Chapter 4 Projected Gradient Descent

Algorithm

- Gradient Step: $\mathbf{y}_{t+1} := \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t)$
- Projection Step: $\mathbf{x}_{t+1} := \Pi_X(\mathbf{y}_{t+1}) := \underset{\mathbf{x} \in X}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{y}_{t+1}\|^2$
  - We assume the minimization in projection step is easy to solve.
  - $d_{\mathbf{y}}(\mathbf{x}) := \|\mathbf{x} - \mathbf{y}\|^2$ is strongly convex -> projection unique.
- Fact 4.1 Let $X \subseteq \mathbb{R}^d$ be closed and convex, $\mathbf{x} \in X, \mathbf{y} \in \mathbb{R}^d$. Then
  - (i) $(\mathbf{x} - \Pi_X(\mathbf{y}))^\top (\mathbf{y} - \Pi_X(\mathbf{y})) \leq 0$
  - (ii) $\|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2$ (*Note this is square of distance*)
  - Proof
    - (i) $\nabla d_{\mathbf{y}}(\mathbf{x}) = \mathbf{x} - \mathbf{y}$, With Lemma 2.27 $\forall \mathbf{x} \in X, \nabla f(\mathbf{x}^\star)^\top (\mathbf{x} - \mathbf{x}^\star) \geq 0$ (also holds for closed set) and $\Pi_X(\mathbf{y})$ as minimum, we get $(\Pi_X(\mathbf{y}) - \mathbf{y})^\top (\mathbf{x} - \Pi_X(\mathbf{y})) \geq 0$.
    - (i) -> (ii) By $2\mathbf{v}^\top \mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2$.
- Lemma (Ex 31) If $\mathbf{x}_{t+1} = \mathbf{x}_t$, then $\mathbf{x}_t$ is minimizer.
  - Proof
    - Let $\mathbf{y} \leftarrow \mathbf{x}_t - \gamma \mathbf{g}_t$ we have $\Pi_X(\mathbf{y}) := \mathbf{x}_{t+1} = \Pi_X(\mathbf{x}_t - \gamma \mathbf{g}_t) = \mathbf{x}_t$.
    - By (i) $(\mathbf{x} - \Pi_X(\mathbf{y}))^\top (\mathbf{y} - \Pi_X(\mathbf{y})) = (\mathbf{x} - \mathbf{x}_t)^\top (-\gamma \mathbf{g}_t) \leq 0 \Leftrightarrow \nabla f(\mathbf{x}_t)^\top (\mathbf{x} - \mathbf{x}_t) \geq 0$ -> Lemma 2.27 -> minimizer.

Bounded gradients: $\mathcal{O}(1/\varepsilon^2)$ steps (SAME)

- Theorem 4.2 Same as unbounded case
  - Proof
    - Difference only in gradient step, original procedure gives $\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) = \frac{1}{2\gamma}(\gamma^2 \|\mathbf{g}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{y}_{t+1} - \mathbf{x}^\star\|^2)$
    - But we need $\mathbf{x}_{t+1}$ instead of $\mathbf{y}_{t+1}$. By fact 4.1 (ii) setting $\mathbf{x} = \mathbf{x}^\star, \mathbf{y} = \mathbf{y}_{t+1}$, we get $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{y}_{t+1} - \mathbf{x}^\star\|^2$
      - so we have $\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) \leq \frac{1}{2\gamma}(\gamma^2 \|\mathbf{g}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2)$

Smooth convex function: $\mathcal{O}(1/\varepsilon)$ steps (SAME)

