Low iteration cost algorithms for large-scale nonsmooth learning

Anatoli Juditsky*
joint research with Arkadi Nemirovski†
*University J. Fourier, †ISyE, Georgia Tech, Atlanta

Optimization & Bigdata
Edinburgh, May 1-2, 2013

First-order methods background

The problem of interest here is a convex optimization problem in the form

$$\mathrm{Opt}(P) = \max_{x \in X} f_*(x) \quad \text{(P)}$$

where X is a nonempty closed and bounded subset of Euclidean space E_x, and f_* is concave and Lipschitz continuous function on X.

We are interested in large-scale problems, so that the methods of choice are the first-order (FO) optimization techniques.

The standard FO approach to (P) requires to equip X, E_x with a *proximal setup* $(\|\cdot\|_x, \omega_x(\cdot))$.

First-order methods background

That means to equip the space E_x with a norm $\|\cdot\|_x$, and the domain X of the problem with a *distance-generating function* (d.-g.f.) $\omega_x : X \to \mathbb{R}$.

- $\omega(\cdot)$ should be convex and continuous on X, admit a continuous in $z \in X^o = \{z \in X : \partial\omega_x(z) \neq \emptyset\}$ selection $\omega_x'(z)$ of subgradients, and be strongly convex, modulus 1, w.r.t. $\|\cdot\|_x$:

$$\langle \omega_x'(z) - \omega_x'(z'), z - z' \rangle \geq \|z - z'\|_x^2.$$

- A *step* of a FO method essentially reduces to a single computation of f_*, f_*' at a point and a single computation of the *prox-mapping*

$$\mathrm{Prox}_x(\xi) := \operatorname*{argmax}_{z \in X} \left[\langle \xi - \omega'(x), z \rangle - \omega(z) \right].$$

for a pair $x \in X^0, \xi \in E_x$.

First-order methods background

Then generating an ϵ-solution to the problem – a point $x_\epsilon \in X$ satisfying

$$\mathrm{Opt} - f_*(x_\epsilon) \leq \epsilon$$

costs at most $N(\epsilon)$ *steps* of the method, where

- $N(\epsilon) = O(1)\frac{\Omega_X^2 L^2}{\epsilon^2}$ when f_* is Lipschitz continuous, with constant L w.r.t. $\|\cdot\|_X$ (Mirror Descent algorithm, Nemirovski, Yudin (1978)), and
- $N(\epsilon) = O(1)\Omega_X\sqrt{\frac{D}{\epsilon}}$ in the smooth case, where f_* possesses Lipschitz continuous, with constant D, gradient: $\|f_*'(z) - f_*'(z')\|_{X,*} \leq D\|z - z'\|_X$, where $\|\cdot\|_{X,*}$ is the norm conjugate to $\|\cdot\|$ (Nesterov's optimal algorithm, Nesterov (1983)).

In the above bounds

$$\Omega_X = \Omega[X, \omega(\cdot)] = \sqrt{2[\max_{z\in X}\omega_X(z) - \min_{z\in X}\omega_X(z)]}$$

is the ω-*diameter* of X.

First-order methods background

Another way to process (P) by FO methods originates in Nesterov (2003), Nemirovski (2003). It is based on using Fenchel-type representation of f_*:

$$f_*(x) = \min_{y \in Y} \left[F(x, y) := \langle x, Ay + a \rangle + \psi(y) \right],$$

where Y is a closed and bounded subset of Euclidean space E_y and $\psi(y)$ is a convex function.

- The implementation of the FO method requires proximal setups $(\|\cdot\|_x, \omega_x(\cdot))$ and $(\|\cdot\|_y, \omega_y(\cdot))$ for (E_x, X) and for (E_y, Y);
- A step of the method requires a single computation of $\nabla F(\cdot)$ at a point and computing the values of two prox-mappings, one associated with $(X, \omega_x(\cdot))$, and the other one associated with $(Y, \omega_y(\cdot))$.

First-order methods background

The method finds an ϵ-solution to (P) in

$$N(\epsilon) \leq O(1)\frac{L_{xx}\Omega_X^2 + 2L_{xy}\Omega_X\Omega_Y + L_{yy}\Omega_Y^2}{\epsilon}$$

steps. Here

- $[\Omega_X = \Omega[X, \omega_x(\cdot)], \Omega_Y = \Omega[Y, \omega_y(\cdot)]]$ are the ω-diameters of X and Y;
- L_{xx}, L_{yy}, L_{xy} are the partial Lipschitz constants of $\nabla F(x, y)$, specifically,

$$\begin{array}{l} \forall (x, x' \in X, y, y' \in Y): \\ \qquad \|\nabla_x F(x', y') - \nabla_x F(x, y)\|_{x,*} \leq L_{xx}\|x' - x\|_x + L_{xy}\|y' - y\|_y, \\ \qquad \|\nabla_y F(x', y') - \nabla_y F(x, y)\|_{y,*} \leq L_{xy}\|x' - x\|_x + L_{yy}\|y' - y\|_y, \end{array}$$

and $\|\cdot\|_{x,*}$, $\|\cdot\|_{y,*}$ are the norms conjugate to $\|\cdot\|_x$, $\|\cdot\|_y$, respectively.

