CSIE5410 Optimization algorithms

Lecture 8: Frank-Wolfe Method

Yen-Huan Li (yenhuan.li@csie.ntu.edu.tw)

29.11.2018

Department of Computer Science and Information Engineering
National Taiwan University

Abstract

- Consider the problem of minimizing a convex differentiable function on the Schatten 1-norm ball.
- If we do Bregman proximal gradient descent, then in each iteration, we need to compute the eigenvalue decomposition of the iterates, which is computationally too expensive.
- The Frank-Wolfe algorithm can avoid such a bottleneck.

Recommended reading

- M. Jaggi. 2013. Revisiting Frank-Wolfe: Projection-free sparse convex optimization.
- Greedy algorithms, Frank-Wolfe and friends—A modern perspective (NIPS 2013 Workshop Videos on Youtube).
- R. M. Freund and P. Grigas. 2016. New analysis and results for the Frank-Wolfe method.
- Yu. Nesterov. 2018. Complexity bounds for primal-dual methods minimizing the model of objective function.

Table of contents

Optimization with low-rank matrices

Problem setup

In this lecture, we consider the following problem

$$f^\star = \min_x \{ f(x) \mid x \in \mathcal{X} \},$$

for some convex differentiable function f, where $\mathcal{X}$ is the unit Schatten 1-norm ball, or is given by

$$\mathcal{X} := \left\{ X \in \mathbb{R}^{p\times p} \;\middle|\; X \geq 0, \operatorname{tr} X = 1 \right\}.$$

Question. Where do we see such a problem?

Problem 1: Quantum state tomography (1/4)

Axioms of quantum mechanics

- A *quantum state* is described by a *density matrix* $\rho \in \mathcal{X}$.
- An *observable* is described by a hermitian matrix $A \in \mathbb{R}^{p \times p}$.
- Let the eigenvalue decomposition of A be $A = \sum_{j=1}^{J} \lambda_j P_j$, where λ_j are eigenvalues, and P_j are projections. The *measurement outcome* is a random variable η, satisfying

$$\mathsf{P}\left(\eta = \lambda_j\right) = \mathrm{tr}(P_j \rho), \quad j = 1, \ldots, J.$$

Indeed, in quantum mechanics, $\mathbb{R}$ should be replaced by $\mathbb{C}$. We consider the real case for simplicity.

Problem. Let $\rho^\natural \in \mathbb{R}^{p \times p}$ be an unknown density matrix. Suppose we have n independent copies of $\rho^\natural$. We measure each of them using possibly different observables $A_1, \ldots, A_n$, and obtain independent random variables $\eta_1, \ldots, \eta_n$ as measurement outcomes.

How do we estimate $\rho^\natural$ given the observables and measurement outcomes?

Problem 1: Quantum state tomography (3/4)

Linear approximation approach. For every i, we write the eigenvalue decomposition $A_i = \sum_j \lambda_{i,j} P_{i,j}$. Then we have

$$\mathsf{E}\left[\eta_i\right] = \sum_j \lambda_{i,j} \operatorname{tr}(P_{i,j}\rho^\natural) = \operatorname{tr}\left[\left(\sum_j \lambda_{i,j} P_{i,j}\right)\rho^\natural\right] = \operatorname{tr}(A_i \rho^\natural).$$

Therefore, we can consider the estimator

$$\hat{\rho}_1 \in \arg\min_\rho \left\{ \frac{1}{2n} \sum_{i=1}^n \left(\eta_i - \operatorname{tr}(A_i \rho)\right)^2 \,\middle|\, \rho \in \mathcal{X} \right\}.$$

M. A. Nielsen and I. L. Chuang. 2010. *Quantum Computation and Quantum Information*.

Problem 1: Quantum state tomography (4/4)

Maximum-likelihood estimation approach. Suppose that η_i corresponds to the j_i-th eigenvalue of A_i. For every i, the likelihood function is given by

$$L_i(\rho) = \mathrm{tr}(P_{i,j_i}\rho), \quad \forall \rho \in \mathcal{X}.$$

Then, the maximum-likelihood estimator is given by

$$\hat{\rho}_2 \in \arg\min_{\rho} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log \mathrm{tr}\left(P_{i,j_i}\rho\right) \;\middle|\; \rho \in \mathcal{X} \right\}.$$

Z. Hradil. 1997. Quantum state estimation.

Problem 2: Low-rank matrix estimation (1/2)

Recall the matrix estimation problem:

Problem. Let $X^\natural \in \mathbb{R}^{p_1 \times p_2}$. Suppose that we observe

$$y_i := \operatorname{tr}(A_i X^\natural) + w_i, \quad i = 1, \dots, n,$$

for some matrices $A_1, \dots, A_n$, where w_i denote the additive noise. How do we estimate $X^\natural$ given $y_1, \dots, y_n$ and $A_1, \dots, A_n$?

