1. Introduction

The Map of Learning

2. Convex Functions

Lemma 2.20

Suppose there exists $\mathbf{y} \in \mathbf{dom}(f)$ such that $f(\mathbf{y}) < f\left(\mathbf{x}^\star\right)$.

Define $\mathbf{y}' := \lambda \mathbf{x}^\star + (1-\lambda)\mathbf{y}$ for $\lambda \in (0,1)$

From convexity, we get that that $f(\mathbf{y}') < f\left(\mathbf{x}^\star\right)$. Choosing λ so close to 1 that $\left\|\mathbf{y}' - \mathbf{x}^\star\right\| < \varepsilon$ yields a contradiction to $\mathbf{x}^\star$ being a local minimum.

Lemma 2.21

Suppose that $\nabla f(\mathbf{x}) = \mathbf{0}$. According to the first-order characterization of convexity, we have

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) = f(\mathbf{x})$$

for all $\mathbf{y} \in \mathrm{dom}(f)$, so $\mathbf{x}$ is a global minimum.

Thm 2.29 (Weierstrass Theorem)

We know that f-as a continuous function-attains a minimum over the closed and bounded (= compact) set $f^{\leq \alpha}$ at some $\mathbf{x}^\star$. This $\mathbf{x}^\star$ is also a global minimum as it has value $f\left(\mathbf{x}^\star\right) \leq \alpha$, while any $\mathbf{x} \notin f^{\leq \alpha}$ has value $f(\mathbf{x}) > \alpha \geq f\left(\mathbf{x}^\star\right)$. Generalizes to suitable domains $\mathrm{dom}(f) \neq \mathbb{R}^d$.

Lemma 2.45

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \underbrace{\sum_{i=1}^m \lambda_i f_i(\mathbf{x})}_{\leq 0} + \underbrace{\sum_{i=1}^p \nu_i h_i(\mathbf{x})}_{=0} \leq f_0(\mathbf{x})$$

Lemma 2.49 & 2.50

Master Equation

$$\begin{aligned} f_0(\tilde{\mathbf{x}}) &= g(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}) \\ &= \inf_{\mathbf{x} \in \mathcal{D}} \left(f_0(\mathbf{x}) + \sum_{i=1}^m \tilde{\lambda}_i f_i(\mathbf{x}) + \sum_{i=1}^p \tilde{\nu}_i h_i(\mathbf{x}) \right) \\ &\leq f_0(\tilde{\mathbf{x}}) + \underbrace{\sum_{i=1}^m \tilde{\lambda}_i f_i(\tilde{\mathbf{x}})}_{\leq 0} + \underbrace{\sum_{i=1}^p \tilde{\nu}_i h_i(\tilde{\mathbf{x}})}_{0} \\ &\leq f_0(\tilde{\mathbf{x}}). \end{aligned}$$

All inequalities are equalities!

Lemma 2.49 follows from $\tilde{\lambda}_i f_i(\tilde{\mathbf{x}}) = 0$ in the Master Equation.

By equality in the third line of the Master Equation, $\tilde{\mathbf{x}}$ minimizes the differentiable function

$$f_0(\mathbf{x}) + \sum_{i=1}^m \tilde{\lambda}_i f_i(\mathbf{x}) + \sum_{i=1}^p \tilde{\nu}_i h_i(\mathbf{x})$$

Hence its gradient vanishes by Lemma 2.22.

3. Gradient Descent

Vanilla Analysis

$\mathbf{g}_t^\top \left(\mathbf{x}_t - \mathbf{x}^\star\right) = \frac{1}{\gamma} (\mathbf{x}_t - \mathbf{x}_{t+1})^\top \left(\mathbf{x}_t - \mathbf{x}^\star\right)$

Apply $2\mathbf{v}^\top \mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2$ (cosine theorem) to rewrite

$$\begin{aligned} \mathbf{g}_t^\top \left(\mathbf{x}_t - \mathbf{x}^\star\right) &= \frac{1}{2\gamma} \left(\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \left\|\mathbf{x}_t - \mathbf{x}^\star\right\|^2 - \left\|\mathbf{x}_{t+1} - \mathbf{x}^\star\right\|^2 \right) \\ &= \frac{\gamma}{2} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \left(\left\|\mathbf{x}_t - \mathbf{x}^\star\right\|^2 - \left\|\mathbf{x}_{t+1} - \mathbf{x}^\star\right\|^2 \right) \end{aligned}$$

Sum this up over the first T iterations:

$$\sum_{t=0}^{T-1} \mathbf{g}_t^\top \left(\mathbf{x}_t - \mathbf{x}^\star\right) = \frac{\gamma}{2} \sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \left(\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|^2 - \left\|\mathbf{x}_T - \mathbf{x}^\star\right\|^2 \right)$$

Remember: $f(\mathbf{x}_t) - f\left(\mathbf{x}^\star\right) \leq \mathbf{g}_t^\top \left(\mathbf{x}_t - \mathbf{x}^\star\right)$.Plug this lower bound into Vanilla Analysis:

$$\begin{aligned} \sum_{t=0}^{T-1} \left(f(\mathbf{x}_t) - f\left(\mathbf{x}^\star\right) \right) &\leq \sum_{t=0}^{T-1} \mathbf{g}_t^\top \left(\mathbf{x}_t - \mathbf{x}^\star\right) \\ &= \frac{\gamma}{2} \sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \left(\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|^2 - \left\|\mathbf{x}_T - \mathbf{x}^\star\right\|^2 \right) \\ &\leq \frac{\gamma}{2} \sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|^2 \end{aligned}$$

Thm 3.1 (Lipschitz Cvx Func $\mathcal{O}\left(1/\varepsilon^2\right)$ Steps)

Plug $\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\| \leq R$ and $\|\mathbf{g}_t\| \leq B$ into Vanilla Analysis || :

$$\sum_{t=0}^{T-1} \left(f(\mathbf{x}_t) - f\left(\mathbf{x}^\star\right) \right) \leq \frac{\gamma}{2} \sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|^2 \leq \frac{\gamma}{2} B^2 T + \frac{1}{2\gamma} R^2$$

Choose γ such that$q(\gamma) = \frac{\gamma}{2} B^2 T + \frac{R^2}{2\gamma}$is minimized.