First-order methods background

To summarize, FO methods of this type rely upon “good” proximal setups –

- those resulting in “moderate” values of Ω_X (and Ω_X, if needed) and
- not too difficult to compute prox-mappings.

This is indeed the case for domains X arising in numerous applications.

The question we address here is

what to do when X does not admit a “good” proximal setup?

First-order methods background

Here are two instructive examples:

A X is the unit ball of the nuclear norm $\|\cdot\| = \sum_i \sigma_i(\cdot)$ (sum of singular values) in the space $\mathbb{R}^{p\times q}$ of $p \times q$ matrices.

In this case, X *does admit* a proximal setup with moderate Ω_X:

$$\Omega_X = O(1)\sqrt{\ln(pq)}.$$

However, computing the prox-mapping becomes prohibitively time-consuming when p, q are large.

B X is a high-dimensional box – the $\|\cdot\|_\infty$-norm of $\mathbb{R}^m$ with large m.

More generally, the unit ball of the ℓ_∞/ℓ_2 norm $\|x\|_{\infty|2} = \max_{1\le i\le m}\|x^i\|_2$, where $x = [x^1; \ldots; x^m] \in E = \mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_m}$.

Here it is easy to point out a proximal setup with easy-to-compute prox mapping – the *Euclidean setup*

$$\|\cdot\| = \|\cdot\|_2, \quad \omega(x) = \frac{1}{2}\langle x, x\rangle.$$

However, it is easily seen that for such X one has $\Omega_X \ge O(1)\sqrt{m}$ for every proximal setup.

Motivating examples: Matrix completion

Let $x \in \mathbb{R}^{p\times q}$, and let $a \in \mathbb{R}^m$,

$$a = \mathcal{A}x + \xi.$$

Here $\mathcal{A}\cdot : \ \mathbb{R}^{p\times q} \to \mathbb{R}^m$ is a given affine mapping, and $\xi \in \mathbb{R}^m, \quad \|\xi\| \leq \delta$ is an unknown perturbation, and $\|\cdot\|$ is a norm on $\mathbb{R}^m$.

The underlying matrix is assumed to be low-rank and is recovered from the observed entries by norm minimization:

$$\hat{x} \in \operatorname*{Argmin}_{z}\{\|x\|_x : \ \|a - \mathcal{A}x\| \leq \delta\}$$

where $\|\cdot\|_x$ is usually chosen to be the nuclear norm: $\|u\|_x = \|\sigma(u)\|_1$.

The problem above can be reduced to a small sequence of problems, parameterized with $R > 0$:

$$\mathrm{Opt}(R) = \min_{\|x\|_x \leq R}[-f^*(x) = \|a - \mathcal{A}x\|] : \qquad (\mathrm{TR}_1).$$

Matrix completion

Note that (TR$_1$) can be rewritten as a linear matrix game:

$$\begin{aligned}\mathrm{Opt}(R) &= -\max_{x:\,\|x\|_x\le R}\left[f^*(x)=\min_{y:\,\|y\|_*\le 1}\langle y,\mathcal{A}x-a\rangle\right]\\ &= \min_{y\in Y}\max_{x\in X}\langle y,\mathcal{A}x\rangle+\psi(y)\quad (\mathrm{TR'}_1)\end{aligned}$$

where

$$Y=\{y\in\mathbb{R}^m:\ \|y\|_*\le 1\},\quad X=\{x\in\mathbb{R}^{p\times q}:\ \|x\|_x\le R\},\quad \psi(y)=\langle y,a\rangle,$$

and $\|\cdot\|_*$ is the norm, dual to $\|\cdot\|$.

Multi-class classification

Assume that we observe

- N "feature vectors" $z_j \in \mathbb{R}^q$ belonging to one of M non-overlapping classes,
- labels $\chi_j \in \mathbb{R}^M$ which are basic orths in $\mathbb{R}^M$. The index of the (only) nonzero entry in e_j is the number of the class to which z_j belongs.

A multi-class classifier, specified by a matrix $x \in \mathbb{R}^{M\times q}$ and a vector $b \in \mathbb{R}^M$ is

$$i(z) \in \mathop{\mathrm{Argmax}}_{1\leq i\leq M} [xz+b]_i.$$

Denote $\bar{\chi}:\ [\bar{\chi}]_i = 1-[\chi]_i.\ i=1,...,M$. Given a feature vector z and the corresponding label χ, let us set

$$h = h(x,b;z,\chi) = [xz+b] - [\chi^T[xz+b]]\mathbf{1} + \bar{\chi} \in \mathbb{R}^M \qquad [\mathbf{1}=[1;...;1]\in\mathbb{R}^M].$$

Observe that

- h is nonpositive iff the classifier, given by x, b "recovers the class i_* of z with margin 1";
- if some entries in $[xz+b]$ are $\geq$ the entry with index i_*, then $\max_i[h]_i \geq 1$.