- Lemma 4.3 (Sufficient descent under constraint) For L-smooth covnex function f, a step size of $\gamma = L^{-1}$ gives sufficient descent of $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2$.
  - Proof
    - $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 = f(\mathbf{x}_t) - L(\mathbf{y}_{t+1} - \mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2$
    - Use $2\mathbf{v}^\top \mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2$ on term $(\mathbf{y}_{t+1} - \mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t)$ we get
    - $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_t\|^2 + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2$
    - Then we arrive at our destination
  - PS (Ex 32): Since $\mathbf{x}_{t+1}$ is the minimzer of distance to $\mathbf{y}_{t+1}$, we have $\|\mathbf{y}_{t+1} - \mathbf{x}_t\|^2 \geq \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2$, therefore from the last inequality, $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t)$
- Lemma 4.4 (Error Bound) $f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$ (*same as unbounded case*).
  - Proof
    - Since we have an additional term $\frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 \leq f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2$ we have to find some compensate in GD algorithm.
    - We have $\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) = \frac{1}{2\gamma}(\gamma^2 \|\mathbf{g}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{y}_{t+1} - \mathbf{x}^\star\|^2)$,
      - since $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 + \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2 \leq \|\mathbf{y}_{t+1} - \mathbf{x}^\star\|^2$, we can upper bound it by
      - $\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) \leq \frac{1}{2\gamma}(\gamma^2 \|\mathbf{g}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 - \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2)$
    - with convexity and sum over all t, we get $\sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq \sum_{t=0}^{T-1} \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) \leq \frac{1}{2L}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2 + \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 - \frac{L}{2}\sum_{t=0}^{T-1}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2$
    - with new version of sufficient decrease we have $\frac{1}{2L}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2 = f(\mathbf{x}_0) - f(\mathbf{x}_T) + \frac{L}{2}\sum_{t=0}^{T-1}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2$
    - then we can prove the claim.

Smooth and strongly convex f: $\mathcal{O}(\log(1/\varepsilon))$ steps (SAME)

- Theorem 4.5 (*similar to theorem 4.3*)
  - (i) Geometric decrease for $\|\mathbf{x}_t-\mathbf{x}^\star\|^2$, $\|\mathbf{x}_{t+1}-\mathbf{x}^\star\|^2 \le \left(1-\frac{\mu}{L}\right)\|\mathbf{x}_t-\mathbf{x}^\star\|^2$
  - (ii) Exponential decrease for absoulute error $f(\mathbf{x}_T)-f(\mathbf{x}^\star) \le \frac{L}{2}\left(1-\frac{\mu}{L}\right)^T\|\mathbf{x}_0-\mathbf{x}^\star\|^2 + \|\nabla f(\mathbf{x}^\star)\|\left(1-\frac{\mu}{L}\right)^{T/2}\|\mathbf{x}_0-\mathbf{x}^\star\|$
  - Proof
    - with strong convexity, we can bound gradient to $\mathbf{g}_t^\top(\mathbf{x}_t-\mathbf{x}^\star) \le \frac{1}{2\gamma}\left(\gamma^2\|\nabla f(\mathbf{x}_t)\|^2+\|\mathbf{x}_t-\mathbf{x}^\star\|^2-\|\mathbf{x}_{t+1}-\mathbf{x}^\star\|^2-\|\mathbf{y}_{t+1}-\mathbf{x}_{t+1}\|^2\right)-\frac{\mu}{2}\|\mathbf{x}_t-\mathbf{x}^\star\|^2$
    - with convexity $f(\mathbf{x}_t)-f(\mathbf{x}^\star)\le \mathbf{g}_t^\top(\mathbf{x}_t-\mathbf{x}^\star)$ we can bound on $\|\mathbf{x}_{t+1}-\mathbf{x}^\star\|^2$
    - $\|\mathbf{x}_{t+1}-\mathbf{x}^\star\|^2 \le (1-\mu\gamma)\|\mathbf{x}_t-\mathbf{x}^\star\|^2 + 2\gamma(f(\mathbf{x}^\star)-f(\mathbf{x}_t))+\gamma^2\|\nabla f(\mathbf{x}_t)\|^2-\|\mathbf{y}_{t+1}-\mathbf{x}_{t+1}\|^2$ (*geometric decrease with some noise*)
      - The additional term is bound to be non-positive by the adapted version of sufficient descent (Lemma 4.3) $\frac{2}{L}(f(\mathbf{x}^\star)-f(\mathbf{x}_t))+\frac{1}{L^2}\|\nabla f(\mathbf{x}_t)\|^2-\|\mathbf{y}_{t+1}-\mathbf{x}_{t+1}\|^2\le 0.$
      - Then we get (i)
    - (ii) is attained by smoothness, but the gradient term does not vanish, $f(\mathbf{x}_T)-f(\mathbf{x}^\star)\le \nabla f(\mathbf{x}^\star)^\top(\mathbf{x}_T-\mathbf{x}^\star)+\frac{L}{2}\|\mathbf{x}^\star-\mathbf{x}_T\|^2$ $\le \|\nabla f(\mathbf{x}^\star)\|\|\mathbf{x}_T-\mathbf{x}^\star\|+\frac{L}{2}\|\mathbf{x}^\star-\mathbf{x}_T\|^2 \le \|\nabla f(\mathbf{x}^\star)\|\left(1-\frac{\mu}{L}\right)^{T/2}\|\mathbf{x}_0-\mathbf{x}^\star\|+\frac{L}{2}\left(1-\frac{\mu}{L}\right)^T\|\mathbf{x}_0-\mathbf{x}^\star\|^2$