Solving $q'(\gamma) = 0$ yields the minimum $\gamma = \frac{R}{B\sqrt{T}}$, and $q(R/(B\sqrt{T})) = RB\sqrt{T}$. Dividing by T, the result follows.

Lemma 3.3

g being convex is by the first-order characterization equivalent to

$$g(\mathbf{y}) \geq g(\mathbf{x}) + \nabla g(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}), \quad \mathbf{x}, \mathbf{y} \in \mathrm{dom}(g)$$

Using the definition of g, this is equivalent to

$$\frac{L}{2} y^\top \mathbf{y} - f(y) \geq \frac{L}{2} \mathbf{x}^\top \mathbf{x} - \mathrm{f}(\mathbf{x}) + (\mathrm{L}\mathrm{x} - \nabla f(x))^\top (\mathbf{y} - \mathrm{x})$$

Reordering terms, this is in turn equivalent to

$$f(y) \leq f(x) + \nabla f(x)(y - x) + \frac{L}{2} y^\top y - \frac{L}{2} x^\top x - L x^\top (y - x)$$

Since $\frac{L}{2} y^\top y - \frac{L}{2} x^\top x - L x^\top (y - x) = \frac{L}{2} y^\top y + \frac{L}{2} x^\top x - L x^\top y = \frac{L}{2} \|x - y\|^2$ we get the definition of smoothness, so the statement follows.

Lemma 3.4

As the function $\mathbf{x} \mapsto \mathbf{b}^\top \mathbf{x} + \mathrm{c}$ is affine and hence smooth with parameter 0 , it suffices by Lemma 3.6 to restrict ourselves to the case $f(x) := x^\top \mathbf{Q} x$.

Because $\mathbf{Q}$ is symmetric, $x^\top \mathbf{Q} y = y^\top \mathbf{Q} x$ for any x and y. Thus, a simple calculation shows that

$$\begin{aligned} f(y) = y^\top \mathbf{Q} y &= x^\top \mathbf{Q} x + 2x^\top \mathbf{Q}(y - x) + (x - y)^\top \mathbf{Q}(x - y) \\ &= f(x) + 2x^\top \mathbf{Q}(y - x) + (x - y)^\top \mathbf{Q}(x - y) \end{aligned}$$

Cauchy-Schwarz for $(\mathbf{x} - \mathbf{y})^\top \mathbf{Q}(\mathrm{x} - \mathrm{y}) \leq \|\mathrm{x} - \mathrm{y}\| \|\mathrm{Q}(\mathrm{x} - \mathrm{y})\|$, and using and the definition of spectral norm for $\|\mathbf{Q}(x - y)\| \leq \|\mathbf{Q}\| \|x - y\|$ we get

$$f(y) \leq f(x) + 2x^\top \mathbf{Q}(y - x) + \|\mathbf{Q}\| \|x - y\|^2,$$

Because $\|x - y\|^2$ vanishes as $(x - y)$ goes to 0 , differentiability of f (Definition 2.5) implies that $\nabla f(x)^\top = 2x^\top Q$, so we further get

$$f(y) \leq f(x) + \nabla f(x)(y - x) + \frac{2\|Q\|}{2} \|x - y\|^2,$$

That is, f is smooth with parameter $2\|Q\|$.

Lemma 3.6

For (1), we sum up the weighted smoothness conditions for all the f_i to obtain

$$\sum_{i=1}^m \lambda_i f_i(x) \leq \sum_{i=1}^m \lambda_i f_i(y) + \sum_{i=1}^m \lambda_i \nabla f_i(x)^\top (y - x) + \sum_{i=1}^m \lambda_i \frac{L_i}{2} \|x - y\|^2.$$

As the gradient is a linear operator, this equivalently reads as

$$f(x) \leq f(y) + \nabla f(x)^\top (y - x) + \frac{\sum_{i=1}^m \lambda_i L_i}{2} \|x - y\|^2$$

and the statement follows.

For (2), we apply smoothness of f at $x' = Ax + b$ and $y' = Ay + b$ to obtain

$$f(Ax + b) \leq f(Ay + b) + \nabla f(Ax + b)^\top (A(y - x)) + \frac{L}{2} \|A(x - y)\|^2$$

As $\nabla(\mathrm{f} \circ \mathrm{g})(\mathbf{x})^\top = \nabla f(\mathbf{A}\mathbf{x} + \mathbf{b})^\top A$ (chain rule (Lemma 2.6), using that $\mathrm{Dg}(\mathbf{x}) = A$, an easy consequence of Definition 2.5). This equivalently reads as

$$(f \circ g)(x) \leq (f \circ g)(y) + \nabla(f \circ g)(x)^\top (y - x) + \frac{L}{2} \|A(x - y)\|^2$$

The statement now follows from $\|A(x - y)\| \leq \|A\| \|x - y\|$.

