An Introduction to Conditional Gradient

Yaoliangorov Yu
University of Alberta

Tea-Time-Talk, August 8th, 2012

A bit of motivation

Linear programming (Kantorovich, 1939)

$$\min_x \langle c, x \rangle, \text{s.t. } x \geq 0, Ax = b$$

The simplex algorithm (Dantzig, 1947)

You know what it is. Famous story happened in 1939.

What next?

Quadratic programming:

$$\min_x \left\langle x, \tfrac{1}{2}Px + c \right\rangle, \text{s.t. } x \geq 0, Ax = b$$

How would you solve it?

A bit of history

"(Bob) Dorfman (at the time with UC Berkeley) used the then very new Kuhn-Tucker theory ... if you wrote the Kuhn-Tucker conditions ... you had a big set of linear equations that looked like a simplex method tableau ... you could show that the solution of the quadratic problem was an extreme point of this tableau problem."

We submitted to the Naval Research Logistics Quarterly. Alan Hoffman was an editor ... he received a manuscript from Harry Markowitz on portfolio selection by parametric quadratic minimization .... he sent the Markowitz manuscript to me and our manuscript to Markowitz to referee. Talking to Harry much later, I found that we did the same sort of thing: looked at the other's paper, couldn't understand it very well, but it sounded like competent mathematics. So we turned in our reports saying 'publish this.' " — Philip Wolfe

Conditional gradient (Frank-Wolfe, 1956)

Convex program

$$\min_{x \in Q} f(x),$$

where Q is convex (bounded), f is convex.

Smoothness assumption

$\exists C_f \geq 0, \forall x, y \in Q, \forall \eta \in (0,1)$

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

Note that: $C_f \leq L \cdot \|y - x\|^2$, if $f \in \mathcal{C}_L^1$.

The power of linearization

Step 1 is a linear program if Q is polyhedra.

1. $y_k \in \arg\min_{y \in Q} \langle y, \nabla f(x_k) \rangle$;
2. choose $\eta_k \in [0, 1]$;
3. $x_{k+1} \leftarrow (1 - \eta_k)x_k + \eta_k y_k$.

Convergence rate

$$
\begin{aligned}
f(x_{k+1}) - f(x^*) &= f((1-\eta_k)x_k + \eta_k y_k) - f(x^*) \\
&\le f(x_k) - f(x^*) + \eta_k \langle y_k - x_k, \nabla f(x_k) \rangle + \tfrac{C_f}{2}\eta_k^2 \\
&\le f(x_k) - f(x^*) + \eta_k \langle x^* - x_k, \nabla f(x_k) \rangle + \tfrac{C_f}{2}\eta_k^2 \\
&\le (1-\eta_k)(f(x_k) - f(x^*)) + \tfrac{C_f}{2}\eta_k^2.
\end{aligned}
$$

Choose $\eta_k = \frac{2}{k+2}$ and use induction proves

$$f(x_k) - f(x^*) \le \frac{C_f}{k+2}.$$

Note: (Dem'yanov and Rubinov, 1963) independently discovered the conditional gradient (although without rates).

What about f is additionally strongly convex?

Will we get linear rate, after all that is what we get for gradient methods?

A bit of digestion

Adavantage

Simple; ideal for infinite-dim.

Disadvantage

Not very fast; hard to apply in nonsmooth/stochastic settings.

A bit of digestion

Adavantage

Simple; ideal for infinite-dim.

Disadvantage

Not very fast; hard to apply in nonsmooth/stochastic settings.

Exact $1/k$ rate (Polyak, 1987)

$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \le 1, b \ge 0$$

Step 1 yields $y_k = (-\mathrm{sign}(a_k), 0)$;
Step 2 chooses η optimally: $\eta_k = \frac{|a_k|+b_k+a_k^2+b_k^2}{(|a_k|+1)^2+b_k^2} \wedge 1$;

For k large, $\eta_k < 1$ and $f_{k+1} = f_k - \frac{1}{4}\|x_k - y_k\|^2(f_k - f^*)^2$. Since $\|x_k - y_k\|^2 \to 1$ we obtain $f_k - f^* = 4/k + o(1/k)$.

