Chapter 8 The Frank-Wolfe Algorithm

- Definition (LMO) *linear minimization oracle* $\mathrm{LMO}_X(\mathbf{g}) := \underset{\mathbf{z}\in X}{\operatorname{argmin}}\, \mathbf{g}^\top \mathbf{z}$.
  - This exists when X is bounded and closed.
- Algo $\mathbf{s} := \mathrm{LMO}_X\left(\nabla f\left(\mathbf{x}_t\right)\right)$ and $\mathbf{x}_{t+1} := (1-\gamma_t)\mathbf{x}_t + \gamma_t \mathbf{s}$ and $\gamma_t \in [0,1]$.
  - Reduce non-linear to linear problem.
- Properties
  - If X convex, then iterates are *always feasible*, $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_t \in X$.
  - *Projection-free* solving a linear program instead of quadratic program
  - *sparse representation* $\mathbf{x}_t$ is always a convex combination of initial iterate and minimizers used so far.

Cases when LMO is simple to compute

- FW algo is useful when X can be described as a convex hull of a finite or other nise set of *atom* points $\mathcal{A}$, $X := \mathbf{conv}(\mathcal{A})$.
  - $\mathbf{s} = \sum_{i=1}^n \lambda_i \mathbf{a}_i$, where $\sum_{i=1}^n \lambda_i = 1$ and all non-negative.
  - Then if $\mathbf{s}$ minimize $\mathbf{g}^\top \mathbf{z}$, then there is also an atomic minimizer.
  - $\mathcal{A} = X$ is the trivial case, we are interested in *extreme points* where $\mathbf{x} \notin \mathbf{conv}(X \backslash \{\mathbf{x}\})$.

LASSO with ℓ_1-ball

- Problem: $\min_{\mathbf{x}\in\mathbb{R}^d} \|A\mathbf{x}-\mathbf{b}\|^2$ s.t. $\|\mathbf{x}\|_1 \le 1$, we see that $X = \mathbf{conv}\left(\{\pm \mathbf{e}_1, \ldots, \pm \mathbf{e}_d\}\right)$.
  - It is easy to show $\mathrm{LMO}_X(\mathbf{g}) = -\operatorname{sgn}\left(g_i\right)\mathbf{e}_i \text{ with } i := \underset{i\in[d]}{\operatorname{argmax}} |g_i|$.

Semidefinite Programming and the Spectahedron

- Problem: $\arg\min_Z G \bullet Z$ s.t. $\mathrm{Tr}(Z) = 1$ and $Z \succeq 0$, where G semetric, and $\bullet$ stands for scalor product.
  - Feasible region X is called *Spectahedron*
- Since every semetric matrix can be decomposed into $C^\top C = \sum_{i\in[d]} z_i z_i^\top$, natually the atom is $\mathbf{Z}\mathbf{Z}^\top$ where $\mathbf{z} \in \mathbb{R}^d, \|\mathbf{z}\| = 1$.
- Lemma 8.1 Let λ_1 be the smallest eigenvalue of G, and let $\mathbf{s}_1$ be a corresponding eigenvector of unit length. Then we can choose $\mathrm{LMO}_X(G) = \mathbf{s}_1 \mathbf{s}_1^\top$.
  - Proof $\min_{\mathrm{Tr}(Z)=1, Z\succeq 0} G \bullet Z = \min_{\|\mathbf{z}\|=1} G \bullet \mathbf{z}\mathbf{z}^\top = \min_{\|\mathbf{z}\|=1} \mathbf{z}^\top G \mathbf{z} = \lambda_1$

Matrix completion (Exercise 54)

- Problem: $\min_{Y\in X\subseteq\mathbb{R}^{n\times m}} \sum_{(i,j)\in\Omega} \left(Z_{ij} - Y_{ij}\right)^2$ where $X := \mathbf{conv}(\mathcal{A})$ with $\mathcal{A} := \left\{\mathbf{u}\mathbf{v}^\top \mid \mathbf{u} \in \mathbb{R}^n, \|\mathbf{u}\|_2 = 1, \mathbf{v} \in \mathbb{R}^m, \|\mathbf{v}\|_2 = 1\right\}$
- F-W step: $\partial_{Y_{ij}} = Y_{ij} - Z_{ij}$, $\mathrm{LMO}_X = \arg\min_X \sum_{ij}(Y_{ij} - Z_{ij})(\mathbf{u}\mathbf{v}^\top)_{i,j} = \mathbf{u}^\top (Y-Z)\mathbf{v}$
  - consider the SVD of $Y-Z$, $Y - Z = U\Sigma V^\top$, where $\Sigma \in \mathbb{R}^{n\times m}$ is diagnal.
  - Then $U^\top u$ and $V^\top v$ also norm-1. This gives solution of $k := \arg\max_{i\in[\min\{n,m\}]} \sigma(\Sigma)_i$, and $u = U\mathbf{e}_k, v = -V\mathbf{e}_k$
  - $\mathbf{u}\mathbf{v}^\top = -U\mathbf{E}_{kk}V^\top$
- PS: Matrix completion has been removed from this course, so I don't know the normal procedure for projection...

