Labs
Optimization for Machine Learning
Spring 2021

EPFL
School of Computer and Communication Sciences
Martin Jaggi & Nicolas Flammarion
github.com/epfml/OptML_course

Problem Set 8 — Solutions (Frank-Wolfe)

Convergence of Frank-Wolfe

Exercise 1:

Assuming $h_0 \leq 2C$, and the sequence $h_0, h_1, \dots$ satisfies

$$h_{t+1} \leq (1-\gamma)h_t + \gamma^2 C \qquad t = 0, 1, \dots$$

for $\gamma = \frac{2}{t+2}$, prove that

$$h_t \leq \frac{4C}{t+2} \qquad t = 0, 1, \dots$$

Solution: Proof by induction. First $t = 0$, $h_0 \leq 2C$. For $t \geq 1$, we have

$$\begin{aligned} h_{t+1} \leq & \quad (1-\gamma_t)h_t + {\gamma_t}^2 C \\ = & \quad \big(1 - \tfrac{2}{t+2}\big)h_t + \big(\tfrac{2}{t+2}\big)^2 C \\ \leq & \quad \big(1 - \tfrac{2}{t+2}\big)\tfrac{4C}{t+2} + \big(\tfrac{2}{t+2}\big)^2 C \ , \end{aligned}$$

where in the last inequality we have used the induction hypothesis for h_t. Simply rearranging the terms gives

$$\begin{aligned} h_{t+1} \leq & \quad \tfrac{4C}{t+2}\big(1 - \tfrac{1}{t+2}\big) \ = \ \tfrac{4C}{t+2}\tfrac{t+2-1}{t+2} \\ \leq & \quad \tfrac{4C}{t+2}\tfrac{t+2}{t+3} \ = \ \tfrac{4C}{t+3} \ , \end{aligned}$$

which is our claimed bound for $t \geq 1$.

Applications of Frank-Wolfe

Exercise 2:

Derive the LMO formulation for matrix completion, that is

$$\min_{Y \in X \subseteq \mathbb{R}^{n \times m}} \ \sum_{(i,j) \in \Omega} (Z_{ij} - Y_{ij})^2$$

when $\Omega \subseteq [n] \times [m]$ is the set of observed entries from a given matrix Z.

Where our optimization domain X is the unit ball of the trace norm (or nuclear norm), which is defined the convex hull of the rank-1 matrices

$$X := conv(\mathcal{A}) \ \text{ with } \ \mathcal{A} := \left\{ \mathbf{u}\mathbf{v}^\top \ \middle| \ \begin{smallmatrix} \mathbf{u} \in \mathbb{R}^n, \, \|\mathbf{u}\|_2 = 1 \\ \mathbf{v} \in \mathbb{R}^m, \, \|\mathbf{v}\|_2 = 1 \end{smallmatrix} \right\} .$$

1. Derive the LMO for this set X for a gradient at iterate $Y \in \mathbb{R}^{n \times m}$.
2. Dervie the *projection* step onto X. How does the computational operations (or costs) needed to compute the LMO and the projection step compare?

Solution: Without loss of generality we assume $n \leq m$.

1. The gradient of the objective function is

$$\frac{\partial F}{\partial Y_{ij}} = \begin{cases} 2(Y_{ij} - Z_{ij}), & (i,j) \in \Omega \\ 0, & \text{otherwise.} \end{cases}$$

First $\forall\ \mathbf{X} \in \text{conv}(\mathcal{A})$, it can be written as a linear combination $\mathbf{X} = \sum_i c_i \mathbf{u}_i \mathbf{v}_i^\top$ where the sum of singular values $\sum_i c_i = 1$ and $c_i \geq 0\ \forall\ i$. Then for any $\mathbf{S} \in \mathbb{R}^{n \times m}$

$$\langle \mathbf{X}, \mathbf{S} \rangle = \sum_i c_i \langle \mathbf{u}_i \mathbf{v}_i^\top, \mathbf{S} \rangle \geq \sum_i c_i \min_{\mathbf{A} \in \mathcal{A}} \langle \mathbf{A}, \mathbf{S} \rangle = \min_{\mathbf{A} \in \mathcal{A}} \langle \mathbf{A}, \mathbf{S} \rangle.$$