Multi-class classification

Let us set

$$\eta(x,b;z,\chi)=\max_{1\leq j\leq M}(0,[h(x,b;z,\chi)]_j),$$

and $H(x,b)=\mathbf{E}\{\eta(x,b;z,\chi)\}$.

We conclude that $H(x,b)$ is an upper bound on the risk of the classifier (x,b)

For the sake of simplicity, we assume from now on that $b=0$.

When replacing the true expectation by its empirical counterpart

$$H_N(x,b)=N^{-1}\sum_{i=1}^{N}\eta(x,b;z_i,\chi_i),$$

we come to the problem

$$\min_{x\in X}\left[-f_*(x)=N^{-1}\sum_{i=1}^{N}\max_{1\leq j\leq M}(0,[xz_i-[\chi_i^T xz_i]\mathbf{1}+\bar{\chi}_i]_j)\right], \quad X=\{x:\|x\|_x\leq R\}. \quad (\mathrm{CL}_1)$$

Multi-class classification

Noting that $\max_i(0, h_i) = \max_u\{u^T h : u \geq 0, \sum_i u_i \leq 1\}$, we write

$$N^{-1} \max_{1\leq j\leq M}(0, [xz_i - [\chi_i^T xz_i]\mathbf{1} + \bar{\chi}_i]_j) = \max_{y^i\in Y^i}\langle y^i, xz_i - [\chi_i^T xz_i]\mathbf{1} + \bar{\chi}_i\rangle,$$

where $Y^i = \{y \in \mathbb{R}^M_+,\ \sum_j [y]_j \leq N^{-1}\}$.

Now (CL$_1$) transforms into

$$\max_{x\in X}\left[f_*(x) = \min_{y\in Y}[\langle y, Bx\rangle + \psi(y)]\right] \quad [X = \{x :\ \|\cdot\|_x \leq R\}] \ \ (\text{CL'}_1)$$

where

$$\begin{array}{l} Y = \left\{y = [y^1; ...; y^N],\ y^i \in \mathbb{R}^M_+,\ \sum_j [y^i]_i \leq N^{-1},\ 1 \leq i \leq N\right\} \in E_y = \mathbb{R}^{MN} \\ Bx = [B^1 x; ...; B^N x],\ B^i x = \left[z_i^T[x^{j(i)} - x^1]; ...; z_i^T[x^{j(i)} - x^M]\right],\ 1 \leq i \leq N \\ \psi(y) = \psi(y^1, ..., y^N) = -\sum_{i=1}^N [y^i]^T \bar{\chi}_i; \end{array}$$

here $i(j)$ is the class of z_j (the index of the only nonzero entry in χ_j) and x^k is the kth row of x.

General problem setting

Let E_x be an Euclidean space, $X \subset E_x$ closed and bounded convex set, equipped with *LO oracle*.

Suppose that $f_*(x)$ is a concave function given by *Fenchel-type representation*:

$$f_*(x) = \min_{y \in Y} \left[\langle x, Ay + a \rangle + \psi(y) \right],$$

where Y is a closed compact subset of an Euclidean space E_y and ψ is a convex Lipschitz-continuous function on Y given by a First Order oracle.

Our standing assumption is as follows:

- Y (but not X!) does admit a proximal setup $(\|\cdot\|_y, \omega_y(\cdot))$.
- an efficient Linear Optimization oracle (LO) is available for X – a routine which, given a linear form ξ returns a point

$$x[\xi] \in \mathop{\mathrm{Argmax}}_{z \in X} \langle \xi, z \rangle.$$

General problem setting

In the sequel we set

$$f(y) = \max_{x \in X} \left[\langle x, Ay + a \rangle + \psi(y) \right],$$

and consider two optimization problems

$$\begin{array}{rcll} \mathrm{Opt}(P) & = & \max\limits_{x \in X} f_*(x) & \quad \text{(P)} \\ \mathrm{Opt}(D) & = & \min\limits_{y \in Y} f(y) & \quad \text{(D)} \end{array}$$

By the standard saddle-point argument, we have $\mathrm{Opt}(P) = \mathrm{Opt}(D)$.

Duality approach

Let $f(y)$ be the dual objective:

$$f(y) = \max_{x \in X} [\langle x, Ay + a \rangle] + \psi(y).$$

Note that the LO oracle for X along with the FO oracle for ψ provide a FO oracle for (D):

$$f'(y) = A^T x(y) + \psi'(y), \ x(y) := x[Ay + a].$$

Since Y admits a good proximal setup we may act as follows:

- find an ϵ-solution to the dual problem (D),
- recover a “good approximate solution” to (P) (problem of interest) from the just found solution to (D).

Accuracy certificates: a summary

- When solving (D) by a FO method, we generate *search points* $y_\tau \in Y$ where the subgradients $f'(y_\tau)$ of f are computed.