Slow rate (Canon and Cullum, 1968)

Consider the QP

$$\min_x \left\langle x, \tfrac{1}{2}Px + c \right\rangle, \text{ s.t. } x \in Q := \mathrm{conv}\{z_1, \ldots, z_m\}, \quad \text{where} \quad P \succ 0$$

Theorem (For the usual conditional gradient)

Suppose $x^* \in \partial Q - \{z_1, \ldots, z_m\}$ and $x_k \in \mathrm{ri}\, Q$ indefinitely, then $\forall \alpha > 0, \epsilon > 0$, $f(x_k) - f(x^*) \geq \alpha / k^{1+\epsilon}$ indefinitely.

In particular, if $x_1 \in \mathrm{ri}\, Q$ and $f(x_1) < \min_i f(z_i)$, then $x_k \in \mathrm{ri}\, Q$.

In retrospect

(Frank-Wolfe, 1956) deliberately (or accidentally?) went through messy primal-dual transforms so that x^* becomes an extreme point, hence avoiding the above theorem (in fact, proved termination in finite steps).

Fast rate I (Levtin and Polyak, 1966)

Uniformly convex set

A convex set C is (mid-point) δ-uniformly convex iff $\forall x, y \in C$, $\forall \|z\| \leq \delta(\|x - y\|)$, $\frac{x+y}{2} + z \in C$; μ-strongly convex iff $\delta(t) = \frac{\mu}{2}t^2$.

Theorem

If in addition, $\inf_{x \in Q} \|\nabla f(x)\| > 0$, *$Q$ is strongly convex,* $\eta_k = \frac{\langle \nabla f(x_k), x_k - y_k \rangle}{L\|x_k - y_k\|^2} \wedge 1$, *then $x_k \to x^*$ at a linear rate.*

Theorem (Sharp minima (Polyak, 1987))

If in addition, $\forall x \in Q, f(x) \geq f(x^*) + \alpha\|x - x^*\|$, then $x_k \to x^*$ in finite steps.

Page 26, Notes 1: "The asymptotic $O(1/k)$ rate cannot be improved without extra assumption ... even if f is strongly convex."

Fast rate II (Beck and Teboulle, 2004)

Problem

Find a point in $\{x : Ax = b\} \cap Q$, where Q is (weakly) compact. Assume feasibility, equivalent as $\min_{x \in Q} \|Ax - b\|^2$.

Theorem

Assume feasibility and Slater's condition, then the conditional gradient (applied to the above problem) converges at a linear rate.

Kernel herding (Chen, Welling and Smola, 2010)

Subsampling: $\min_{\phi \in Q} \|\mu_{\mathbb{P}} - \phi\|_{\mathcal{H}}^2, \mathrm{where} Q := \overline{\mathrm{conv}}\{\Phi(X)\}, \sup_{x \in X} \|\Phi(x)\|_{\mathcal{H}} < \infty.$

$1/k^2$ rate, assuming $\mu_{\mathbb{P}} \in \mathrm{ri}\, Q$. Note that *i.i.d.* sampling yields $1/k$ rate.

Fast rate II (Beck and Teboulle, 2004)

Problem

Find a point in $\{x : Ax = b\} \cap Q, \quad \text{where} \quad Q$ is (weakly) compact. Assume feasibility, equivalent as $\min_{x\in Q} \|Ax - b\|^2$.

Theorem

Assume feasibility and Slater's condition, then the conditional gradient (applied to the above problem) converges at a linear rate.

Kernel herding (Chen, Welling and Smola, 2010)

Subsampling: $\min_{\phi\in Q} \|\mu_{\mathbb{P}} - \phi\|_{\mathcal{H}}^2, \mathrm{where} Q := \overline{\mathrm{conv}}\{\Phi(X)\}, \sup_{x\in X} \|\Phi(x)\|_{\mathcal{H}} < \infty.$

$1/k^2$ rate, assuming $\mu_{\mathbb{P}} \in \mathrm{ri}\, Q$. Note that *i.i.d.* sampling yields $1/k$ rate.

But see (Bach, Lacoste-Julien and Obozinski, 2012)!

Fast rate II (Beck and Teboulle, 2004)

Problem

Find a point in $\{x : Ax = b\} \cap Q, \quad$ where $\quad Q$ is (weakly) compact. Assume feasibility, equivalent as $\min_{x \in Q} \|Ax - b\|^2$.

Theorem

Assume feasibility and Slater's condition, then the conditional gradient (applied to the above problem) converges at a linear rate.