It means the minimizer of LMO must be from $\mathcal{A}$

$$\begin{aligned} \text{LMO}(\nabla F(Y)) &= \operatorname*{argmin}_{\mathbf{X} \in X} \langle \mathbf{X}, \nabla F(Y) \rangle = \operatorname*{argmin}_{\mathbf{u}\mathbf{v}^\top \in \mathcal{A}} \langle \mathbf{u}\mathbf{v}^\top, \nabla F(Y) \rangle = 2 \operatorname*{argmax}_{\mathbf{u}\mathbf{v}^\top \in \mathcal{A}} \sum_{(i,j) \in \Omega} u_i v_j (Z_{ij} - Y_{ij}) \\ &= \operatorname*{argmax}_{\mathbf{u}\mathbf{v}^\top \in \mathcal{A}} \mathbf{u}^\top B \mathbf{v}, \end{aligned}$$

where the matrix B is

$$B_{ij} = \begin{cases} 2(Z_{ij} - Y_{ij}), & (i,j) \in \Omega, \\ 0, & \text{otherwise.} \end{cases}$$

Taking SVD-decomposition of $B = UDV^\top$, we get that

$$\mathbf{u}^\top B \mathbf{v} = \mathbf{u}^\top U D V^\top \mathbf{v} = \mathbf{u}^\top \sum_{i=1}^n \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \mathbf{v} = \sum_{i=1}^n \sigma_i \langle \mathbf{u}, \mathbf{u}_i \rangle \langle \mathbf{v}, \mathbf{v}_i \rangle$$

Since $\mathbf{u} = \sum_i \langle \mathbf{u}, \mathbf{u}_i \rangle \mathbf{u}_i$ and $1 = \|\mathbf{u}\|_2^2 \geq \sum_{i=1}^n \langle \mathbf{u}, \mathbf{u}_i \rangle^2$, then

$$\mathbf{u}^\top B \mathbf{v} \leq \sqrt{\sum_{i=1}^n \sigma_i \langle \mathbf{u}, \mathbf{u}_i \rangle^2 \sum_{i=1}^n \sigma_i \langle \mathbf{v}, \mathbf{v}_i \rangle^2} \leq \sigma_1 \sqrt{\sum_{i=1}^n \langle \mathbf{u}, \mathbf{u}_i \rangle^2 \sum_{i=1}^n \langle \mathbf{v}, \mathbf{v}_i \rangle^2} = \sigma_1$$

Hence the largest possible value is achieved by taking singular vectors corresponding to the largest singular value: $\mathbf{u} = \mathbf{u}_1$, $\mathbf{v} = \mathbf{v}_1$, then $\mathbf{u}^\top U D V^\top \mathbf{v} = \sigma_1$.

2. By the definition of projection, we have

$$\Pi_X(S) = \operatorname*{argmin}_{C \in X} \|C - S\|_F^2.$$

For any $C \in X = \text{conv}(\mathcal{A})$ we can perform SVD on $C = U_C \Sigma V_C^\top$ and represent C as the sum of rank 1 matrices

$$C = \sum_{i=1}^n \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$

where $\sum_{i=1}^n \sigma_i \leq 1$ and $\sigma_i \geq 0$, $\|\mathbf{u}_i\|^2 = 1$, $\|\mathbf{v}_i\|^2 = 1$ for all $i > 0$. So it is equivalent to consider the optimization over trace norm $\|C\|_* = \sum_{i=1}^n \sigma_i$,

$$\Pi_X(S) = \operatorname*{argmin}_{C \in X} \|C - S\|_F^2 = \operatorname*{argmin}_{\|C\|_* \leq 1} \|C - S\|_F^2.$$