  As a byproduct of the latter computation, we have at our disposal the corresponding primal solutions – the points $x_\tau = x(y_\tau)$.

- As a result, after t steps we have at our disposal the execution protocol $y^t = \{y_\tau, f'(y_\tau)\}_{\tau=1}^t$,
- and an accuracy certificate associated with this protocol is a collection $\lambda^t = \{\lambda_\tau^t\}_{\tau=1}^t, \ \ \lambda_\tau^t \geq 0, \ \ \sum_{\tau=1}^t \lambda_\tau^t = 1$.

  The resolution of the certificate is the quantity

$$\epsilon(y^t, \lambda^t) = \max_{y \in Y} \sum_{\tau=1}^t \lambda_\tau^t \langle f'(y_\tau), y_\tau - y \rangle.$$

Main observation

Proposition

Let y^t be search points, generated by the method solving (D), λ^t be a certificate, and

$$\begin{aligned} y(y^t,\lambda^t) &= \sum_{\tau=1}^{t}\lambda_\tau y_\tau, \\ x(y^t,\lambda^t) &= \sum_{\tau=1}^{t}\lambda_\tau x[Ay_\tau + a], \\ \epsilon(y^t,\lambda^t) &= \max_{y\in Y}\sum_{\tau=1}^{t}\lambda_\tau\langle f'(y_\tau), y_\tau - y\rangle. \end{aligned}$$

Then $\widehat{x} := x(y^t,\lambda^t)$, $\widehat{y} := y(y^t,\lambda^t)$ are feasible solutions to problems (P), (D), respectively, and

$$f(\widehat{y}) - f_*(\widehat{x}) = \left[f(\widehat{y}) - \mathrm{Opt}(D)\right] + \left[\mathrm{Opt}(P) - f_*(\widehat{x})\right] \le \epsilon(y^t,\lambda^t).$$

Proof:

Denote $F(x,y) = \langle x, Ay+a\rangle + \psi(x)$ and $x(y) = x[Ay+a]$, so that $f(y) = F(x(y),y)$. Observe that $f'(y) = F'_y(x(y),y)$, where

$$F'_y(x,y) \in \partial_y F(x,y) \text{ for all } x \in X,\ y \in Y.$$

Setting $x_\tau = x(y_\tau)$, we have for any $y \in Y$:

$$\begin{array}{rl}
\multicolumn{2}{l}{\epsilon(y^t,\lambda^t) = \sum\limits_{\tau=1}^{t} \lambda_\tau \langle f'(y_\tau), y_\tau - y\rangle = \sum\limits_{\tau=1}^{t} \lambda_\tau \langle F'_y(x_\tau,y_\tau), y_\tau - y\rangle} \\
\geq & \sum\limits_{\tau=1}^{t} \lambda_\tau \left[F(x_\tau,y_\tau) - F(x_\tau,y)\right] \text{ [by convexity of } F \text{ in } y] \\
= & \sum\limits_{\tau=1}^{t} \lambda_\tau \left[f(y_\tau) - F(x_\tau,y)\right] \text{ [since } x_\tau = x(y_\tau)\text{, so that } F(x_\tau,y_\tau) = f(y_\tau)] \\
\geq & f(\widehat{y}) - F(\widehat{x},y) \text{ [by convexity of } f \text{ and concavity of } F(x,y) \text{ in } x] \\
\Rightarrow & \epsilon(y^t,\lambda^t) \geq \max\limits_{y \in y} \left[f(\widehat{y}) - F(\widehat{x},y)\right] = f(\widehat{y}) - f_*(\widehat{x}).
\end{array}$$

The inclusions $\widehat{x} \in X$, $\widehat{y} \in Y$ are evident. □

Accuracy certificates: a summary

Thus, assuming that

- the FO method for minimizing $f(y)$ produces, in addition to search points, accuracy certificates for the resulting execution protocols, and that
- the resolution of these certificates goes to 0 as $t \to \infty$ at some rate,

the certificates can be used to build feasible approximate solutions to (D) and to (P) with accuracies (in terms of the objectives of the respective problems) going to 0 at the same rate.

The question is

> *Are we able to equip known methods with a computationally cheap accuracy certificates with the resolutions identical to the standard efficiency estimates of the methods?*

Proximal setup

Proximal setup for Y is given by a norm $\|\cdot\|$ on E_y and a distance-generating function (d.-g.f.) $\omega: \; Y \to \mathbb{R}$.

ω should be continuous and convex on Y, should admit a continuous in $y \in Y^o = \{y \in Y : \partial\omega(y) \neq \emptyset\}$ selection of subdifferentials $\omega'(y)$, and should be strongly convex, modulus 1, w.r.t. $\|\cdot\|$:

$$\forall y, y' \in Y^o : \langle \omega'(y) - \omega'(y'), y - y' \rangle \geq \|y - y'\|^2.$$

It gives rise to

- ω-center $y_\omega = \mathrm{argmin}_{y \in Y}\, \omega(y)$ of Y;
- ω-diameter $\Omega = \Omega[Y, \omega(\cdot)] := \sqrt{2\left[\max_{y\in Y} \omega(y) - \min_{y \in Y} \omega(y)\right]}$.