Kernel herding (Chen, Welling and Smola, 2010)

Subsampling: $\min_{\phi \in Q} \|\mu_{\mathbb{P}} - \phi\|_{\mathcal{H}}^2, \mathrm{where} Q := \overline{\mathrm{conv}}\{\Phi(X)\}, \sup_{x \in X} \|\Phi(x)\|_{\mathcal{H}} < \infty.$

$1/k^2$ rate, assuming $\mu_{\mathbb{P}} \in \mathrm{ri}\, Q$. Note that *i.i.d.* sampling yields $1/k$ rate.

But see (Bach, Lacoste-Julien and Obozinski, 2012)!

Similar idea appeared in (White, Yu, Zhang and Schuurmans, 2012).

Connection to boosting (Zhang, 2003)

Obvious: Q corresponds to the set of hypotheses, and Step 1 corresponds to the oracle that selects "weak" hypotheses.

Totally corrective (Meyer, 1974)

Had we kept all y_k's, perhaps we should choose

$$x_{k+1} \leftarrow \arg \min_{x \in \mathrm{conv}\{y_1, \ldots, y_k\}} f(x).$$

Progressively more expensive. At least converge at $O(1/k)$.

Connection to function approximation (Temlyakov, 2012)

Problem

Let $(X, \|\cdot\|)$ be a Banach space, D an arbitrary bounded subset having dense span. Given $z \in X$, want to approximate it by linear combinations of elements of D, *i.e.* $\min_{x \in \mathrm{span}\{D\}} \|z - x\|$.

Similar to the Herding problem, except the norm need not be Hilbertien.

The algorithm

1. $x_k^* \in \mathrm{argmax}_{\|x^*\|=1} \langle z - x_k; x^* \rangle$;
2. $y_k \in \mathrm{argmin}_{y \in D} \langle y; x_k^* \rangle$;
3. choose η_k;
4. $x_{k+1} \leftarrow (1 - \eta_k) x_k + \eta_k y_k$.

Theorem

Suppose $\|\cdot\|$ is uniformly smooth with type $q \in (1, 2]$, then the above algorithm converges at $k^{-1/p}$.

Fast rate III (Shalev-Shwartz, Srebro and Zhang, 2010)

Problem

$$\min_{x:\|x\|_1 \leq B} f(x), \quad \text{where} \quad f \in C_L^1.$$

Uniformly convex function

f is δ-uniformly convex iff $\forall x, y \in \operatorname{dom} f, \forall \lambda \in (0,1)$,
$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \lambda(1-\lambda)\delta(\|x-y\|)$;
μ-strongly convex iff $\delta(t) = \frac{\mu}{2}t^2$.

Theorem

If in addition f is μ-strongly convex, then the totally corrective conditional gradient algorithm converges at a linear rate.

Remains true for ℓ_1 regularization (Zhang, Yu and Schuurmans, 2012).

Generalized conditional gradient (Bredies, Lorenz and Maass, 2009)

Problem

$$\min_{x \in Q} g(x), \quad \text{where} \quad g(x) := f(x) + h(x).$$

As usual, $f \in \mathcal{C}_L^1$ is convex while h is convex but not necessarily smooth.

Lesson learned recently

Carry the nonsmooth part with you!

The generalized algorithm

1. $y_k \in \arg\min_{x \in Q} \langle x, \nabla f(x_k) \rangle + h(x)$;
2. choose η_k;
3. line search if Q is unbounded;
4. $x_{k+1} \leftarrow (1 - \eta_k) x_k + \eta_k y_k$;

Mirror descent as g.c.g.

Mirror descent (Nemirovski and Yudin, 1979)

The problem: $\min_{y \in Q} \; g(y)$;

The algorithm: $y_{k+1} \leftarrow \operatorname*{argmin}_{y \in Q} g(y_k) + \langle y - y_k, \nabla g(y_k) \rangle + LD(y, y_k)$.

Reinterpretation as g.c.g

$$\min_{x \in Q} \; \underbrace{g(x) - Ld(x)}_{f(x)} + Ld(x)$$

Step 1: $y \leftarrow \min_{x \in Q} \langle x, \nabla f(x_k) \rangle + Ld(x)$

Note that $y = y_{k+1}$.

Step 3: $x_{k+1} \leftarrow (1 - \eta) x_k + \eta y$.

Choose $\eta \equiv 1$ recovers mirror descent.