Assumption. Assume that $X^\natural$ is low-rank.

Problem 2: Low-rank matrix estimation (2/2)

Penalized estimation approach. We have seen the estimator given by

$$\hat{X}_1 \in \arg\min_X \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathrm{tr}(A_i X))^2 + \lambda_n \|X\|_{S^1} \;\middle|\; X \in \mathbb{R}^{p_1 \times p_2} \right\},$$

for some properly chosen penalization parameter $\lambda_n > 0$.

Constrained estimation approach. Another closely-related estimator is given by

$$\hat{X}_2 \in \arg\min_X \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathrm{tr}(A_i X))^2 \;\middle|\; X \in \mathbb{R}^{p_1 \times p_2}, \|X\|_{S^1} \le C \right\},$$

for some properly chosen $C > 0$.

What do we already know?

Fact. If we adopt the proximal gradient descent, then we can find an ε-approximate solution in $O(1/\varepsilon)$ iterations.

Fact. If we adopt an accelarated proximal gradient method, then we can find an ε-approximate solution in $O(1/\sqrt{\varepsilon})$ iterations.

Remark. The above two arguments *do not apply* to maximum-likelihood quantum state tomography. (Why?)

Question. Why do we seek for another algorithm?

Scalability issue

Observation. Computing projection onto $\mathcal{X}$, with respect to either the 2-norm or a Bregman divergence, requires computing the singular value decomposition of the input matrix first. Then, the per-iteration computational complexity is cubic in dimension.

Observation. The issue lies in scalability with the dimension. Notice that for both quantum state tomography and low-rank matrix completion, typically, the dimension is large.

Example. Suppose that the quantum state we would like to estimate consists of m qubits (quantum bits). Then, $p = 2^m$.

Frank-Wolfe method

Frank-Wolfe method

Consider the optimization problem

$$f^\star = \min_x \{\, f(x) \mid x \in \mathcal{X} \,\},$$

for some convex differentiable function f, and closed convex set $\mathcal{X}$ in a finite-dimensional real vector space E.

Algorithm Frank-Wolfe method (aka conditional gradient method)

1: Set $x_0 \in \mathcal{X}$.
2: **for** $t = 0, 1, \ldots, T$ **do**
3: $\quad v_t \leftarrow \arg\min_v \{\, \langle \nabla f(x_t), v \rangle \mid v \in \mathcal{X} \,\}$
4: $\quad x_{t+1} \leftarrow (1-\tau_t)x_t + \tau_t v_t$, $\tau_t \in [0,1]$
5: **end for**

M. Frank and P. Wolfe. 1956. An algorithm for quadratic programming.
E. S. Levitan and B. T. Polyak. 1966. Constrained minimization methods.

Original form

Algorithm Original Frank-Wolfe method

1: Set $x_0 \in \mathcal{X}$.
2: **for** $t = 0, 1, \ldots, T$ **do**
3: $\quad v_t \leftarrow \arg\min_v \{ \langle \nabla f(x_t), v \rangle \mid v \in \mathcal{X} \}$
4: $\quad \tau_t \leftarrow \arg\min_\tau \{ f((1-\tau)x_t + \tau v_t) \mid \tau \in [0,1] \}$
5: $\quad x_{t+1} \leftarrow (1-\tau_t)x_t + \tau_t v_t$
6: **end for**

Remark. We say that the first step calls a *linear minimization oracle (LMO)*.

Remark. The fourth step is called *exact line search*.

Interpretations (1/2)

First interpretation. The Frank-Wolfe method linearizes the objective function at each iterate, and then solve the corresponding linear minimization problem.

Second interpretation. Let g be the indicator function of $\mathcal{X}$. Then we have $x^\star \in \mathcal{X}$ is a minimizer, if and only if

$$-\nabla f(x^\star) \in \partial g(x^\star).$$

This is equivalently to

$$x^\star \in (\partial g)^{-1}(-\nabla f(x^\star)),$$

also equivalent to, for any $\tau \in]0,1[$,

$$x^\star \in (1-\tau)x^\star + \tau(\partial g)^{-1}(-\nabla f(x^\star)).$$

Interpretations (2/2)

Proposition. We have

$$y \in (\partial g)^{-1}(-\nabla f(x)) \quad \Leftrightarrow \quad y \in \arg\min_z \left\{ \langle \nabla f(x), z \rangle \mid z \in \mathcal{X} \right\}.$$