Lemma 3.7 Sufficient Decrease

Use smoothness and definition of gradient descent $(\mathbf{x}_{t+1} - \mathbf{x}_t = -\nabla f(\mathbf{x}_t)/L)$:

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\ &= f(\mathbf{x}_t) - \frac{1}{L} \|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2L} \|\nabla f(\mathbf{x}_t)\|^2 \\ &= f(\mathbf{x}_t) - \frac{1}{2L} \|\nabla f(\mathbf{x}_t)\|^2 \end{aligned}$$

Thm 3.8 Smooth Convex Func $\mathcal{O}(1/\varepsilon)$ Steps
Vanilla Analysis II:

$$\sum_{t=0}^{T-1}\left(f(\mathbf{x}_t)-f\left(\mathbf{x}^\star\right)\right) \le \frac{\gamma}{2}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2+\frac{1}{2\gamma}\left\|\mathbf{x}_0-\mathbf{x}^\star\right\|^2.$$

This time, we can bound the squared gradients by sufficient decrease:

$$\frac{1}{2L}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 \le \sum_{t=0}^{T-1}\left(f(\mathbf{x}_t)-f(\mathbf{x}_{t+1})\right)=f(\mathbf{x}_0)-f(\mathbf{x}_T).$$

Putting it together with $\gamma = 1/L$:

$$\begin{aligned}\sum_{t=0}^{T-1}\left(f(\mathbf{x}_t)-f\left(\mathbf{x}^\star\right)\right) &\le \frac{1}{2L}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2+\frac{L}{2}\left\|\mathbf{x}_0-\mathbf{x}^\star\right\|^2\\ &\le f(\mathbf{x}_0)-f(\mathbf{x}_T)+\frac{L}{2}\left\|\mathbf{x}_0-\mathbf{x}^\star\right\|^2\end{aligned}$$

Rewriting: $\sum_{t=1}^{T}\left(f(\mathbf{x}_t)-f\left(\mathbf{x}^\star\right)\right) \le \frac{L}{2}\left\|\mathbf{x}_0-\mathbf{x}^\star\right\|^2$.
As last iterate is the best (sufficient decrease!):

$$f(\mathbf{x}_T)-f\left(\mathbf{x}^\star\right) \le \frac{1}{T}\left(\sum_{t=1}^{T}\left(f(\mathbf{x}_t)-f\left(\mathbf{x}^\star\right)\right)\right) \le \frac{L}{2T}\left\|\mathbf{x}_0-\mathbf{x}^\star\right\|^2.$$

Lemma 3.11 Strongly Cvx Func
g being convex is by the first-order characterization equivalent to

$$g(y) \ge g(x)+\nabla g(x)^\top(\mathbf{y}-\mathbf{x}), \quad \mathbf{x}, \mathrm{y} \in \mathrm{dom}(g).$$

Using the definition of g, this is equivalent to

$$f(y)-\frac{\mu}{2}y^\top y \ge f(x)-\frac{\mu}{2}x^\top x+(\nabla f(x)-\mu x)^\top(y-x)$$

Reordering terms, this is in turn equivalent to

$$f(y) \ge f(x)+\nabla f(x)(y-x)+\frac{\mu}{2}y^\top y-\frac{\mu}{2}x^\top x-\mu x^\top(y-x).$$

Since

$$\frac{\mu}{2}y^\top y-\frac{\mu}{2}x^\top x-\mu x^\top(y-x)=\frac{\mu}{2}y^\top y+\frac{\mu}{2}x^\top x-\mu x^\top y=\frac{\mu}{2}\|x-y\|^2$$

we get the definition of strong convexity, so the statement follows.

4a. Projected Gradient Descent

Fact 4a.1 Properties of Projection
(i) $\Pi_X(\mathbf{y})$ is minimizer of (differentiable) convex function $d_{\mathbf{y}}(\mathbf{x}) = \|\mathbf{x}-\mathbf{y}\|^2$ over X. By first-order characterization of optimality (Lemma 2.27),

$$\begin{aligned}0 &\le \nabla d_{\mathbf{y}}(\Pi_X(\mathbf{y}))^\top(\mathbf{x}-\Pi_X(\mathbf{y}))\\ &= 2(\Pi_X(\mathbf{y})-\mathbf{y})^\top(\mathbf{x}-\Pi_X(\mathbf{y}))\end{aligned}$$

(ii)

$$\mathbf{v} := (\mathbf{x}-\Pi_X(\mathbf{y})), \quad \mathbf{w} := (\mathbf{y}-\Pi_X(\mathbf{y}))$$

By (i),

$$\begin{aligned}0 \ge 2\mathbf{v}^\top\mathbf{w} &= \|\mathbf{v}\|^2+\|\mathbf{w}\|^2-\|\mathbf{v}-\mathbf{w}\|^2\\ &= \|\mathbf{x}-\Pi_X(\mathbf{y})\|^2+\|\mathbf{y}-\Pi_X(\mathbf{y})\|^2-\|\mathbf{x}-\mathbf{y}\|^2.\end{aligned}$$