Proximal setup

- Bregman distance

$$V_y(z) = \omega(z) - \omega(y) - \langle \omega'(y), z - y \rangle \quad (y \in Y^o,\ z \in Y).$$

Due to strong convexity of ω, we have

$$\forall (z \in Y, y \in Y^o) : V_y(z) \geq \frac{1}{2}\|z - y\|^2;$$

Observe that $\langle \omega'(y_\omega), y - y_\omega \rangle \geq 0$, so that

$$V_{y_\omega}(z) \leq \omega(z) - \omega(y_\omega) \leq \frac{1}{2}\Omega^2,\ \forall z \in Y, \quad \text{and} \quad \|y - y_\omega\| \leq \Omega,\ \forall y \in Y.$$

- Prox-mapping

$$\mathrm{Prox}_y(\xi) = \operatorname*{argmin}_{z \in Y} \left[\langle \xi, z \rangle + V_y(z) \right],$$

where $\xi \in E_y$ and $y \in Y^o$.

Proximal setup

$\mathrm{Prox}_y(\cdot)$ takes its values in Y^o and satisfies the inequality: $\forall (y \in Y^o, \xi \in E_y)$

$$y_+ = \mathrm{Prox}_y(\xi) : \ \langle \xi, y_+ - z \rangle \le V_y(z) - V_{y_+}(z) - V_y(y_+) \ \ \forall z \in Y.$$

Indeed, by the optimality condition

$$\forall z \in Y, \ \ 0 \le \langle \xi + V'_y(y_+), z - y_+ \rangle = \langle \xi + \omega'(y_+) - \omega'(y), z - y_+ \rangle.$$

So

$$\begin{array}{l} V_{y_+}(z) - V_y(z) = [\omega(z) - \langle \omega'(y_+), z - y_+ \rangle - \omega(y_+)] - [\omega(z) - \langle \omega'(y), z - y \rangle - \omega(y)] \\ = \langle \xi, z - y_+ \rangle - \langle \xi + \omega'(y_+) - \omega'(y), z - y_+ \rangle - [\omega(y_+) - \langle \omega'(y), y_+ - y \rangle - \omega(y)] \\ \le \langle \xi, z - y_+ \rangle - V_y(y_+) \end{array}$$

□

Generating certificates: Mirror Descent

We assume to be given an oracle represented vector field

$$y \mapsto g(y) : Y \to E_y, \tag{1}$$

From now on we assume that g is bounded:

$$\|g(y)\|_* \leq L[g] < \infty, \forall y \in Y,$$

where $\|\cdot\|_*$ is the norm conjugate to $\|\cdot\|$.

Mirror Descent algorithm (MD) is given by the recurrence

$$\begin{array}{l} \text{Set } y_1 = y_\omega; \\ \text{Given } y_\tau \text{ compute } g_\tau := g(y_\tau) \text{ and } y_{\tau+1} := \mathrm{Prox}_{y_\tau}(\gamma_\tau g_\tau), \end{array} \tag{MD}$$

where $\gamma_\tau > 0$ are stepsizes.

We equip this recurrence with the accuracy certificate

$$\lambda^t = \left(\sum_{\tau=1}^t \gamma_\tau\right)^{-1} [\gamma_1; \ldots; \gamma_t]. \tag{2}$$

Generating certificates: Mirror Descent

Proposition

For every t, the resolution

$$\epsilon(y^t,\lambda^t) := \max_{y\in Y}\sum_{\tau=1}^{t}\lambda_\tau\langle g(y_\tau), y_\tau - y\rangle$$

of λ^t on the execution protocol $y^t = \{y_\tau, g(y_\tau)\}_{\tau=1}^t$ satisfies the standard MD efficiency estimate

$$\epsilon(y^t,\lambda^t) \le \frac{\Omega^2 + \sum_{\tau=1}^t \gamma_\tau^2 L^2[g]}{2\sum_{\tau=1}^t \gamma_\tau}.$$

In particular, if $\gamma_\tau = \frac{\Omega}{\sqrt{t}\|g(y_\tau)\|_*}$ for $1 \le \tau \le t$ then

$$\epsilon(y^t,\lambda^t) \le \frac{\Omega L[g]}{\sqrt{t}}.$$

Proof

is given by the standard MD rate-of-convergence derivation: by the prox-mapping identity,

$$\forall z \in Y : \langle \gamma_\tau g_\tau, y_{\tau+1} - z \rangle \leq V_{y_\tau}(z) - V_{y_{\tau+1}}(z) - V_{y_\tau}(y_{\tau+1}).$$