Duality gap, A certificate for optimization quality

- Definition (Duality gap) Given $\mathbf{x} \in X$ the *duality gap(Hearn gap)* is $g(\mathbf{x}) := \nabla f(\mathbf{x})^\top(\mathbf{x} - \mathbf{s})$ where $\mathbf{s} := \mathrm{LMO}_X(\nabla f(\mathbf{x}))$.
- Lemma 8.2 Suppose there is a minimizer for F-W algo,$\mathbf{x}^\star$, f-convex. Let $\mathbf{x} \in X$. Then $g(\mathbf{x}) \ge f(\mathbf{x}) - f\left(\mathbf{x}^\star\right)$.
  - Proof $g(\mathbf{x}) = \nabla f(\mathbf{x})^\top(\mathbf{x}-\mathbf{s}) \ge \nabla f(\mathbf{x})^\top\left(\mathbf{x}-\mathbf{x}^\star\right) \ge f(\mathbf{x}) - f\left(\mathbf{x}^\star\right) \ge 0$
  - $g(\mathbf{x}^\star) = 0$ since $\nabla f(\mathbf{x}^\star)^\top\left(\mathbf{x}-\mathbf{x}^\star\right) \ge 0, \forall \mathbf{x} \in X$ and this means $g(\mathbf{x}^\star) \le 0$.
  - Note that $g(\mathbf{x}) \le \|\nabla f(\mathbf{x})\|_{a*}\|\mathbf{x}-\mathbf{s}\|_a$.

Convegence in $\mathcal{O}(1/\varepsilon)$ Steps

- Interestingly, step size can be set to be unrelated to smooth constant.

Case for $\gamma_t = 2/(t+2)$

- Lemma 8.4 (Descent Lemma) For a step step $\mathbf{x}_{t+1} := \mathbf{x}_t + \gamma_t (\mathbf{s} - \mathbf{x}_t)$ with $\gamma_t \in [0,1]$, we have $f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - \gamma_t g(\mathbf{x}_t) + \gamma_t^2 \frac{L}{2} \|\mathbf{s} - \mathbf{x}_t\|^2$, where $\mathbf{s} = \mathrm{LMO}_X(\nabla f(\mathbf{x}_t))$.
  - Proof
    - Smoothness $f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) + \gamma_t \nabla f(\mathbf{x}_t)^\top (\mathbf{s} - \mathbf{x}_t) + \frac{L}{2}\gamma_t^2 \|\mathbf{s} - \mathbf{x}_t\|_a^2 = f(\mathbf{x}_t) - \gamma_t g(\mathbf{x}_t) + \frac{L}{2}\gamma_t^2 \|\mathbf{s} - \mathbf{x}_t\|_a^2$.
- Theorem 8.3 If f convex and L-smooth, X convex and bounded. With any start $\mathbf{x}_0 \in X$ and stepsize $\gamma_t = 2/(t+2)$ F-W algo gives $f(\mathbf{x}_T) - f(\mathbf{x}^\star) \le \frac{2L\,\mathrm{diam}(X)^2}{T+1}$ where $\mathrm{diam}(X) := \max_{\mathbf{x},\mathbf{y}\in X} \|\mathbf{x} - \mathbf{y}\|$.
  - Proof
    - By certificate property, $g(\mathbf{x}_t) \ge f(\mathbf{x}_t) - f(\mathbf{x}^\star)$, so $\Delta f(\mathbf{x}_{t+1}) \le (1-\gamma_t)\Delta f(\mathbf{x}_t) + \gamma_t^2 C$, where $C := \frac{L}{2}\|\mathbf{s} - \mathbf{x}_t\|^2$.
    - We assme $\Delta f(\mathbf{x}_t) \le \frac{4C}{t+1}$, this is true for $t=0$, since by smoothness $f(\mathbf{x}_0) \le f(\mathbf{x}^\star) + \frac{L}{2}\|\mathbf{x}^\star - \mathbf{x}_0\|_a^2$.
    - By spritis of induction, we assmue this holds for t and smaller, then for $t+1$ we have $\Delta f(\mathbf{x}_{t+1}) \le (1 - \frac{2}{t+2})\frac{4C}{t+1} + \frac{4}{(t+2)^2}C = \frac{4C}{t+2}\frac{t(t+2)+(t+1)}{(t+2)(t+1)} = \frac{4C}{t+2}\frac{t^2+3t+1}{t^2+3t+2} \le \frac{4C}{t+2}$