Lemma 4a.3 Projected Sufficient Decrease
Use smoothness, $\mathbf{y}_{t+1}-\mathbf{x}_t = -\nabla f(\mathbf{x}_t)/L, 2\mathbf{v}\mathbf{w} = \|\mathbf{v}\|^2+\|\mathbf{w}\|^2-\|\mathbf{v}-\mathbf{w}\|^2$:

$$\begin{aligned}f(\mathbf{x}_{t+1}) &\le f(\mathbf{x}_t)+\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_{t+1}-\mathbf{x}_t)+\frac{L}{2}\|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2\\ &= f(\mathbf{x}_t)-L(\mathbf{y}_{t+1}-\mathbf{x}_t)^\top(\mathbf{x}_{t+1}-\mathbf{x}_t)+\frac{L}{2}\|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2\\ &= f(\mathbf{x}_t)-\frac{L}{2}\left(\|\mathbf{y}_{t+1}-\mathbf{x}_t\|^2+\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2-\|\mathbf{y}_{t+1}-\mathbf{x}_{t+1}\|^2\right)\\ &\quad+\frac{L}{2}\|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2\\ &= f(\mathbf{x}_t)-\frac{L}{2}\|\mathbf{y}_{t+1}-\mathbf{x}_t\|^2+\frac{L}{2}\|\mathbf{y}_{t+1}-\mathbf{x}_{t+1}\|^2\\ &= f(\mathbf{x}_t)-\frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2+\frac{L}{2}\|\mathbf{y}_{t+1}-\mathbf{x}_{t+1}\|^2\end{aligned}$$

Thm 4a.4 Smooth Convex Func over X : $\mathcal{O}(1/\varepsilon)$ Steps
Constrained Vanilla Analysis
Replace $\mathbf{x}_{t+1}$ in the vanilla analysis with $\mathbf{y}_{t+1}$ (the unprojected):

$$\mathbf{g}_t^\top\left(\mathbf{x}_t-\mathbf{x}^\star\right)=\frac{1}{2\gamma}\left(\gamma^2\|\mathbf{g}_t\|^2+\left\|\mathbf{x}_t-\mathbf{x}^\star\right\|^2-\left\|\mathbf{y}_{t+1}-\mathbf{x}^\star\right\|^2\right).$$

Use Fact 4.1 (ii): $\|\mathbf{x}-\Pi_X(\mathbf{y})\|^2+\|\mathbf{y}-\Pi_X(\mathbf{y})\|^2 \le \|\mathbf{x}-\mathbf{y}\|^2$.
With $\mathbf{x}=\mathbf{x}^\star, \mathbf{y}=\mathbf{y}_{t+1}$, we have $\Pi_X(\mathbf{y})=\mathbf{x}_{t+1}$, and hence

$$\left\|\mathbf{x}^\star-\mathbf{x}_{t+1}\right\|^2+\underline{\|\mathbf{y}_{t+1}-\mathbf{x}_{t+1}\|^2} \le \left\|\mathbf{x}^\star-\mathbf{y}_{t+1}\right\|^2$$

We get back to the standard vanilla analysis, but with a saving!

$$\mathbf{g}_t^\top\left(\mathbf{x}_t-\mathbf{x}^\star\right) \le \frac{1}{2\gamma}\left(\gamma^2\|\mathbf{g}_t\|^2+\left\|\mathbf{x}_t-\mathbf{x}^\star\right\|^2-\left\|\mathbf{x}_{t+1}-\mathbf{x}^\star\right\|^2-\underline{\|\mathbf{y}_{t+1}-\mathbf{x}_{t+1}\|^2}\right)$$

Proof
Use $f(\mathbf{x}_t)-f\left(\mathbf{x}^\star\right) \le \mathbf{g}_t^\top\left(\mathbf{x}_t-\mathbf{x}^\star\right)$ (convexity), vanilla analysis with saving, $\gamma = 1/L$:

$$\begin{aligned}\sum_{t=0}^{T-1}\left(f(\mathbf{x}_t)-f\left(\mathbf{x}^\star\right)\right) &\le \sum_{t=0}^{T-1}\mathbf{g}_t^\top\left(\mathbf{x}_t-\mathbf{x}^\star\right)\\ &\le \frac{1}{2L}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2+\frac{L}{2}\left\|\mathbf{x}_0-\mathbf{x}^\star\right\|^2-\frac{L}{2}\sum_{t=0}^{T-1}\|\mathbf{y}_{t+1}-\mathbf{x}_{t+1}\|^2\end{aligned}$$

Use projected sufficient decrease to bound $\frac{1}{2L}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2$ by

$$\begin{aligned}&\sum_{t=0}^{T-1}\left(f(\mathbf{x}_t)-f(\mathbf{x}_{t+1})+\frac{L}{2}\|\mathbf{y}_{t+1}-\mathbf{x}_{t+1}\|^2\right)\\ &= f(\mathbf{x}_0)-f(\mathbf{x}_T)+\frac{L}{2}\sum_{t=0}^{T-1}\|\mathbf{y}_{t+1}-\mathbf{x}_{t+1}\|^2\end{aligned}$$

Putting it together: extra terms cancel, and as in unconstrained case, we get

$$\sum_{t=1}^{T}\left(f(\mathbf{x}_t)-f\left(\mathbf{x}^\star\right)\right) \le \frac{L}{2}\|\mathbf{x}_0-\mathbf{x}^\star\|^2.$$

Exercise 32: again, we make progress in every step (not immediate from projected sufficient decrease). Hence,

$$f(\mathbf{x}_T)-f\left(\mathbf{x}^\star\right) \le \frac{1}{T}\sum_{t=1}^{T}\left(f(\mathbf{x}_t)-f\left(\mathbf{x}^\star\right)\right) \le \frac{L}{2T}\left\|\mathbf{x}_0-\mathbf{x}^\star\right\|^2$$