PGD on ℓ_1-Ball

- Definition An ℓ_1-Ball of radius R is $X=B_1(R):=\left\{\mathbf{x}\in\mathbb{R}^d:\|\mathbf{x}\|_1=\sum_{i=1}^d|x_i|\le R\right\}$
- Fact 4.6 By suitable scaling and sign flipping of coordinates, we can assume $R=1$ and for the point $\mathbf{v}$ to be projected, each component $v_i\ge 0$, and the non-trivil case is when $\sum_i v_i>1$.
- Fact 4.7 Under Fact 4.6, the projected point $\mathbf{x}=\Pi_X(\mathbf{v})$ satisfies (i) $x_i\ge 0$ (ii) $\sum_{i=1}^d x_i=1$.
  - Proof
    - (i) Otherwise $(-x_i-v_i)^2\le(x_i-v_i)^2$ if $x_i<0$, then sign-flipping can get better result.
    - (ii) If $\sum_{i=1}^d x_i<1$, then for some small $\lambda>0$ still $\mathbf{x}'=\mathbf{x}+\lambda(\mathbf{v}-\mathbf{x})\in X$, then $\|\mathbf{x}'-\mathbf{v}\|=(1-\lambda)\|\mathbf{x}-\mathbf{v}\|$ is smaller.
- Collary 4.8 (4.6 + 4.7) $\Pi_X(\mathbf{v})=\underset{\mathbf{x}\in\Delta_d}{\operatorname{argmin}}\|\mathbf{x}-\mathbf{v}\|^2$ where $\Delta_d:=\left\{\mathbf{x}\in\mathbb{R}^d:\sum_{i=1}^d x_i=1, x_i\ge 0\forall i\right\}$ is the standard simplex.
- Fact 4.9 By switching coordinates, we can assume $v_1\ge v_2\ge\cdots\ge v_d$.
- Lemma 4.10 Let $\mathbf{x}^\star:=\operatorname{argmin}_{\mathbf{x}\in\Delta_d}\|\mathbf{x}-\mathbf{v}\|^2$. Under Fact 4.9, there exists (a unique) $p\in\{1,\dots,d\}$ such that $x_i^\star>0, i\le p$ and $x_i^\star=0, i>p$.
  - Proof
    - By Lemma 2.27, the optimal condition is $\forall x in\Delta_d$, $\nabla d_{\mathbf{v}}(\mathbf{x}^\star)^\top(\mathbf{x}-\mathbf{x}^\star)=2(\mathbf{x}^\star-\mathbf{v})^\top(\mathbf{x}-\mathbf{x}^\star)\ge 0$
    - Since $\sum_{i=1}^d x_i^\star=1$, at least one $x_i>0$.
    - If we have $x_i^\star=0$ and $x_{i+1}^\star>0$, we construct an $\mathbf{x}$ s.t. $x_{i+1}^\star-x_{i+1}=x_i-x_i^\star=\varepsilon$, for small enough ε, we can ensure $\mathbf{x}\in\Delta_d$.
      - then we take this into optimial condition and get $(\mathbf{x}^\star-\mathbf{v})^\top(\mathbf{x}-\mathbf{x}^\star)=(0-v_i)\varepsilon-(x_{i+1}^\star-v_{i+1})\varepsilon=\varepsilon(\underbrace{v_{i+1}-v_i}_{\le 0}-\underbrace{x_{i+1}^\star}_{>0})<0$ which leads to contradictory.