Other step size

- *Line search* $\gamma_t := \operatorname*{argmin}_{\gamma\in[0,1]} f((1-\gamma)\mathbf{x}_t + \gamma\mathbf{s})$. This can be guranteed to be faster than previous step size.
- *Gap-based* $\gamma_t := \min\left(\frac{g(\mathbf{x}_t)}{L\|\mathbf{s}-\mathbf{x}_t\|^2}, 1\right)$, this is the quadratic function minimizer for $\gamma_t \in [0,1]$, so definitely, it is better than $2/(t+2)$
  - $h(\mathbf{x}_{t+1}) \le \begin{cases} h(\mathbf{x}_t)\left(1 - \frac{\gamma_t}{2}\right), & \gamma_t < 1 \\ h(\mathbf{x}_t), & \gamma_t = 1 \end{cases}$. (This can be proved easily)

Affine invariance

- The upper bound seems to depends on the coordinate and changes under affine transform, but in reality, the algorithm objective $\nabla f'(\mathbf{x}')^\top \mathbf{z}'$ is unchanged under affine transform.
- This contradiction can be solved by defining a new *curvature constant* $C_{(f,X)} := \sup_{\substack{\mathbf{x},\mathbf{s}\in X, \gamma\in(0,1] \\ \mathbf{y}=(1-\gamma)\mathbf{x}+\gamma\mathbf{s}}} \frac{1}{\gamma^2}\left(f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x})\right)$ which is affine invariant. (PS: this is similar to Bregman div)
  - By this definition, we have $f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)^\top \gamma_t (\mathbf{x}_t - \mathbf{s}) + \gamma_t^2 C_{(f,X)}$
- Theorem 8.5 (*proof is similar*) $f(\mathbf{x}_T) - f(\mathbf{x}^\star) \le \frac{4C_{(f,X)}}{T+1}$.
  - no smoothness assumed.
- Lemma 8.6 (Exercise 52) Let f convex and L-smooth, then $C_{(f,X)}$ is a tighter constant, $C_{(f,X)} \le \frac{L}{2}\mathrm{diam}(X)^2$.
  - Proof $f(\mathbf{x} + \gamma(\mathbf{s}-\mathbf{x})) \le f(\mathbf{x}) + \gamma\nabla f(\mathbf{x})^\top(\mathbf{s}-\mathbf{x}) + \frac{\gamma^2 L}{2}\|\mathbf{s}-\mathbf{x}\|_a^2$
- All of the stepsize holds the following inequality $f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)^\top \mu_t (\mathbf{x}_t - \mathbf{s}) + \mu_t^2 C_{(f,X)}$ where $\mu_t := 2/(t+2)$

Convergence of duality gap

- Theorem 8.7 f convex and L-smooth, then choosing any of stepsize in $2/(t+2)$, line search or gap-based stepsize, F-W algo gives duality gap minimum such that $\exists t \in [1:T]$ s.t. $g(\mathbf{x}_t) \le \frac{27/2 \cdot C_{(f,X)}}{T+1}$.
  - Proof See Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization Appendix B for detail.
  - Looser Proof:
    - By descent lemma $\mu_t g(x_t) \le f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \mu_t^2 C_{(f,X)}$
    - sum over $t \in [\lfloor T/2 \rfloor : T]$, by the fact of $2(\ln(t+2) - \ln(t+3)) \le \mu_t = \frac{2}{t+2} \le (\ln(t+1) - \ln(t+2))$
      - $\sum_{t=\lfloor T/2\rfloor}^{T} \mu_t \ge 2\ln\frac{T+2}{\lfloor T/2\rfloor + 3}$
    - And the fact of $\mu_t^2 \le \frac{4}{(t+2)(t+1)} = \frac{4}{t+1} - \frac{4}{t+2}$, so that $\sum_{t=\lfloor T/2\rfloor}^{T} \mu_t^2 \le \frac{4}{T+1} - \frac{4}{\lfloor T/2\rfloor + 2}$
    - Also $f(\mathbf{x}_{\lfloor T/2\rfloor}) - f(\mathbf{x}_T) \le f(\mathbf{x}_{\lfloor T/2\rfloor}) - f(\mathbf{x}^\star) \le 4C_{(f,X)}/(\lfloor T/2\rfloor + 1)$
    - we have $\min_{t\in[\lfloor T/2\rfloor, T]} g(x_t) \le \dots \le \frac{2C_{(f,X)}}{\ln\frac{T+2}{\lfloor T/2\rfloor+3}}\left(\frac{1}{\lfloor T/2\rfloor+1} + \frac{1}{T+1} - \frac{1}{\lfloor T/2\rfloor+2}\right) = \frac{2C_{(f,X)}}{T+1}\frac{1 + \frac{T+1}{(\lfloor T/2\rfloor+1)(\lfloor T/2\rfloor+2)}}{\ln\frac{T+2}{\lfloor T/2\rfloor+3}}$
    - We get a similar conclusion, but loser, the constant is about, when $T > 3$, coefficient ≤ 15.