4b. Coordinate Descent

Lemma 4b.2 Strong convexity $\Rightarrow$ PL inequality

$$\begin{aligned}f\left(\mathbf{x}^\star\right) &\ge f(\mathbf{x})+\nabla f(\mathbf{x})^\top\left(\mathbf{x}^\star-\mathbf{x}\right)+\frac{\mu}{2}\left\|\mathbf{x}^\star-\mathbf{x}\right\|^2 \quad \text{(strong convexity)}\\ &\ge f(\mathbf{x})+\min_{\mathbf{y}}\left(\nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x})+\frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|^2\right)\\ &= f(\mathbf{x})-\frac{1}{2\mu}\|\nabla f(\mathbf{x})\|^2\end{aligned}$$

Thm 4b.3 GD on Smooth Func with PL Ineq
For all t:

$$\begin{aligned}f(\mathbf{x}_{t+1}) &\le f(\mathbf{x}_t)-\tfrac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 \quad \text{(sufficient decrease, Lemma 3.7)}\\ &\le f(\mathbf{x}_t)-\tfrac{\mu}{L}\left(f(\mathbf{x}_t)-f\left(\mathbf{x}^\star\right)\right)\end{aligned}$$

Subtract $f\left(\mathbf{x}^\star\right)$ on both sides:

$$f(\mathbf{x}_{t+1})-f\left(\mathbf{x}^\star\right) \le \left(1-\frac{\mu}{L}\right)\left(f(\mathbf{x}_t)-f\left(\mathbf{x}^\star\right)\right)$$

Lemma 4b.5 Coordinate-wise Sufficient Decrease
Apply coordinate-wise smoothness with $\lambda = -\nabla_i f(\mathbf{x}_t)/L_i$ and $\mathbf{x}_{t+1} = \mathbf{x}_t+\lambda\mathbf{e}_i$

$$\begin{aligned}f(\mathbf{x}_{t+1}) &\le f(\mathbf{x}_t)+\lambda\nabla_i f(\mathbf{x}_t)+\frac{L_i}{2}\lambda^2\\ &= f(\mathbf{x}_t)-\frac{1}{L_i}|\nabla_i f(\mathbf{x}_t)|^2+\frac{1}{2L_i}|\nabla_i f(\mathbf{x}_t)|^2\\ &= f(\mathbf{x}_t)-\frac{1}{2L_i}|\nabla_i f(\mathbf{x}_t)|^2\end{aligned}$$

Thm 4b.6

Coordinate-wise sufficient decrease:

$$f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - \frac{1}{2L}|\nabla_i f(\mathbf{x}_t)|^2.$$

Taking expectations with respect to the choice of the active coordinate i:

$$\begin{aligned}\mathbb{E}[f(\mathbf{x}_{t+1}) \mid \mathbf{x}_t] &\le f(\mathbf{x}_t) - \frac{1}{2L}\sum_{i=1}^{d}\frac{1}{d}|\nabla_i f(\mathbf{x}_t)|^2 \\ &= f(\mathbf{x}_t) - \frac{1}{2dL}\|\nabla f(\mathbf{x}_t)\|^2 \\ &\le f(\mathbf{x}_t) - \frac{\mu}{dL}\left(f(\mathbf{x}_t) - f\left(\mathbf{x}^\star\right)\right) \quad \text{(PL inequality)}\end{aligned}$$

Subtracting $f\left(\mathbf{x}^\star\right)$ from both sides:

$$\mathbb{E}\left[f(\mathbf{x}_{t+1}) - f\left(\mathbf{x}^\star\right) \mid \mathbf{x}_t\right] \le \left(1 - \frac{\mu}{dL}\right)\left(f(\mathbf{x}_t) - f\left(\mathbf{x}^\star\right)\right)$$

Taking expectations with respect to $\mathbf{x}_t$:

$$\mathbb{E}\left[f(\mathbf{x}_{t+1}) - f\left(\mathbf{x}^\star\right)\right] \le \left(1 - \frac{\mu}{dl}\right)\mathbb{E}\left[f(\mathbf{x}_t) - f\left(\mathbf{x}^\star\right)\right]$$

Thm 4b.7 Importance Sampling

Sufficient decrease according to Lemma 5.5 yields

$$f(x_{t+1}) \le f(x_t) - \frac{1}{2L_i}|\nabla_i f(x_t)|^2$$

if coordinate i is chosen. By taking the expectation of both sides with respect to the choice of i, we have

$$\begin{aligned}\mathbb{E}[f(x_{t+1}) \mid x_t] &\le f(x_t) - \sum_{i=1}^{d}\frac{L_i}{\sum_{j=1}^{d} L_j}\frac{1}{2L_i}|\nabla_i f(x_t)|^2 \\ &= f(x_t) - \frac{1}{2\sum_{j=1}^{d} L_j}\sum_{i=1}^{d}|\nabla_i f(x_t)|^2 \\ &= f(x_t) - \frac{1}{2d\bar{L}}\|\nabla f(x_t)\|^2 \\ &\le f(x_t) - \frac{\mu}{d\bar{L}}\left(f(x_t) - f\left(x^\star\right)\right) \quad \text{(PL inequality (5.1)).}\end{aligned}$$

Subtracting $f\left(x^\star\right)$ from both sides, we therefore obtain

$$\mathbb{E}\left[f(x_{t+1}) - f\left(x^\star\right) \mid x_t\right] \le \left(1 - \frac{\mu}{d\bar{L}}\right)\left(f(x_t) - f\left(x^\star\right)\right)$$