Proof. The left-hand side holds, if and only if

$$-\nabla f(x) \in \partial g(y),$$

or

$$0 \in \nabla f(x) + \partial g(y).$$

By Fermat's rule, this is equivalent to

$$y \in \arg\min_z \left\{ \langle \nabla f(x), z \rangle + g(z) \mid z \in E \right\}.$$

Y. Yu *et al.* 2017. Generalized conditional gradient for sparse estimation.

Linear minimization oracle (1/3)

Definition. For every $s \in E^*$ (dual space), we define

$$v(s) := \arg\min_x \{ \langle s, x \rangle \mid x \in \mathcal{X} \} .$$

Proposition. Let $E = \mathbb{R}^p$ and $\mathcal{X}$ be the unit 1-norm ball. Then, for every $s \in \mathbb{R}^p$,

$$[v(s)]^{(i)} = \begin{cases} -\operatorname{sign}(s^{(i)}), & \text{if } |s^{(i)}| = \|s\|_\infty, \\ 0, & \text{otherwise.} \end{cases}$$

Proof. Notice that

$$\langle s, x \rangle \geq -\|s\|_\infty \|x\|_1 = -\|s\|_\infty.$$

The lower bound is obviously achievable.

Linear minimization oracle (2/3)

Proposition. Let $E = \mathbb{R}^{p_1 \times p_2}$, and $\mathcal{X}$ be the unit Schatten 1-norm ball. Then, for every $s \in \mathbb{R}^{p_1 \times p_2}$,

$$v(s) = -\operatorname{sign}(\sigma) u_1 u_2^{\mathrm{T}},$$

where σ is the largest singular value of s, and u_1 and u_2 are the corresponding left- and right-singular vectors.

Remark. Here we use the Hilbert-Schmidt inner product:

$$\langle A, B \rangle_{\mathsf{HS}} := \operatorname{tr}(A^{\mathrm{T}} B), \quad \forall A, B \in \mathbb{R}^{p_1 \times p_2}.$$

It is easily checked that $\langle A, B \rangle_{\mathsf{HS}} = \langle \operatorname{vec}(A), \operatorname{vec}(B) \rangle$.

Linear minimization oracle (3/3)

Proposition. Let $E = \mathbb{R}^p$, and $\mathcal{X}$ be the probability simplex. Then, for every $s \in \mathbb{R}^p$,

$$[v(s)]^{(i)} = \delta_{i,i^\star}, \quad i = 1, \ldots, p,$$

where $i^\star$ is the index of the smallest entry of s.

Proposition. Let $E = \mathbb{R}^{p \times p}$, and $\mathcal{X}$ be the set of positive semi-definite matrices of unit trace. Then, for every $s \in \mathbb{R}^{p \times p}$,

$$v(s) = uu^{\mathrm{T}},$$

where u is an eigenvector corresponding to the smallest eigenvalue of s.

Theorem. For any matrix $M \in \mathbb{R}^{p_1 \times p_2}$ and $\varepsilon > 0$, *Lanczos' algorithm* returns a pair of unit vectors (u, v), such that

$$\langle u, Mv \rangle \geq \sigma_{\max}(M) - \varepsilon,$$

with high probability, where $\sigma_{\max}$ denotes the largest singular value of M. The number of required arithmetic operations is $O\left(\mathrm{nnz}(M)\frac{\sqrt{L}\log(p_1+p_2)}{\sqrt{\varepsilon}}\right)$, where L is an upper bound of $\sigma_{\max}(M)$.

Remark. In MATLAB and NumPy, the corresponding functions are eigs and svds.

M. Jaggi. 2013. Revisiting Frank-Wolfe: Projection-free sparse convex optimization.

Convergence

Curvature (1/2)

Definition. Let f be a convex differentiable function and $\mathcal{X}$ be a bounded closed convex set. The *curvature* of f with respect to $\mathcal{X}$ is given by

$$C_f := \max_{x,v,\tau} \frac{2}{\tau^2} \left\{ f((1-\tau)x + \tau v) - [f(x) + \langle \nabla f(x), -\tau x + \tau v \rangle] \right\},$$

subject to the constraint that $x, v \in \mathcal{X}$ and $\tau \in [0, 1]$.