Thus ($\forall z \in Y$)

$$\begin{array}{rcl} \langle \gamma_\tau g_\tau, y_\tau - z \rangle & \leq & V_{y_\tau}(z) - V_{y_{\tau+1}}(z) + [\ \underbrace{\langle g_\tau, y_\tau - y_{\tau+1} \rangle}_{\leq \gamma_\tau \|g_\tau\|_* \|y_\tau - y_{\tau+1}\|} - \underbrace{V_{y_\tau}(y_{\tau+1})}_{\geq \frac{1}{2}\|y_{\tau+1} - y_\tau\|^2}\ ] \\ & \leq & V_{y_\tau}(z) - V_{y_{\tau+1}}(z) + \frac{1}{2}\gamma_\tau^2 \|g_\tau\|_*^2 . \end{array}$$

Then ($\forall z \in Y$) we obtain, when summing up,

$$\begin{array}{rcl} \sum_{\tau=1}^{t} \gamma_t \langle g_\tau, y_\tau - z \rangle & \leq & \underbrace{V_{y_w}(z)}_{\leq \frac{1}{2}\Omega^2} - \underbrace{V_{y_{t+1}}(z)}_{\geq 0} + \frac{1}{2} \sum_{\tau=1}^{t} \gamma_\tau^2 \underbrace{\|g_\tau\|_*^2}_{\leq L^2[g]} \\ & \leq & \frac{\Omega^2 + \sum_{\tau=1}^{t} \gamma_\tau^2 L^2[g]}{2\sum_{\tau=1}^{t} \gamma_\tau}, \end{array}$$

what implies the first statement of the proposition. □

Solving (P) and (D) with Mirror Descent

In order to solve (D), we apply MD to the vector field $g = f'$. Assuming that

$$L_f = \sup_{y \in Y} \{ \|f'(y)\|_* \equiv \|A^* x(y) + \psi'(y)\|_* \} < \infty \quad [x(y) = x[Ay + a]]$$

we can set $L[g] = L_f$. The described certificate λ^t and MD execution protocol $y^t = \{y_\tau, g(y_\tau) = f'(y_\tau)\}_{\tau=1}^t$ yield feasible approximate solutions to (P) and to (D):

$$\widehat{x}^t = \sum_{\tau=1}^t \lambda_\tau^t x_\tau, \quad \widehat{y}^t = \sum_{\tau=1}^t \lambda_\tau^t y_\tau,$$

such that

$$f(\widehat{y}^t) - f_*(\widehat{x}^t) \leq \epsilon(y^t, \lambda^t).$$

Solving (P) and (D) with Mirror Descent

Corollary

For every $t = 1, 2, \ldots$ the t-step MD with the stepsize policy

$$\gamma_\tau = \frac{\Omega}{\sqrt{t}\|f'(x_\tau)\|_*},\ 1 \le \tau \le t$$

as applied to (D) yields feasible approximate solutions $\widehat{x}^t$, $\widehat{y}^t$ to (P), (D) such that

$$[f(\widehat{y}^t) - \mathrm{Opt}(P)] + [\mathrm{Opt}(D) - f_*(\widehat{x}^t)] \le \frac{\Omega L_f}{\sqrt{t}} \qquad [\Omega = \Omega[Y, \omega(\cdot)]] .$$

In particular, given $\epsilon > 0$, it takes at most $\mathrm{Ceil}\left(\frac{\Omega^2 L_f^2}{\epsilon^2}\right)$ steps of the MD algorithm to ensure that

$$[f(\widehat{y}^t) - \mathrm{Opt}(P)] + [\mathrm{Opt}(D) - f_*(\widehat{x}^t)] \le \epsilon.$$

Generating certificates: Mirror Level method

Mirror Level algorithm (ML) is aimed to process an oracle represented vector bounded field $g: \ Y \to E_y$ ($\|g(y)\|_* \leq L[g] < \infty, \ \forall y \in Y$).

We associate with $y_\tau \in Y$ the affine function

$$h_\tau(z) = \langle g(y_\tau), y_\tau - z \rangle,$$

and with a finite set $S \subset \{1, ..., t\}$ the family $\mathcal{F}_S$ of affine functions on E_y. Denote

$$h(z) = \sum_{\tau \in S} \lambda_\tau h_\tau(z), \ \text{ with } \lambda_\tau \geq 0, \ \sum_{\tau \in S} \lambda_\tau = 1.$$

The goal of the algorithm is, given a tolerance $\epsilon > 0$, find a finite set $\{y_\tau, \tau \in S\} \subset Y$ and $h \in \mathcal{F}_S$ such that

$$\max_{z \in Y} h(z) \leq \epsilon.$$

In other words, we are to build an “incomplete” execution protocol $y^S = \{y_\tau, g(y_\tau)\}_{\tau \in S}$ and an associated accuracy certificate λ^S such that

$$\epsilon(y^S, \lambda^S) = \max_{z \in Y} \sum_{\tau \in S} \lambda_\tau \langle g(y_\tau), y_\tau - z \rangle \leq \epsilon.$$

Construction

Steps of ML are split into subsequent phases, numbered $s = 1, 2, \ldots$. We associate to each phase s the optimality gap $\Delta_s > 0$.