Taking expectations (over x_t), we obtain

$$\mathbb{E}\left[f(x_{t+1}) - f\left(x^\star\right)\right] \le \left(1 - \frac{\mu}{d\bar{L}}\right)\mathbb{E}\left[f(x_t) - f\left(x^\star\right)\right]$$

Lemma 4b.9

The main step is to show that

$$\min_{\mathbf{y}}\left(\nabla f(\mathbf{x})^\top\underbrace{(\mathbf{y} - \mathbf{x})}_{=:\mathbf{z}} + \frac{\mu_1}{2}\|\mathbf{y} - \mathbf{x}\|_1^2\right) = -\frac{1}{2\mu_1}\|\nabla f(\mathbf{x})\|_\infty^2,$$

the rest of the proof is the same as Lemma 5.2. Let

$$g(z) = \nabla f(x)^\top z + \frac{\mu}{2}\|z\|_1^2.$$

Fix $K \in \mathbb{R}$. Among all z such that $\|z\|_1 = K$, the ones minimizing g are exactly the ones that have nonzero entries z_i only where $|\nabla_i f(x)| = \|\nabla f(x)\|_\infty$. To see this, first observe that every such z that minimizes g has $\operatorname{sgn}(z_i) \neq \operatorname{sgn}(\nabla_i f(x))$ whenever both signs are nonzero (otherwise, we could decrease g by flipping the sign of z_i). Now suppose there is $z_i \neq 0$ for some i such that $|\nabla_i f(x)| < \|\nabla f(x)\|_\infty$, and let j be such that $\left|\nabla_j f(\mathbf{x})\right| = \|\nabla f(\mathbf{x})\|_\infty$. Then we can decrease $|z_i|$ and increase $\left|z_j\right|$ accordingly such that g decreases. On the other hand, having nonzero values only where $|\nabla_i f(x)| = \|\nabla f(\mathbf{x})\|_\infty$, we have $\nabla f(x)^\top z = K\|\nabla f(x)\|_\infty$. Knowing this, it follows that the minimum of g under the constraint $\|z\|_1 = K$ is

$$q(K) = K\|\nabla f(x)\|_\infty + \frac{\mu_1}{2}K^2$$

This is minimized by $K^* = -\|\nabla f(\mathbf{x})\|_\infty/\mu_1$ and

$$q\left(K^\star\right) = -\frac{1}{2\mu_1}\|\nabla f(x)\|_\infty^2$$

Thm 4b.10 Steeper coordinate descent

For all t:
Coordinate-wise sufficient decrease for $i = \operatorname{argmax}_{i\in[d]}|\nabla_i f(\mathbf{x}_t)|$:

$$\begin{aligned}f(\mathbf{x}_{t+1}) &\le f(\mathbf{x}_t) - \frac{1}{2L}|\nabla_i f(\mathbf{x}_t)|^2 = f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|_\infty^2 \\ &\le f(\mathbf{x}_t) - \frac{\mu_1}{L}\left(f(\mathbf{x}_t) - f\left(\mathbf{x}^\star\right)\right). \quad (\text{PLineq wrt}\ell_\infty\text{-norm})\end{aligned}$$

Now it continues as for GD (subtracting $f\left(\mathrm{x}^\star\right)$ from both sides):

$$f(\mathbf{x}_{t+1}) - f\left(\mathbf{x}^\star\right) \le \left(1 - \frac{\mu_1}{L}\right)\left(f(\mathbf{x}_t) - f\left(\mathbf{x}^\star\right)\right),$$

5. Coordinate Descent

Lemma 5.2 Subgradients of Differentiable Func

Let g be a subgradient at $\mathbf{x}$. Suppose by contradiction that $\mathbf{g} \neq \nabla f(\mathbf{x})$. From the definition of g, for every $y \in \operatorname{dom}(f)$ we have

$$f(\mathbf{y}) \ge f(\mathbf{x}) + g^\top(\mathbf{y} - \mathbf{x}).$$

Since f is differentiable at $\mathbf{x}$, for every $\mathbf{y} \in \operatorname{dom}(f)$, we have

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + r(\mathbf{y} - \mathbf{x}),$$

where r is the error function s.t. $r(\mathbf{v}) \to 0$ as $\mathbf{v} \to 0$. Combining this two formulas, we have

$$(\mathbf{g} - \nabla f(\mathbf{x}))^\top(\mathbf{y} - \mathbf{x}) \le r(\mathbf{y} - \mathbf{x})$$

Take $\epsilon > 0$ small enough s.t. $\mathbf{y} = \mathbf{x} + \epsilon(\mathbf{g} - \nabla f(\mathbf{x})) \in \operatorname{dom}(f)$. Applying $\mathbf{y} = \mathbf{x} + \epsilon(\mathbf{g} - \nabla f(\mathbf{x}))$ to the formula above, we have

$$\epsilon\|\mathbf{g} - \nabla f(\mathbf{x})\|^2 \le r(\epsilon(\mathbf{g} - \nabla f(\mathbf{x}))).$$

Divide the inequality above by $\epsilon\|\mathbf{g} - \nabla f(\mathbf{x})\|$ and we have

$$\|\mathbf{g} - \nabla f(\mathbf{x})\| \le \frac{r(\epsilon(\mathbf{g} - \nabla f(\mathbf{x})))}{\epsilon\|\mathbf{g} - \nabla f(\mathbf{x})\|}$$

Note that the right hand side goes to 0 as $\epsilon \to 0$. Thus, by taking $\epsilon \to 0$, we have

$$\|\mathbf{g} - \nabla f(\mathbf{x})\| \le 0$$

This just shows that $\|\mathbf{g} - \nabla f(\mathbf{x})\| = 0$, which implies that $\mathbf{g} = \nabla f(\mathbf{x})$. Contradiction. Thus, we have $\partial f(\mathbf{x}) \subseteq \{\nabla f(\mathbf{x})\}$.