Regularization

The rise of regularization

$$\min_x g(x) \quad \text{where} \quad g(x) := f(x) + h(x).$$

As usual, $f \in \mathcal{C}_L^1$ is convex; h, the regularizer, is convex but not necessarily smooth.
Fits nicely into the generalized conditional gradient framework.

Trace norm regularization

$$\min_X f(X) + \lambda \cdot \|X\|_{\mathrm{tr}}$$

(Optimal) gradient needs to solve: $\min_Y \frac{1}{2}\|Y - Z\|_F + \lambda \cdot \|Y\|_{\mathrm{tr}}$.

G.c.g. needs to solve: $\min_Y \langle Z, Y \rangle + \lambda \cdot \|Y\|_{\mathrm{tr}}$,

Regularization

The rise of regularization

$$\min_x g(x) \quad \text{where} \quad g(x) := f(x) + h(x).$$

As usual, $f \in \mathcal{C}_L^1$ is convex; h, the regularizer, is convex but not necessarily smooth.

Fits nicely into the generalized conditional gradient framework.

Trace norm regularization

$$\min_X f(X) + \lambda \cdot \|X\|_{\mathrm{tr}}$$

(Optimal) gradient needs to solve: $\min_Y \frac{1}{2}\|Y - Z\|_F + \lambda \cdot \|Y\|_{\mathrm{tr}}$.

G.c.g. needs to solve: $\min_Y \langle Z, Y \rangle + \lambda \cdot \|Y\|_{\mathrm{tr}}$,
and the line search... Wooooops!

Gauge (Minkowski) function (Tewari, Ravikumar and Dhillon, 2012)

Given any convex, balanced, absorbing and bounded set A, define the gauge function $p_K(x) := \inf\{\sigma > 0 : x \in \sigma K\}$, which is a norm.

Trace norm

$\|X\|_{\mathrm{tr}} = p_K(X), \mathrm{where}\ K := \mathrm{conv}(W), W := \{uv' : \|u\|_2 = 1, \|v\|_2 = 1\}$.

Reformulation

$$\min_X f(X) + \lambda \|X\|_{\mathrm{tr}} \Leftrightarrow \min_{\boldsymbol{\sigma} \in c_{00}^+} f(\langle \mathbf{w}; \boldsymbol{\sigma} \rangle) + \lambda \sum_i \sigma_i$$

Cons: Infinite-dim!

Gauge (Minkowski) function (Tewari, Ravikumar and Dhillon, 2012)

Given any convex, balanced, absorbing and bounded set A, define the gauge function $p_K(x) := \inf\{\sigma > 0 : x \in \sigma K\}$, which is a norm.

Trace norm

$\|X\|_{\mathrm{tr}} = p_K(X), \text{where } K := \mathrm{conv}(W), W := \{uv' : \|u\|_2 = 1, \|v\|_2 = 1\}.$

Reformulation

$$\min_X f(X) + \lambda \|X\|_{\mathrm{tr}} \Leftrightarrow \min_{\boldsymbol{\sigma} \in c_{00}^+} f(\langle \mathbf{w}; \boldsymbol{\sigma} \rangle) + \lambda \sum_i \sigma_i$$

Cons: Infinite-dim!

Pros: Infinite-dim! Line search much cheaper.

The power of local search (Zhang, Yu and Schuurmans, 2012)

Recall

Serious drawback of conditional gradient: sublinear rate!

Yet another reformulation

$$\min_X f(X) + \lambda\|X\|_{\mathrm{tr}} \Leftrightarrow \min_{U,V} f(UV) + \tfrac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2)$$

RHS is smooth unconstrained, albeit nonconvex...

Hybridize!

1. one step conditional gradient on the LHS;
2. construct initializer for RHS;
3. run lbfgs on RHS until local convergence;
4. construct initializer for LHS;

The power of local search (Zhang, Yu and Schuurmans, 2012)

Recall

Serious drawback of conditional gradient: sublinear rate!

Yet another reformulation

$$\min_X f(X) + \lambda \|X\|_{\mathrm{tr}} \Leftrightarrow \min_{U,V} f(UV) + \tfrac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2)$$

RHS is smooth unconstrained, albeit nonconvex...

Hybridize!

1. one step conditional gradient on the LHS;
2. construct initializer for RHS;
3. run lbfgs on RHS until local convergence;
4. construct initializer for LHS;

It works extremely well!