Frank-Wolfe Method

Ryan Tibshirani
Convex Optimization 10-725

Last time: ADMM

For the problem

$$\min_{x,z} \; f(x) + g(z) \quad \text{subject to} \quad Ax + Bz = c$$

we form augmented Lagrangian (scaled form):

$$L_\rho(x, z, w) = f(x) + g(z) + \frac{\rho}{2}\|Ax - Bx + c + w\|_2^2 - \frac{\rho}{2}\|w\|_2^2$$

Alternating direction method of multipliers or ADMM:

$$\begin{aligned} x^{(k)} &= \underset{x}{\operatorname{argmin}} \; L_\rho(x, z^{(k-1)}, w^{(k-1)}) \\ z^{(k)} &= \underset{z}{\operatorname{argmin}} \; L_\rho(x^{(k)}, z, w^{(k-1)}) \\ w^{(k)} &= w^{(k-1)} + Ax^{(k)} + Bz^{(k)} - c \end{aligned}$$

Converges like a first-order method. Very flexible framework

Projected gradient descent

Consider constrained problem

$$\min_x \ f(x) \ \text{ subject to } \ x \in C$$

where f is convex and smooth, and C is convex. Recall projected gradient descent chooses an initial $x^{(0)}$, repeats for $k = 1, 2, 3, \ldots$

$$x^{(k)} = P_C\big(x^{(k-1)} - t_k \nabla f(x^{(k-1)}\big)$$

where P_C is the projection operator onto the set C. Special case of proximal gradient, motivated by local quadratic expansion of f:

$$x^{(k)} = P_C\left(\underset{y}{\operatorname{argmin}} \ \nabla f(x^{(k-1)})^T (y - x^{(k-1)}) + \frac{1}{2t}\|y - x^{(k-1)}\|_2^2\right)$$

Motivation for today: projections are not always easy!

Frank-Wolfe method

The Frank-Wolfe method, also called conditional gradient method, uses a local linear expansion of f:

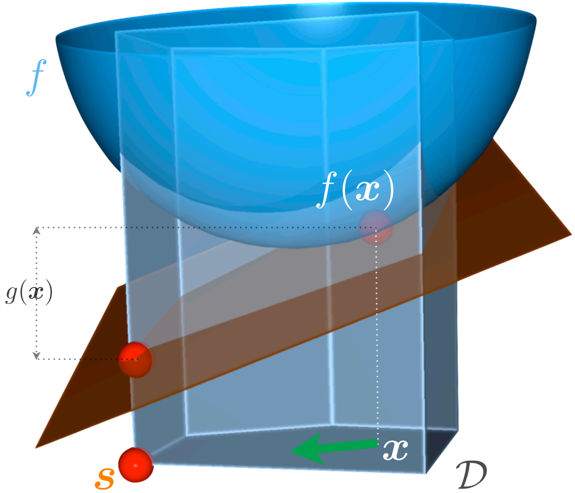
$$s^{(k-1)} \in \underset{s \in C}{\operatorname{argmin}} \ \nabla f(x^{(k-1)})^T s$$
$$x^{(k)} = (1-\gamma_k) x^{(k-1)} + \gamma_k s^{(k-1)}$$

Note that there is no projection; update is solved directly over C

Default step sizes: $\gamma_k = 2/(k+1)$, $k = 1, 2, 3, \ldots$. Note for any $0 \le \gamma_k \le 1$, we have $x^{(k)} \in C$ by convexity. Can rewrite update as

$$x^{(k)} = x^{(k-1)} + \gamma_k (s^{(k-1)} - x^{(k-1)})$$

i.e., we are moving less and less in the direction of the linearization minimizer as the algorithm proceeds



(From Jaggi 2011)

Norm constraints

What happens when $C = \{x : \|x\| \leq t\}$ for a norm $\|\cdot\|$? Then

$$
\begin{aligned}
s &\in \underset{\|s\|\leq t}{\operatorname{argmin}} \; \nabla f(x^{(k-1)})^T s \\
&= -t \cdot \left(\underset{\|s\|\leq 1}{\operatorname{argmax}} \; \nabla f(x^{(k-1)})^T s \right) \\
&= -t \cdot \partial \|\nabla f(x^{(k-1)})\|_*
\end{aligned}
$$

where $\|\cdot\|_*$ denotes the corresponding dual norm. That is, 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 $C = \{x : \|x\| \leq t\}$

Outline

Today:

- Examples
- Convergence analysis
- Properties and variants
- Path following

Example: ℓ_1 regularization

For the ℓ_1-regularized problem

$$\min_x \; f(x) \;\; \text{subject to} \;\; \|x\|_1 \le t$$

we have $s^{(k-1)} \in -t\partial\|\nabla f(x^{(k-1)})\|_\infty$. Frank-Wolfe update is thus

$$\begin{aligned} i_{k-1} &\in \operatorname*{argmax}_{i=1,\ldots,p} \; \big|\nabla_i f(x^{(k-1)})\big| \\ x^{(k)} &= (1-\gamma_k)x^{(k-1)} - \gamma_k t \cdot \text{sign}\big(\nabla_{i_{k-1}} f(x^{(k-1)})\big) \cdot e_{i_{k-1}} \end{aligned}$$

Like greedy coordinate descent! (But with diminshing steps)

Note: this is a lot simpler than projection onto the ℓ_1 ball, though both require $O(n)$ operations

Example: ℓ_p regularization

For the ℓ_p-regularized problem

$$\min_x \; f(x) \;\; \text{subject to} \;\; \|x\|_p \leq t$$

for $1 \leq p \leq \infty$, we have $s^{(k-1)} \in -t\partial \|\nabla f(x^{(k-1)})\|_q$, where p, q are dual, i.e., $1/p + 1/q = 1$. Claim: can choose

$$s_i^{(k-1)} = -\alpha \cdot \text{sign}\big(\nabla f_i(x^{(k-1)})\big) \cdot \big|\nabla f_i(x^{(k-1)})\big|^{p/q}, \quad i = 1, \ldots, n$$

where α is a constant such that $\|s^{(k-1)}\|_q = t$ (check this!), and then Frank-Wolfe updates are as usual

Note: this is a lot simpler projection onto the ℓ_p ball, for general p! Aside from special cases ($p = 1, 2, \infty$), these projections cannot be directly computed (must be treated as an optimization)

Example: trace norm regularization

For the trace-regularized problem

$$\min_X \; f(X) \;\; \text{subject to} \;\; \|X\|_{\text{tr}} \le t$$

we have $S^{(k-1)} \in -t\partial\|\nabla f(X^{(k-1)})\|_{\text{op}}$. Claim: can choose

$$S^{(k-1)} = -t \cdot uv^T$$

where u, v are leading left and right singular vectors of $\nabla f(X^{(k-1)})$ (check this!), and then Frank-Wolfe updates are as usual

Note: this substantially simpler and cheaper than projection onto the trace norm ball, which requires a singular value decomposition!

Constrained and Lagrange forms

Recall that solution of the constrained problem

$$\min_x \; f(x) \;\; \text{subject to} \;\; \|x\| \le t$$

are equivalent to those of the Lagrange problem

$$\min_x \; f(x) + \lambda \|x\|$$

as we let the tuning parameters t and λ vary over $[0, \infty]$. Typically in statistics and ML problems, we would just solve whichever form is easiest, over wide range of parameter values, then use CV

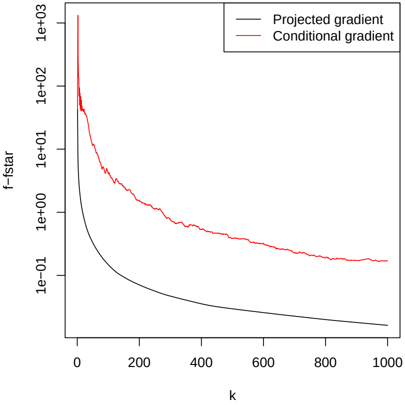
So we should also compare the Frank-Wolfe updates under $\| \cdot \|$ to the proximal operator of $\| \cdot \|$

- ℓ_1 norm: Frank-Wolfe update scans for maximum of gradient; proximal operator soft-thresholds the gradient step; both use $O(n)$ flops
- ℓ_p norm: Frank-Wolfe update computes raises each entry of gradient to power and sums, in $O(n)$ flops; proximal operator not generally directly computable
- Trace norm: Frank-Wolfe update computes top left and right singular vectors of gradient; proximal operator soft-thresholds the gradient step, requiring a singular value decomposition

Various other constraints yield efficient Frank-Wolfe updates, e.g., special polyhedra or cone constraints, sum-of-norms (group-based) regularization, atomic norms. See Jaggi (2011)

Example: lasso comparison

Comparing projected and conditional gradient for constrained lasso problem, with $n = 100$, $p = 500$:



Note: FW uses standard step sizes, line search would probably help

Duality gap

Frank-Wolfe iterations admit a very natural duality gap:

$$g(x^{(k)}) = \nabla f(x^{(k)})^T (x^{(k)} - s^{(k)})$$

Claim: it holds that $f(x^{(k)}) - f^\star \le g(x^{(k)})$

Proof: by the first-order condition for convexity

$$f(s) \ge f(x^{(k)}) + \nabla f(x^{(k)})^T (s - x^{(k)})$$

Minimizing both sides over all $s \in C$ yields

$$\begin{aligned} f^\star &\ge f(x^{(k)}) + \min_{s \in C} \; \nabla f(x^{(k)})^T (s - x^{(k)}) \\ &= f(x^{(k)}) + \nabla f(x^{(k)})^T (s^{(k)} - x^{(k)}) \end{aligned}$$