We initialize the method with $y_1 = y_\omega$ $S_0 = \emptyset$, $\Delta_0 = +\infty$, and $\gamma \in (0, 1)$. At a step t we act as follows:

- given y_t compute $g(y_t)$, thus getting $h_t(z) = \langle g(y_t), y_t - z\rangle$, and set $S_t^+ = S_{t-1} \cup \{t\}$.
- solve the auxiliary problem

$$\epsilon_t = \max_{y \in Y} \min_{\tau \in S_t^+} h_\tau(y) = \max_{y \in Y} \min_{\lambda} \left\{ \sum_{\tau \in S_t^+} \lambda_\tau h_\tau(y) : \lambda_\tau \geq 0, \sum_{\tau \in S_t^+} \lambda_\tau = 1 \right\} \quad (\text{AUX}_1)$$

We assume that as a result of solving (AUX_1), both ϵ_t and the optimal solution $\lambda_\tau^t, \ \tau \in S_t^+$, become known.

We set $\lambda_\tau = 0$ for $\tau \leq t$ which are not in S_t^+ to get the certificate λ^t for the protocol $y^t = \{y_\tau, g(y_\tau)\}_{\tau=1}^t$ and $h^t(\cdot) = \sum_{\tau=1}^t \lambda_\tau^t h_\tau(\cdot)$.

Note that, by construction,

$$\epsilon(y^t, \lambda^t) = \max_{y \in Y} h^t(y) = \epsilon_t.$$

Construction

If $\epsilon_t \leq \epsilon$, we terminate: we have $\max_{z \in Y} h^t(z) \leq \epsilon$. Otherwise we proceed as follows:

- If $\epsilon_t \leq \gamma\Delta_{s-1}$, we say that step t starts phase s (e.g., step $t = 1$ starts phase 1), and set
$$\Delta_s = \epsilon_t, \quad S_t = \{\tau : 1 \leq \tau \leq t : \lambda_\tau^t > 0\} \cup \{1\}, \, \widehat{y}_t = y_\omega.$$
  Otherwise we set $S_t = S_t^+$, $\widehat{y}_t = y_t$.

  Note that in both cases we have
$$\epsilon_t = \max_{y \in Y} \min_{\tau \in S_t} h_\tau(y).$$

- Finally, we define the t-th level as $\ell_t = \gamma\epsilon_t$ and associate with this quantity the level set
$$U_t = \{y \in Y : h_\tau(y) \geq \ell_t \, \forall t \in S_t\};$$
  we specify y_{t+1} as Bregman projection of $\widehat{y}_t$ on U_t:
$$y_{t+1} = \operatorname*{argmin}_{y \in U_t} V_{\widehat{y}_t}(y). \qquad (\mathrm{AUX}_2)$$
  and loop to step $t + 1$.

Construction

Remark The outlined algorithm requires solving at every step two nontrivial auxiliary optimization problems, specifically, (AUX$_1$) and (AUX$_2$). These problems may be solved efficiently, provided that

- Y and ω are "simple and fit each other," meaning that we can easily solve problems of the form
$$\min_{z \in Y} \left[\omega(z) + \langle a, z \rangle\right];$$
that is, our proximal setup for Y results in easy-to-compute prox-mapping;
- t is moderate.

On the other hand, the auxiliary problems become difficult when t increases.

Algorithm Non-Euclidean Restricted Memory Level (NERML), which originates in Ben-Tal, Nemirovski (2005), allows to control the complexity of the linear functions in the models, which never exceed a given memory control parameter of the method.

ML efficiency estimate

Proposition

Given on input a target tolerance $\epsilon > 0$, the ML algorithm terminates after finitely many steps, with the output

$$y^t = \{y_\tau, g(y_\tau)\}_{\tau=1}^t, \quad \lambda^t = \{\lambda_\tau^t \geq 0\}_{\tau=1}^t, \ \sum_\tau \lambda_\tau^t = 1,$$

such that

$$\epsilon_t(y^t, \lambda^t) \leq \epsilon.$$

The number of steps of the algorithm does not exceed

$$N = \frac{\Omega^2 L^2[g]}{\gamma^4(1-\gamma^2)\epsilon^2} + 1 \qquad [\Omega = \Omega[Y, \omega(\cdot)]].$$

Solving (P) and (D) with Mirror Level

is completely similar to the case of MD:

- given a desired tolerance $\epsilon > 0$, we apply ML to the vector field $g(y) = f'(y)$ until the target

  $$\max_{z\in Y}[h(z) = \sum_{\tau=1}^{t} \lambda_\tau \langle g(y_\tau), y_\tau - z\rangle] \leq \epsilon.$$

  is satisfied. We set $L[g] = L_f$, so that by the proposition it will be achieved in

  $$\mathrm{Ceil}\left(\frac{\Omega^2 L_f^2}{\gamma^4(1-\gamma^2)\epsilon^2} + 1\right) \quad [\Omega = \Omega[Y, \omega(\cdot)]]$$

  steps.