Lemma 5.6 Convex and Lipschitz continuity = bounded subgradients

We assume that $\operatorname{dom}(f) = \mathbb{R}^d$ and hint at the general case.
$(ii) \Longrightarrow (i)$: Given any $x \in \mathbb{R}^d$ (harder alternative: x in a convex domain $D = \operatorname{dom}(f)$), consider $\mathbf{g}$ an element of $\partial f(\mathbf{x})$. Let $\mathbf{z} = \mathbf{x} + \mathbf{g}$ (alternative: let $\eta > 0$ such that $\mathbf{z} = \mathbf{x} + \eta\mathbf{g}$ is still in D).
Since f is B-Lipschitz, we have

$$f(z) - f(x) \le B \cdot \|z - x\| = B \cdot \|g\|$$

(Alternative $\cdots \le \eta \cdot \|\mathbf{g}\|$)
Using the definition of subgradient, we have:

$$f(\mathbf{z}) - f(\mathbf{x}) \ge \mathbf{g}^\top(\mathbf{z} - \mathbf{x}) = \|\mathbf{g}\|^2$$

(Alternative: $\cdots \ge \eta \cdot \|\mathbf{g}\|^2$)
Combining the inequalities, we have $\|\mathbf{g}\| \le B$ (the η is simplified on both sides in the alternative situation when x is drawn from a domain D and not from all $\mathbb{R}^d$ and we get the same result.)
$(i) \Longrightarrow (ii)$: Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and let $\mathbf{g}$ be any element in $\partial f(\mathbf{x})$, by definition of subgradient we have: $f(\mathbf{y}) - \mathbf{f}(\mathbf{x}) \ge \mathbf{g}^\top(\mathbf{y} - \mathbf{x})$, therefore, by inversing the signs in the inequality, then using Cauchy-Schwarz and finally the bound on the norm of the subgradient, we have:

$$\begin{aligned}f(\mathbf{x}) - f(\mathbf{y}) &\le g^\top(\mathbf{x} - \mathbf{y}) \\ &\le \|\mathbf{g}\| \cdot \|\mathbf{x} - \mathbf{y}\| \\ &\le B \cdot \|\mathbf{x} - \mathbf{y}\|.\end{aligned}$$

Note that $f(\mathbf{y}) - f(\mathbf{x}) \le B \cdot \|\mathbf{y} - \mathbf{x}\|$ follows from a similar proof. Using these two inequalities, we can conclude that (ii) holds.
Note: in the case where f is defined on a convex domain D, the latter is assumed to be open in the alternative situation described above. If not, the reasoning applies for any $\mathbf{x}$ in the interior of D.

Lemma 5.7 Subgradient optimality condition

By definition of subgradients, $\mathbf{g} = \mathbf{0} \in \partial f(\mathbf{x})$ gives

$$f(\mathbf{y}) \ge f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y} - \mathbf{x}) = f(\mathbf{x})$$

for all $\mathbf{y} \in \operatorname{dom}(f)$, so $\mathbf{x}$ is a global minimum.

Lemma 5.8 Basic Descent Lemma

Asymptotic Convergence under Different Stepsizes

Take constant stepsize $\gamma_t \equiv \gamma$ as an example. By Thm 5.9,

$$\begin{aligned}\lim_{T\to\infty}\min_{1\le t\le T} f(\mathbf{x}_t) - f^* &\le \frac{\|\mathbf{x}_1-\mathbf{x}^*\|_2^2 + \sum_{t=1}^T \gamma_t^2 \|\mathbf{g}_t\|_2^2}{2\sum_{t=1}^T \gamma_t} \\ &\le \lim_{T\to\infty} \frac{R^2}{2\gamma T} + \frac{\gamma^2 B^2 T}{2\gamma T} \\ &= \lim_{T\to\infty} \frac{R^2}{2\gamma T} + \frac{\gamma B^2}{2} \\ &= \frac{\gamma B^2}{2}\end{aligned}$$

Corollary 5.10 Convergence Rate for Convex Lipschitz Problem

At first, we want to prove $\min_{1\le t\le T} f(\mathbf{x}_t) - f^* \le \mathcal{O}\left(\frac{BR\ln(T)}{\sqrt{T}}\right)$

From Thm 5.9, we know that

$$\min_{1\le t\le T} f(x_t) - f^* \le \frac{\|x_1 - x^*\|^2 + \sum_{t=1}^T \gamma_t^2 \|\mathbf{g}_t\|^2}{2\sum_{t=1}^T \gamma_t}$$

Replacing $\|g_t\|^2$ by the upper bound $\frac{R}{B\sqrt{t}}$ and then using the fact that $\sum_{t=1}^T 1/\sqrt{t} = \mathcal{O}(\sqrt{T})$ and $\sum_{t=1}^T 1/t = \mathcal{O}(\ln T)$, we can derive the first.

Then we want to prove $\min_{1\le t\le T} f(x_t) - f^* \le \mathcal{O}\left(\frac{BR}{\sqrt{T}}\right)$

We can simply ignore the contribution of the first $T/2$ steps. Since all the iterates are inside X, we know that $\|\mathbf{x}_{T/2} - \mathbf{x}^*\|^2 \le R^2$. Then, we apply the equation above on the last $T/2$ iterates and get the result.