Rearranged, this gives the duality gap above

Why do we call it "duality gap"? Rewrite original problem as

$$\min_x \; f(x) + I_C(x)$$

where I_C is the indicator function of C. The dual problem is

$$\max_u \; -f^*(u) - I_C^*(-u)$$

where I_C^* is the support function of C. Duality gap at x, u is

$$f(x) + f^*(u) + I_C^*(-u) \geq x^T u + I_C^*(-u)$$

Evaluated at $x = x^{(k)}$, $u = \nabla f(x^{(k)})$, this gives

$$\nabla f(x^{(k)})^T x^{(k)} + \max_{s \in C} \; -\nabla f(x^{(k)})^T s = \nabla f(x^{(k)})^T (x^{(k)} - s^{(k)})$$

which is our gap

Convergence analysis

Following Jaggi (2011), define the curvature constant of f over C:

$$M = \max_{\substack{\gamma\in[0,1] \\ x,s,y\in C \\ y=(1-\gamma)x+\gamma s}} \frac{2}{\gamma^2}\Big(f(y) - f(x) - \nabla f(x)^T(y-x)\Big)$$

Note that $M = 0$ for linear f, and $f(y) - f(x) - \nabla f(x)^T(y-x)$ is called the Bregman divergence, defined by f

Theorem: The Frank-Wolfe method using standard step sizes $\gamma_k = 2/(k+1)$, $k = 1, 2, 3, \ldots$ satisfies

$$f(x^{(k)}) - f^\star \le \frac{2M}{k+2}$$

Thus number of iterations needed for $f(x^{(k)}) - f^\star \le \epsilon$ is $O(1/\epsilon)$

This matches the sublinear rate for projected gradient descent for Lipschitz ∇f, but how do the assumptions compare?

For Lipschitz ∇f with constant L, recall

$$f(y) - f(x) - \nabla f(x)^T (y - x) \leq \frac{L}{2} \|y - x\|_2^2$$

Maximizing over all $y = (1-\gamma)x + \gamma s$, and multiplying by $2/\gamma^2$,

$$\begin{aligned} M &\leq \max_{\substack{\gamma \in [0,1] \\ x,s,y \in C \\ y = (1-\gamma)x + \gamma s}} \frac{2}{\gamma^2} \cdot \frac{L}{2} \|y - x\|_2^2 \\ &= \max_{x,s \in C} L \|x - s\|_2^2 = L \cdot \mathrm{diam}^2(C) \end{aligned}$$

Hence assuming a bounded curvature is basically no stronger than what we assumed for projected gradient

Basic inequality

The key inequality used to prove the Frank-Wolfe convergence rate:

$$f(x^{(k)}) \leq f(x^{(k-1)}) - \gamma_k g(x^{(k-1)}) + \frac{\gamma_k^2}{2} M$$

Here $g(x) = \max_{s \in C} \nabla f(x)^T (x - s)$ is duality gap defined earlier

Proof: write $x^+ = x^{(k)}$, $x = x^{(k-1)}$, $s = s^{(k-1)}$, $\gamma = \gamma_k$. Then

$$\begin{aligned} f(x^+) &= f\big(x + \gamma(s - x)\big) \\ &\leq f(x) + \gamma \nabla f(x)^T (s - x) + \frac{\gamma^2}{2} M \\ &= f(x) - \gamma g(x) + \frac{\gamma^2}{2} M \end{aligned}$$

Second line used definition of M, and third line the definition of g

The proof of the convergence result is now straightforward. Denote by $h(x) = f(x) - f^\star$ the suboptimality gap at x. Basic inequality:

$$\begin{aligned} h(x^{(k)}) &\leq h(x^{(k-1)}) - \gamma_k g(x^{(k-1)}) + \frac{\gamma_k^2}{2} M \\ &\leq h(x^{(k-1)}) - \gamma_k h(x^{(k-1)}) + \frac{\gamma_k^2}{2} M \\ &= (1 - \gamma_k) h(x^{(k-1)}) + \frac{\gamma_k^2}{2} M \end{aligned}$$

where in the second line we used $g(x^{(k-1)}) \geq h(x^{(k-1)})$

To get the desired result we use induction:

$$h(x^{(k)}) \leq \left(1 - \frac{2}{k+1}\right) \frac{2M}{k+1} + \left(\frac{2}{k+1}\right)^2 \frac{M}{2} \leq \frac{2M}{k+2}$$

