsymbol{y}_t = \arg\min_{\boldsymbol{y}\in\Omega} \hat{f}^{lin} \tag{46}$$

If $\boldsymbol{y_t}$ is the minimizer, we move a bit in the direction of the minimizer and call it say $\boldsymbol{x_{t+1}}$

$$\boldsymbol{x_{t+1}} = \boldsymbol{x_t} + \mu_t(\boldsymbol{y_t} - \boldsymbol{x_t}) \tag{47}$$

It can be clearly seen from the following figure that the direction of descent is $(\boldsymbol{y_t} - \boldsymbol{x_t})$ and multiply it by some step size μ_t and move along that direction. Equation (48) is just restatement of (44)with little rearrangement. But it can be seen from (48) that, as t increases μ_t decreases and we move less and less aggressively in the direction of the linearization minimizer.



The following convergence result can be obtained for the Franke-Wolf algorithm:

Theorem 6 *Let f be convex and L-smooth with arbitrary norms, i.e., $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_* \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$ ($\|.\|$ and $\|.\|_*$ are dual norms), and let $R = \sup_{\boldsymbol{x},\boldsymbol{y}\in\Omega} \|\boldsymbol{x} - \boldsymbol{y}\|$, $\mu_t = \frac{2}{t+1}$, then, we have*

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leq \frac{2LR^2}{t+1} \tag{48}$$

Proof From equation (4) we can write the quadratic bound over x_t as:

$$f(\boldsymbol{x_{t+1}}) - f(\boldsymbol{x_t}) \leq \nabla f(\boldsymbol{x_t})^T(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \frac{L}{2}\|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\|^2 \tag{49}$$

from (26), we have

$$\boldsymbol{x_{t+1}} - \boldsymbol{x_t} = \mu_t(\boldsymbol{y_t} - \boldsymbol{x_t}) \tag{50}$$

Therefore,

$$\begin{align} f(\boldsymbol{x_{t+1}}) - f(\boldsymbol{x_t}) &\leq \mu_t\langle\nabla f(\boldsymbol{x_t}), (\boldsymbol{y_t} - \boldsymbol{x_t})\rangle + \frac{L\mu_t^2R^2}{2} \tag{51}\\ &\leq \mu_t\langle\nabla f(\boldsymbol{x_t}), (\boldsymbol{x^*} - \boldsymbol{x_t})\rangle + \frac{L\mu_t^2R^2}{2} \tag{52}\end{align}$$

where x^* is the optimal solution
And also we have from the fact that a function is above its first order approximation,

$$\langle\nabla f(\boldsymbol{x_t}), (\boldsymbol{x_*} - \boldsymbol{x_t})\rangle \leq f(\boldsymbol{x^*}) - f(\boldsymbol{x_t}) \tag{53}$$

$$\implies f(\boldsymbol{x_{t+1}}) - f(\boldsymbol{x_t}) \leq \mu_t[f(\boldsymbol{x^*}) - f(\boldsymbol{x_t})] + \frac{L\mu_t^2R^2}{2} \tag{54}$$

$$\implies f(\boldsymbol{x_{t+1}}) - f(\boldsymbol{x^*}) - [f(\boldsymbol{x_t}) - f(\boldsymbol{x^*})] \leq -\mu_t[f(\boldsymbol{x_t}) - f(\boldsymbol{x^*})] + \frac{L\mu_t^2R^2}{2} \tag{55}$$

$$\implies f(\boldsymbol{x_{t+1}}) - f(\boldsymbol{x^*}) \leq (1-\mu_t)[f(\boldsymbol{x_t}) - f(\boldsymbol{x^*})] + \frac{L\mu_t^2R^2}{2} \tag{56}$$

Now, let δ indicate the error and substituting for μ_t, (36) becomes

$$\delta_{t+1} \leq \frac{t-1}{t+1}\delta_t + \frac{L\mu_t^2R^2}{2} \tag{57}$$

Also (37) indicates that the error decreases as the number of iterations increase. Further, when t = 0,

$$\delta_1 = f(\boldsymbol{x_1}) - f(\boldsymbol{x^*}) \leq \frac{L}{2}\|\boldsymbol{x_1} - \boldsymbol{x^*}\|^2 \tag{58}$$

From one of the consequences of the quadratic upper bound, $f(\boldsymbol{x_1})$ can be bounded as in (38)

$$\implies \delta_1 \leq \frac{L}{2}R^2 \tag{59}$$

Continuing like this, by induction we can prove that

$$\delta_t \leq \frac{2}{t+1}LR^2 \tag{60}$$

ie.,

$$f(\boldsymbol{x_t}) - f(\boldsymbol{x^*}) \leq \frac{2L}{t+1}R^2 \tag{61}$$

■

Note that there is no projection, update is solved directly over the constrained set Ω
Let us consider an example of using the Franke-Wolf algorithm.

Example:

$$\begin{aligned} &\min f(\boldsymbol{x}) = \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_{\ell_2}^2 \\ &\text{subject to } \ \|\boldsymbol{x}\|_{\ell_1} \leq \tau \end{aligned} \tag{62}$$

Note that after rescaling (dividing by τ), the above problem is equivalent to

$$\begin{aligned} &\min f(\boldsymbol{x}) = \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_{\ell_2}^2 \\ &\text{subject to } \ \|\boldsymbol{x}\|_{\ell_1} \leq 1 \end{aligned} \tag{63}$$

Solution: Consider $f(\boldsymbol{x}) = \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_{\ell_2}^2$, and let $\nabla f(\boldsymbol{x}) = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})$. Then, the update is of the form

$$\boldsymbol{y}_t = \underset{\|\boldsymbol{y}\|_{\ell_1} \leq 1}{\operatorname{argmin}} \ \nabla f(\boldsymbol{x})^T \boldsymbol{y}.$$

Let $i_{\max} = \arg\max_i \left|[\nabla f(\boldsymbol{x})]_i\right|$ denote the index of the entry of $\nabla f(\boldsymbol{x})$ with the largest absolute value. Then the update is of the form

$$\boldsymbol{y}_t = -\text{sign}\left([\nabla f(\boldsymbol{x})]_{i_{\max}}\right) \boldsymbol{e}_{i_{\max}}, \tag{64}$$

where $\boldsymbol{e}_i$ is the standard basis vector with all entries zeros except a 1 at entry i. For example, if $\nabla f(\boldsymbol{x}) = [2, -3, 1]^T$, then, $\boldsymbol{y}_t = \ [0, 1, 0]^T$. Hence, for the ℓ_1-norm constraint, we clearly see how this method is projection-free. If we start with $\boldsymbol{x}_0 = \boldsymbol{0}$, then after t iterations, we have t non-zero entries. It is also norm-free because the L-smoothness condition for $f(\boldsymbol{x}) = \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_{\ell_2}^2$, i.e.,

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_{\ell_\infty} = \|\boldsymbol{A}^T\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{y})\|_{\ell_\infty} \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_{\ell_1}, \tag{65}$$

where $L = \max_i \|\boldsymbol{A}_i\|_{\ell_2}^2$. In comparison, for the ℓ_2 norm, the same condition

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_{\ell_2} = \|\boldsymbol{A}^T\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{y})\|_{\ell_2} \leq \tilde{L}\|\boldsymbol{x} - \boldsymbol{y}\|_{\ell_2}, \tag{66}$$

holds for $\tilde{L} = \|\boldsymbol{A}^T\boldsymbol{A}\|$, which can be much greater than L.