Thm 5.12

6. Stochastic Optimization

Thm 6.1 Convex, weighted averaging

First, $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 = \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\gamma_t \nabla f(\mathbf{x}_t, \xi_t)^T (\mathbf{x}_t - \mathbf{x}^*) + \gamma_t^2 \|\nabla f(\mathbf{x}_t, \xi_t)\|_2^2$. By law of total expectation,

$$\begin{aligned}\mathbb{E}\left[\nabla f(\mathbf{x}_t,\xi_t)^T(\mathbf{x}_t-\mathbf{x}^*)\right] &= \mathbb{E}\left[\mathbb{E}\left[\nabla f(\mathbf{x}_t,\xi_t)^T(\mathbf{x}_t-\mathbf{x}^*) \mid \mathbf{x}_t\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\nabla f(\mathbf{x}_t,\xi_t)\mid \mathbf{x}_t\right]^T(\mathbf{x}_t-\mathbf{x}^*)\right] \\ &= \mathbb{E}\left[\nabla F(\mathbf{x}_t)^T(\mathbf{x}_t-\mathbf{x}^*)\right] \\ &\ge \mathbb{E}\left[F(\mathbf{x}_t) - F(\mathbf{x}^*)\right]\end{aligned}$$

This leads to the recursion:

$$\gamma_t \mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] \le \frac{1}{2}\mathbb{E}\left[\|\mathbf{x}_t-\mathbf{x}^*\|_2^2\right] - \frac{1}{2}\mathbb{E}\left[\|\mathbf{x}_{t+1}-\mathbf{x}^*\|_2^2\right] + \frac{1}{2}\gamma_t^2 B^2$$

The result follows by telescoping the sum from $t = 1$ to T.

Thm 6.2 Strong convex, diminishing stepsize, last iterate

First, $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 = \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\gamma_t \nabla f(\mathbf{x}_t, \xi_t)^T (\mathbf{x}_t - \mathbf{x}^*) + \gamma_t^2 \|\nabla f(\mathbf{x}_t, \xi_t)\|_2^2$. By law of total expectation and strong convexity,

$$\mathbb{E}\left[\nabla f(\mathbf{x}_t,\xi_t)^T(\mathbf{x}_t-\mathbf{x}^*)\right] = \mathbb{E}\left[\nabla F(\mathbf{x}_t)^T(\mathbf{x}_t-\mathbf{x}^*)\right] \ge \mu\mathbb{E}\left[\|\mathbf{x}_t-\mathbf{x}^*\|_2^2\right]$$

This leads to the recursion:

$$\mathbb{E}\left[\|\mathbf{x}_{t+1}-\mathbf{x}^*\|_2^2\right] \le \left(1 - \frac{2\mu\gamma}{t}\right)\mathbb{E}\left[\|\mathbf{x}_t-\mathbf{x}^*\|_2^2\right] + \frac{\gamma^2 B^2}{t^2}$$

The result follows by induction.

7. Variance-reduced Stochastic Methods

Thm 7.1 Convergence of SVRG

Handout07 Pages 40-41.

Lemma 7.2 Property of Smoothness

Hw7 Ex1

Lemma 7.3 Bound of Variance

Hw7 Ex1

8. Nonconvex Functions

Lemma 8.1 Bounded Hessians ⇒ smooth

Handout08 Pages 8-10

Thm 8.2 Gradient descent on smooth (not necessarily convex) functions

Handout08 Pages 15

Corollary of Thm 8.2

Hw8 Ex1

Lemma 8.3 No overshooting

Hw8 Ex2

Lemma 8.5 Balanced iterates

Handout08 Pages 36

Lemma 8.6

Handout08 Pages 38

Lemma 8.7

Handout08 Pages 39

Lemma 8.8

Handout08 Pages 40

Thm 8.9 Convergence of Balanced Iterates

Handout08 Pages 42

Corollary of Thm 8.9

Hw8 Ex4

9. The Frank-Wolfe Algorithm

Lemma 9.1

Handout09 Page 11

Lemma 9.2

Handout09 Page 13

Thm 9.3 Convergence in $\mathcal{O}(1/\varepsilon)$ steps

Handout09 Page 16 + Hw9 Ex1

Lemma 9.4 Descent Lemma

Handout09 Page 15

Thm 9.5 Convergence in terms of the curvature constant

Handout09 Page 23

Lemma 9.6 Relating Curvature and Smoothness

Hw9 Ex2

10. Newton's Method and Quasi-Newton Methods

Lemma 10.1 Convergence in one step on quadratic functions

Handout10 Page 8

Lemma 10.3 Minimizing the second-order Taylor approximation

Hw10 Ex2

Thm 10.4 Convergence Thm

Handout10 Pages 15-17

Lemma 10.7 Strong convexity ⇒ Bounded inverse Hessians

Hw10 Ex3

11. Modern Second-Order Methods and Nonconvex Optimization

Lipschitz Hessian

Hw11 Ex4

Global analysis for strongly-convex smooth objectives

Handout11 Page 9

Thm 11.1 Convergence of Nonconvex SGD

Handout11 Page 27

12. Modern Nonsmooth Optimization

Lemma 12.1 Three Point Identity

Handout12 Page 49

Lemma 12.2

Handout12 Page 19

Theorem 12.7 Convergence of PPA

Handout12 Page 41

13. Min-Max Optimization

Thm 12.5 Convergence of GDA for SC-SC Setting

Handout13 Page 49, Pages 25-26

Thm 12.6 EG for C-C Setting

Handout13 Pages 50-53