We perform SVD on S and get $S = U \Lambda V^\top$. Using the properties of trace and unitary matrix U and V

$$\|UMV^\top\|_F^2 = tr((UMV^\top)^\top UMV^\top) = tr(VM^\top MV^\top) = tr(M^\top M) \qquad \forall\ M \in \mathbb{R}^{n \times m}$$

Therefore

$$\Pi_X(S) = \operatorname*{argmin}_{\|C\|_* \leq 1} \|C - S\|_F^2 = \operatorname*{argmin}_{\|C\|_* \leq 1} \|C - U\Lambda V^\top\|_F^2 = \operatorname*{argmin}_{\|C\|_* \leq 1} \|U^\top C V - \Lambda\|_F^2.$$

Let $\tilde{C} = U^\top C V$, we know $\|C\|_* = \|U\tilde{C}V^\top\|_* = \|\tilde{C}\|_*$ because multiplying a unitary matrix does not change its eigenvalues and thus their trace norm.

$$\Pi_X(S) = \operatorname*{argmin}_{\|\tilde{C}\|_* \leq 1} \|\tilde{C} - \Lambda\|_F^2.$$

Using the Von Neumann's trace inequality [1]

$$|\langle A, B\rangle| \leq \sigma_1(A)\sigma_1(B) + \cdots + \sigma_n(A)\sigma_n(B) \qquad \forall\ A, B \in \mathbb{R}^{n\times m}$$

and using the definition of $\|A\|_F^2 = \sum_{i=1}^n \sigma_i(A)^2$, we know that

$$\begin{aligned}
\|\tilde{C} - \Lambda\|_F^2 =& \langle \tilde{C} - \Lambda, \tilde{C} - \Lambda\rangle = \|\tilde{C}\|_F^2 - 2\langle \tilde{C}, \Lambda\rangle + \|\Lambda\|_F^2 \\
\geq& \|\tilde{C}\|_F^2 - 2\sum_{i=1}^n \sigma_i\Lambda_{ii} + \|\Lambda\|_F^2 \\
=& \sum_{i=1}^n \sigma_i^2 - 2\sum_{i=1}^n \sigma_i\Lambda_{ii} + \sum_{i=1}^n \Lambda_{ii}^2 \\
=& \sum_{i=1}^n (\sigma_i - \Lambda_{ii})^2 = \|\Sigma - \Lambda\|_F^2
\end{aligned}$$

which means the minimizer must be a diagonal matrix

$$\Pi_X(S) = \operatorname*{argmin}_{\|\tilde{C}\|_* \leq 1} \|\tilde{C} - \Lambda\|_F^2 = \operatorname*{argmin}_{\|\Sigma\|_* \leq 1} \|\Sigma - \Lambda\|_F^2.$$

Now that both Λ and Σ are diagonal matrices, the minimization problem is indeed for vectors

$$\Pi_X(S) = \operatorname*{argmin}_{\|\Sigma\mathbf{1}\|_1 \leq 1} \|\Sigma\mathbf{1} - \Lambda\mathbf{1}\|_2^2.$$

This is a projection of diagonal elements of D to the unit l_1 ball. We already know from Lemma 3.12 (section 3.5) of lecture notes that this is equal to

$$\Sigma_{ii}^* = \begin{cases} \Lambda_{ii} - \theta_p, & i < p \\ 0 & \text{otherwise} \end{cases},$$

where $\theta_p = \frac{1}{p}\left(\sum_{i=1}^p \Lambda_{ii} - 1\right)$ $p = \max\{p' \in \{1, \ldots, n\} : \Sigma_{pp} - \theta_p > 0\}$ (assuming that all Λ_{ii} are sorted in decedent order).

3. For a projection step we need to compute the full SVD-decomposition, which takes $O(mn^2)$, for LMO we need only top 1 singular vectors, which is much faster.

[1] Page 20: `https://www.epfl.ch/labs/anchp/wp-content/uploads/2018/10/lecture1-slides.pdf`