- Lemma 4.11 Under Fact 4.9, we further have $x_i^\star=v_i-\Theta_p, i\le p$ where $\Theta_p=\frac{1}{p}\left(\sum_{i=1}^p v_i-1\right)$.
  - Proof
    - If not all $x_i^\star-v_i, i\le p$ is the same, then we must have $x_i^\star-v_i<x_j^\star-v_j$ for some $i,j\le p$. Similar to 4.10, we set $\mathbf{x}$ to be $x_j^\star-x_j=x_i-x_i^\star=\varepsilon$ for some small enough ε
    - then $(\mathbf{x}^\star-\mathbf{v})^\top(\mathbf{x}-\mathbf{x}^\star)=(x_i^\star-v_i)\varepsilon-\left(x_j^\star-v_j\right)\varepsilon=\varepsilon\underbrace{\left((x_i^\star-v_i)-(x_j^\star-v_j)\right)}_{<0}<0$
    - then we can compute Θ_p by $1=\sum_{i=1}^p x_i^\star=\sum_{i=1}^p(v_i-\Theta_p)=\sum_{i=1}^p v_i-p\Theta_p$.
  - Therefore, the solution is of the form $\mathbf{x}^\star(p):=(v_1-\Theta_p,\dots,v_p-\Theta_p,0,\dots,0)$ and $v_p-\Theta_p>0$.
  - The total sorting and comparison of maximum $\|\mathbf{x}^\star(p)-\mathbf{V}\|^2$ takes $\mathcal{O}(d\log d)$
  - The following lemma show comparing $\|\mathbf{x}^\star(p)-\mathbf{V}\|^2$ is not necessary.
- Lemma 4.12 Finding $p^\star:=\max\left\{p\in\{1,\dots,d\}:v_p-\frac{1}{p}\left(\sum_{i=1}^p v_i-1\right)>0\right\}$ is enough, $\underset{\mathbf{x}\in\Delta_d}{\operatorname{argmin}}\|\mathbf{x}-\mathbf{v}\|^2=\mathbf{x}^\star(p^\star)$.
  - Proof
    - We can show $\|\mathbf{x}^\star(p)-\mathbf{v}\|^2$ is non-increasing w.r.t. p.

- $\|\mathbf{x}^\star(p)-\mathbf{v}\|^2-\|\mathbf{x}^\star(p+1)-\mathbf{v}\|^2=v_{p+1}^2+\sum_{i=1}^p(x_i^\star(p)-v_i)^2-\sum_{i=1}^{p+1}(x_i^\star(p+1)-v_i)^2$
  $=v_{p+1}^2+p\Theta(p)^2-(p+1)\Theta(p+1)^2$
- Denote $\Delta:=\sum_{i=1}^p v_i-1$ then
- $\|\mathbf{x}^\star(p)-\mathbf{v}\|^2-\|\mathbf{x}^\star(p+1)-\mathbf{v}\|^2=v_{p+1}^2+\Delta^2/p-(v_{p+1}+\Delta)^2/(p+1)=\dfrac{(pv_{p+1}-\Delta)^2}{p(p+1)}\geq 0.$
- Then we can simply find the maximum p with $v_p-\Theta_p>0$.