- The target being achieved, we have at our disposal an execution protocol $y^t = \{y_\tau, f'(y_\tau)\}_{\tau=1}^{t}$ along with accuracy certificate $\lambda^t = \{\lambda_\tau^t\}$ such that

  $$\epsilon(y^t, \lambda^t) \leq \epsilon.$$

  Therefore, specifying $\widehat{x}^t$, $\widehat{y}^t$, we ensure

  $$[f(\widehat{y}^t) - \mathrm{Opt}(D)] + [\mathrm{Opt}(P) - f_*(\widehat{x}^t)] \leq \epsilon(y^t, \lambda^t).$$

An alternative: smoothing

We have assumed that the LO oracle is available for X.

If the objective f were smooth, with Lipschitz constant D_f of the gradient,

$$\|f'(x) - f'(z)\|_{X,*} \leq D_f \|x - z\|_X,$$

one could use Conditional Gradient algorithm (CoG) to solve (P)!

Note that in this case solving (P) up to accuracy ϵ would take $8\frac{R^2 D_f}{\epsilon}$ steps of CoG.

Approach by Conditional Gradients with Smoothing (CoGS)

- use the proximal setup for Y to smooth f_*
- use Conditional Gradient method to maximize the resulting approximation

Smoothing

How to smooth (cf. Nesterov (2003)):

assume that ω_Y is strongly convex with respect to $\|\cdot\|_y$, $\min_{y\in Y}\omega_Y(y)=0$, and set

$$f_*^\lambda(x)=\min_{y\in Y}\left[\langle x,Ay+a\rangle+\psi(y)+\lambda\omega_Y(y)\right] \quad \text{(S)}$$

Given a desired tolerance ϵ, choose

$$\lambda=\lambda(\epsilon):=\frac{\epsilon}{\Omega_Y^2}, \quad \Omega_Y=\Omega[Y,\omega_Y(\cdot)].$$

Then

- f_*^λ is concave, and for all $x\in X$, $\quad f_*(x)\le f_*^\lambda(x)\le f_*(x)+\frac{\epsilon}{2}$.
- $\forall\,(z,z'\in X)$, $\|\nabla f_*^\lambda(z)-\nabla f_*^\lambda(z')\|_{x,*}\le\lambda^{-1}\|A\|_{y;x,*}^2\|z-z'\|_x$, where

$$\begin{aligned}\|A\|_{y;x,*} &= \max\{\|A^*z\|_{y,*} : z\in E_x, \|z\|_x\le 1\}\\ &= \max\{\|Av\|_{x,*} : v\in E_y, \|v\|_y\le 1\}.\end{aligned}$$

Smoothing

When assuming that an optimal solution $y(x)$ of (S) is easily available, we have at our disposal a FO oracle for f_*^λ:

$$f_*^\lambda(x) = \langle x, Ay(x) + a + \psi(y(x)) + \lambda\omega_y(y(x)), \quad \nabla f_*^\lambda(x) = A^* y(x).$$

Using this oracle, along with the LO oracle for X, we can solve (P) by Conditional Gradients with Smoothing.

Proposition

Assume X is contained in $\|\cdot\|_X$-ball of radius R of E_X. Then to find an ϵ-minimizer of f_* it suffices to find an $\frac{\epsilon}{2}$-minimizer of f_*^λ, what takes

$$32\frac{\Omega_y^2 R^2 \|A\|_{y;x,*}^2}{\epsilon^2}$$

steps of CoG.

Discussion

Suppose that

- X is contained in centered at the origin ball of the norm $\|\cdot\|_x$ of radius R, and
- subgradients of ψ satisfy the bound $\|\psi'(y)\|_* \leq L_\psi$, and the Lipschitz constant of f satisfies:

$$L_f \leq \|A\|_{y;x,*} + L_\psi.$$

This results in the complexity bounds

$$O(1)\frac{[R\|A\|_{y;x,*} + L_\psi]^2\Omega^2[Y,\omega(\cdot)]}{\epsilon^2}$$

for MD and ML, and

$$O(1)\frac{[R\|A\|_{y;x,*}]^2\Omega^2[Y,\omega(\cdot)]}{\epsilon^2}$$

for CoGS.

When assuming that $L_\psi = O(1)R\|A\|_{y;x,*}$, these complexity bounds are essentially identical.

Discussion

- Formally, CoGS has a more restricted area of applications than MD/ML, since the related optimization problem may be more involved than computing prox-mapping associated with $Y, \omega(\cdot)$.

  At the same time, in vast majority of applications ψ is linear, and in this case the just outlined phenomenon disappears.
- What is in favor of CoGS, is its insensitivity to the Lipschitz constant of ψ. However, in the case of linear ψ the nonsmooth techniques admit simple modifications equally insensitive to L_ψ.
- On the other hand, the convergence pattern of nonsmooth methods utilizing memory (ML) is, at least at the beginning of the solution process, much better than it is predicted by their worst-case efficiency estimates.

  In particular, the nonsmooth approach “most probably,” significantly outperforms the smooth one when E_y is of moderate dimension.