Convex Optimization

Part 3: Frank-Wolfe method

Namhoon Lee

POSTECH

19 Oct 2022

Projected gradient method

Consider constrained minimization problem

$$\min_{x} \; f(x)$$
$$\text{s.t. } x \in \mathbb{C}$$

where f is convex and smooth, and $\mathbb{C}$ is convex.

Projected gradient method repeats the following update

$$x_{t+1} = \mathrm{P}_{\mathbb{C}}(x_t - \eta \nabla f(x_t))$$

where $\mathrm{P}_{\mathbb{C}}$ is the projection operator onto the set $\mathbb{C}$.

We can treat projection as a special case of proximal operation. However, the projection step may not always be easy.

- local quadratic expansion of f

Frank-Wolfe method

Frank-Wolfe method (conditional gradient method) uses a local linear expansion of f

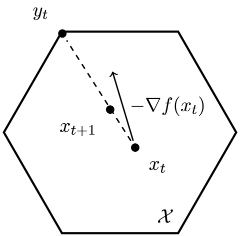
$$y_t \in \arg\min_{y \in \mathbb{C}} \nabla f(x_t)^\top y$$
$$x_{t+1} = (1 - \gamma_t) x_t + \gamma_t y_t$$

where default step size is $\gamma_t = 2/(t+1)$ for $t = 1, 2, \ldots$.

- Unlike projected gradient method, there is no projection; instead Frank-Wolfe minimizes a linear function.
- When the set constraint is easy, then Frank-Wolfe can be more efficient than projected gradient method; for instance, $\mathbb{C}$ is convex polytope, the minimizer is always found in one of the vertices.
- The update is always in the feasible set; for $0 \leq \gamma_t \leq 1$ we have $x_t \in \mathbb{C}$ by convexity.

$$y_t \in \arg\min_{y \in \mathbb{C}} \nabla f(x_t)^\top y$$

$$x_{t+1} = (1 - \gamma_t) x_t + \gamma_t y_t$$



- moving less and less in the direction of the linearization minimizer as the algorithm proceeds

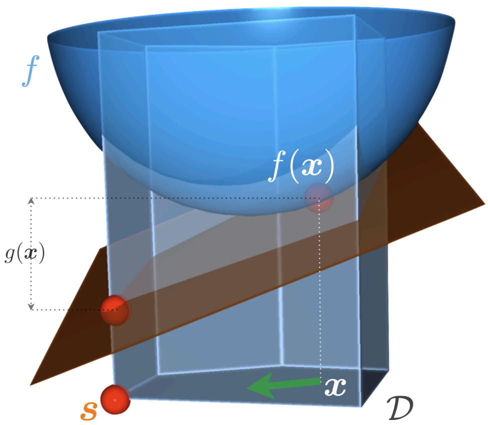


Figure: The algorithm considers the linearization of the objective function and moves towards its minimizer; figure from (Jaggi 2013)

Norm constraint

Consider $\mathbb{C} = \{x : \|x\| \leq t\}$ for an abitrary norm $\|\cdot\|$. Then

$$
\begin{aligned}
y_t &\in \underset{\|y\| \leq t}{\arg\min} \; \nabla f(x_t)^\top y \\
&= -t \cdot \left(\underset{\|y\| \leq 1}{\arg\max} \; \nabla f(x_t)^\top y \right) \\
&= -t \cdot \partial \|\nabla f(x_t)\|_*
\end{aligned}
$$

where $\|\cdot\|_*$ denotes the corresponding dual norm.

- If we know how to compute subgradients of the dual norm, then we can easily perform Frank-Wolfe steps.
- A key to Frank-Wolfe: this can often be simpler or cheaper than projection onto $\mathbb{C} = \{x : \|x\| \leq t\}$

Example

Consider minimizing with 1-norm constraint

$$\min_x \; f(x)$$
$$\text{s. t. } \|x\|_1 \le t$$

We have $y_t = -t\partial\|\nabla f(x_t)\|_\infty$, and thus Frank-Wolf update becomes

$$i_t \in \arg\max_{i=1,\ldots,d} |\nabla_i f(x_t))|$$
$$x_{t+1} = (1-\gamma_t)x_t - \gamma_t t \cdot \text{sign}\left(\nabla_{i_t} f(x_t)\right) \cdot e_{i_t}$$

- Special case of coordinate descent (update one coordinate at a time)
- Simpler than projection onto 1-norm ball

Convergence analysis

Theorem

Let f be a convex and β-smooth function with respect to some norm $\|\cdot\|$, $R = \sup_{x,y\in\mathbb{C}} \|x - y\|$, and $\gamma_t = 2/(t+1)$ for $t \geq 1$. Then for any $t \geq 2$, one has

$$f(x_T) - f(x^*) \leq \frac{2\beta R^2}{T+1}$$

- same convergence rate as gradient descent for smooth function
- smoothness measured in arbitrary norm $\|\cdot\|$ ("norm-free")

Proof.

Using β-smoothness (for arbitrary norms), the definition of the algorithm, and the convexity of f

$$
\begin{aligned}
f(x_{t+1}) - f(x_t) &\leq \nabla f(x_t)^\top (x_{t+1} - x_t) + \frac{\beta}{2}\|x_{t+1} - x_t\|^2 \\
&\leq \gamma_t \nabla f(x_t)^\top (y_t - x_t) + \frac{\beta}{2}\gamma_t^2 R^2 \\
&\leq \gamma_t \nabla f(x_t)^\top (x^* - x_t) + \frac{\beta}{2}\gamma_t^2 R^2 \\
&\leq \gamma_t (f(x^*) - f(x_t)) + \frac{\beta}{2}\gamma_t^2 R^2
\end{aligned}
$$

Rewriting this inequality in terms of $\delta_t = f(x_t) - f(x^*)$ one obtains

$$
\delta_{t+1} \leq (1 - \gamma_t)\delta_t + \frac{\beta}{2}\gamma_t^2 R^2
$$

We prove $\delta_T \leq \frac{2\beta R^2}{T+1}$ by induction. First we show that the base case ($T = 2$) holds true, *i.e.*, $\delta_2 \leq \frac{2}{3}\beta R^2$. With $t = 1$, $\gamma_t = 1$, and we get from the previous inquality that

$$f(x_2) - f(x_1) \leq f(x^*) - f(x_1) + \frac{\beta}{2}R^2$$
$$\iff \delta_2 = f(x_2) - f(x^*) \leq \frac{\beta}{2}R^2 \leq \frac{2}{3}\beta R^2$$

Next assume $\delta_T \leq \frac{2\beta R^2}{T+1}$ for $T = t$, and show it holds true for $T = t + 1$

$$\begin{aligned}
\delta_{t+1} &\leq (1 - \gamma_t)\delta_t + \frac{\beta}{2}\gamma_t^2 R^2 \\
&\leq \left(1 - \frac{2}{t+1}\right)\frac{2\beta R^2}{t+1} + \frac{\beta}{2}\left(\frac{2}{t+1}\right)^2 R^2 \\
&= \frac{2t\beta R^2}{(t+1)^2}
\end{aligned}$$

We finish the proof by noting that $t/(t+1) \leq 1$. □

Any questions?

References I

Jaggi, Martin (2013). "Revisiting Frank-Wolfe: Projection-free sparse convex optimization". In: *ICML*.