Affine invariance

Frank-Wolfe updates are affine invariant: for nonsingular matrix A, define $x = Ax'$, $F(x') = f(Ax')$, consider Frank-Wolfe on F:

$$s' = \underset{z \in A^{-1}C}{\operatorname{argmin}} \; \nabla F(x')^T z$$
$$(x')^+ = (1-\gamma)x' + \gamma s'$$

Multiplying by A produces same Frank-Wolfe update as that from f. Convergence analysis is also affine invariant: curvature constant

$$M = \max_{\substack{\gamma \in [0,1] \\ x',s',y' \in A^{-1}C \\ y' = (1-\gamma)x' + \gamma s'}} \frac{2}{\gamma^2}\Big(F(y') - F(x') - \nabla F(x')^T(y'-x')\Big)$$

matches that of f, because $\nabla F(x')^T(y'-x') = \nabla f(x)^T(y-x)$

Inexact updates

Jaggi (2011) also analyzes inexact Frank-Wolfe updates: suppose we choose $s^{(k-1)}$ so that

$$\nabla f(x^{(k-1)})^T s^{(k-1)} \le \min_{s \in C} \nabla f(x^{(k-1)})^T s + \frac{M\gamma_k}{2} \cdot \delta$$

where $\delta \ge 0$ is our inaccuracy parameter. Then we basically attain the same rate

Theorem: Frank-Wolfe using step sizes $\gamma_k = 2/(k+1)$, $k = 1, 2, 3, \ldots$, and inaccuracy parameter $\delta \ge 0$, satisfies

$$f(x^{(k)}) - f^\star \le \frac{2M}{k+1}(1+\delta)$$

Note: the optimization error at step k is $M\gamma_k/2 \cdot \delta$. Since $\gamma_k \to 0$, we require the errors to vanish

Two variants

Two important variants of Frank-Wolfe:

- Line search: instead of using standard step sizes, use

$$\gamma_k = \underset{\gamma \in [0,1]}{\operatorname{argmin}} \; f\big(x^{(k-1)} + \gamma(s^{(k-1)} - x^{(k-1)})\big)$$

  at each $k = 1, 2, 3, \ldots$. Or, we could use backtracking

- Fully corrective: directly update according to

$$x^{(k)} = \underset{y}{\operatorname{argmin}} \; f(y) \;\; \text{subject to} \;\; y \in \operatorname{conv}\{x^{(0)}, s^{(0)}, \ldots, s^{(k-1)}\}$$

Both variants lead to the same $O(1/\epsilon)$ iteration complexity

Another popular variant: away steps, which get linear convergence under strong convexity

Path following

Given the norm constrained problem

$$\min_x \; f(x) \;\; \text{subject to} \;\; \|x\| \le t$$

Frank-Wolfe can be used for path following, i.e., we can produce an approximate solution path $\hat{x}(t)$ that is ϵ-suboptimal for every $t \ge 0$. Let $t_0 = 0$ and $x^\star(0) = 0$, fix $m > 0$, repeat for $k = 1, 2, 3, \ldots$:

- Calculate
$$t_k = t_{k-1} + \frac{(1 - 1/m)\epsilon}{\|\nabla f(\hat{x}(t_{k-1}))\|_*}$$
and set $\hat{x}(t) = \hat{x}(t_{k-1})$ for all $t \in (t_{k-1}, t_k)$
- Compute $\hat{x}(t_k)$ by running Frank-Wolfe at $t = t_k$, terminating when the duality gap is $\le \epsilon/m$

(This is a simplification of the strategy from Giesen et al., 2012)

Claim: this produces (piecewise-constant) path with

$$f(\hat{x}(t)) - f(x^\star(t)) \le \epsilon \quad \text{for all } t \ge 0$$

Proof: rewrite the Frank-Wolfe duality gap as

$$g_t(x) = \max_{\|s\| \le t} \; \nabla f(x)^T (x - s) = \nabla f(x)^T x + t\|\nabla f(x)\|_*$$

This is a linear function of t. Hence if $g_t(x) \le \epsilon/m$, then we can increase t until $t^+ = t + (1 - 1/m)\epsilon/\|\nabla f(x)\|_*$, because

$$g_{t^+}(x) = \nabla f(x)^T x + t\|\nabla f(x)\|_* + \epsilon - \epsilon/m \le \epsilon$$

i.e., the duality gap remains $\le \epsilon$ for the same x, between t and t^+

References

- K. Clarkson (2010), "Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm"
- J. Giesen and M. Jaggi and S. Laue, S. (2012), "Approximating parametrized convex optimization problems"
- M. Jaggi (2011), "Sparse convex optimization methods for machine learning"
- M. Jaggi (2011), "Revisiting Frank-Wolfe: projection-free sparse convex optimization"
- M. Frank and P. Wolfe (2011), "An algorithm for quadratic programming"
- R. J. Tibshirani (2015), "A general framework for fast stagewise algorithms"