Remark. Therefore, we have

$$f((1-\tau)x + \tau v) \leq f(x) + \langle \nabla f(x), -\tau x + \tau v \rangle + \frac{C_f}{2}\tau^2.$$

K. L. Clarkson. 2010. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm.

Proposition. If f is L-smooth with respect to a norm $\|\cdot\|$ on $\mathcal{X}$, i.e.,

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2, \quad \forall x, y \in \mathcal{X},$$

then,

$$C_f \le L \max_{x,y} \left\{ \|x - y\|^2 \;\middle|\; x, y \in \mathcal{X} \right\}.$$

Proof. By definition, then,

$$C_f \le \max_{x,v,\tau} \left\{ \frac{2}{\tau^2} \frac{L}{2} \| - \tau x + \tau v\|^2 \;\middle|\; x, v \in \mathcal{X}, \tau \in [0,1] \right\}.$$

Convergence guarantee

Theorem. Set $\tau_t = \frac{2}{t+2}$. Then we have

$$f(x_t) - f^\star \leq \frac{2C_f}{t+2}, \quad \forall t \in \mathbb{N}.$$

Remark. Notice that, unlike proximal gradient methods, the choice of τ_t does not require information o C_f, while the convergence speed is comparable to that of the standard gradient descent.

Z. Harchaoui *et al.* 2015. Conditional gradient algorithms for norm-regularized smooth convex optimization.
M. Jaggi. 2013. Revisiting Frank-Wolfe: Projection-free sparse convex optimization.

Proof of convergence (1/2)

Proof. By the definition of the curvature, we write

$$f(x_{t+1}) - f^\star \le f(x_t) - f^\star + \langle \nabla f(x_t), -\tau_t x_t + \tau_t v_t \rangle + \frac{C_f}{2}\tau_t^2.$$

Let $x^\star$ be a minimizer of f on $\mathcal{X}$. By convexity of f, we write

$$\begin{aligned}
\langle \nabla f(x_t), -x_t + v_t \rangle &= -\langle \nabla f(x_t), x_t \rangle + \langle \nabla f(x_t), v_t \rangle \\
&\le \langle \nabla f(x_t), x^* - x_t \rangle \\
&\le -(f(x_t) - f^\star).
\end{aligned}$$

Then, we obtain

$$f(x_{t+1}) - f^\star \le (1 - \tau_t)\,(f(x_t) - f^\star) + \frac{C_f}{2}\tau_t^2.$$

Proof of convergence (2/2)

Proof continued. Define $h_t \coloneqq f(x_t) - f^\star$. We have

$$h_{t+1} \le (1 - \tau_t) h_t + \frac{C_f}{2} \tau_t^2.$$

The theorem follows from the lemma below.

Lemma. Set $\tau_t = \frac{2}{t+2}$. Then $h_t \le \frac{2C_f}{t+2}$.

Proof of the lemma. We prove by induction. When $t = 0$, we have $h_1 \le \frac{C_f}{2} \le \frac{2C_f}{3}$. Assume the induction hypothesis holds for some $t \in \{ 0 \} \cup \mathbb{N}$. Then, we write

$$\begin{aligned} h_{t+1} &\le \left(1 - \frac{2}{t+2} \right) \frac{2C_f}{t+2} + \frac{C_f}{2} \left(\frac{2}{t+2} \right)^2 \\ &= \frac{2C_f(t+1)}{(t+2)^2} = \frac{2C_f(t+1)}{(t+1)(t+3)+1} \le \frac{2C_f}{t+3}. \end{aligned}$$

Extensions

There are more general convergence results.

- Frank-Wolfe-type methods for:
  - Minimizing $\|\cdot\|$ subject to $f(\cdot) \leq 0$.
  - Minimizing $f(\cdot) + \lambda\|\cdot\|$.
- Numerical error bound for an arbitrary sequence of τ_t.
- Convergence under a Hölder condition: There exist some $\nu \in \,]0, 1]$ and $G_\nu > 0$, such that

$$\|\nabla f(y) - \nabla f(x)\|_* \leq G_\nu \|x - y\|^\nu, \quad \forall x, y \in \mathcal{X}.$$

Z. Harchaoui *et al.* 2015. Conditional gradient algorithms for norm-regularized smooth convex optimization.

R. Freund and P. Grigas. 2016. New analysis and results for the Frank-Wolfe method.

Yu. Nesterov. 2018. Complexity bounds for primal-dual methods minimizing the model of objective function.

Conclusions

Summary

Algorithm Frank-Wolfe method (aka conditional gradient method)

1: Set $x_0 \in \mathcal{X}$.
2: **for** $t = 0, 1, \ldots, T$ **do**
3: $\quad v_t \leftarrow \arg\min_v \{ \langle \nabla f(x_t), v \rangle \mid v \in \mathcal{X} \}$
4: $\quad x_{t+1} \leftarrow (1 - \tau_t) x_t + \tau_t v_t$, $\tau_t \in [0, 1]$
5: **end for**

Remark. This method is in particular computationally efficient when $\mathcal{X}$ is a Schatten 1-norm ball.

Theorem. If $\tau_t = \frac{2}{t+2}$, then $f(x_t) - f^\star = O(C_f / t)$.

Next lecture

- Online learning.
- Follow the leader, follow the regularized leader, follow the perturbed leader.