Subgradient Ellipsoid Method for Nonsmooth Convex Problems

Anton Rodomanov (UCLouvain)
(joint work with Y. Nesterov)

May 6, 2022
FGP Conference on Optimization, Porto

Problem Setup

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

- $f : \mathbb{R}^n \to \mathbb{R}$ is a general convex Lipschitz continuous function.
- $Q \subseteq \mathbb{R}^n$ is a compact convex set with nonempty interior.

First-Order Oracle: Returns $f'(x) \in \partial f(x)$ for any $x \in \mathbb{R}^n$.

Separation Oracle: Checks if $x \in Q$. If not, returns $g_Q(x) \in \mathbb{R}^n \setminus \{0\}$:

$$\langle g_Q(x), x - y \rangle \geq 0, \qquad \forall y \in Q.$$

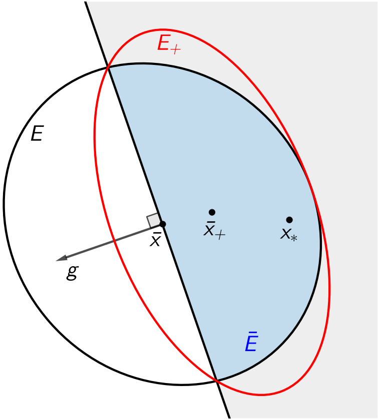
Ex: $Q = \{x : g(x) \leq 0\}$ for convex $g : \mathbb{R}^n \to \mathbb{R} \implies g_Q(x) = g'(x)$.

Combined Oracle: $\mathcal{G}(x) := \begin{cases} f'(x), & \text{if } x \in Q, \\ g_Q(x), & \text{if } x \notin Q. \end{cases}$

Main property: For an optimal solution x^*, we have

$$\langle \mathcal{G}(x), x - x^* \rangle \geq 0, \qquad \forall x \in \mathbb{R}^n.$$

Ellipsoid Method: Geometry



Ellispoid: For $\bar{x} \in \mathbb{R}^n$, $H \in \mathbb{S}^n_{++}$,

$$\mathcal{E}(\bar{x}, H) := \{x : \langle H^{-1}(x-\bar{x}), x-\bar{x}\rangle \leq 1\}.$$

Let $E := \mathcal{E}(\bar{x}, H)$, $g \in \mathbb{R}^n \setminus \{0\}$,

$$\bar{E} := \{x \in E : \langle g, \bar{x} - x\rangle \geq 0\}.$$

Minimum-volume ellipsoid $\supseteq \bar{E}$:

$$\bar{E} \subseteq E_+ := \mathcal{E}(\bar{x}_+, H_+),$$

where

$$\operatorname{vol} E_+ \leq \exp\bigl(-1/(2n)\bigr) \operatorname{vol} E.$$

Ellipsoid E_+ can be easily computed:

$$\bar{x}_+ = \bar{x} - \frac{1}{n+1}\frac{Hg}{\langle g, Hg\rangle^{1/2}}, \quad H_+ = \frac{n^2}{n^2-1}\Bigl(H - \frac{2}{n+1}\frac{Hgg^T H}{\langle g, Hg\rangle}\Bigr).$$

Ellipsoid Method: Algorithm

1. Choose $x_0 \in \mathbb{R}^n$ and $R > 0$ such that $Q \subseteq B(x_0, R)$.
2. Set $H_0 \coloneqq R^2 I$.
3. Iterate for $k \geq 0$:
   1. Query the oracle to obtain $g_k \coloneqq \mathcal{G}(x_k)$.
   2. Compute the center of the new ellipsoid:

      $$x_{k+1} \coloneqq x_k - \frac{1}{n+1} \frac{H_k g_k}{\langle g_k, H_k g_k \rangle^{1/2}},$$

   3. Compute the matrix of the new ellipsoid:

      $$H_{k+1} \coloneqq \frac{n^2}{n^2 - 1}\left(H_k - \frac{2}{n+1} \frac{H_k g_k g_k^T H_k}{\langle g_k, H_k g_k \rangle}\right).$$

Output: $x_k^* \coloneqq \operatorname{argmin}\big\{ f(x) : x \in \{x_0, \dots, x_{k-1}\} \cap Q \big\},\ k \geq 1.$

Ellipsoid Method: Complexity

To obtain $x_k^* \in Q$ such that $f(x_k^*) - f^* \leq \epsilon$, Ellipsoid Method needs

$$K_{\mathrm{Ell}}(\epsilon) = O\left(n^2 \ln \frac{RMD}{r\epsilon}\right)$$

iterations, where

- M is the Lipschitz constant of f.
- D is the diameter of Q.
- r is the inner radius of Q (largest of radii of Euclidean balls $\subseteq Q$).

Comparison with Subgradient Method

Suppose that

$$Q = B(x_0, R).$$

Then, we obtain the following estimate:

$$K_{\mathrm{Ell}}(\epsilon) = O\Big(n^2 \ln \frac{2MR}{\epsilon}\Big)$$

Cf: Subgradient Method (π_Q is the Euclidean projection on Q)

$$x_{k+1} := \pi_Q(x_k - h_k g_k), \qquad k \geq 0,$$

where $h_k := 2R/(\|g_k\|\sqrt{k+1})$, has the "dimension-independent" bound:

$$K_{\mathrm{Subgr}}(\epsilon) = O\Big(\frac{M^2 R^2}{\epsilon^2}\Big).$$

Note: $K_{\mathrm{Ell}}(\epsilon) \ll K_{\mathrm{Subgr}}(\epsilon) \iff n \ll \frac{MR}{\epsilon}$.

Note: $K_{\mathrm{Ell}}(\epsilon) \to \infty$ when $n \to \infty$.

Main Issue

Recall the iteration of the Ellipsoid Method:

$$x_{k+1} = x_k - \frac{1}{n+1} \frac{H_k g_k}{\langle g_k, H_k g_k \rangle^{1/2}},$$
$$H_{k+1} = \frac{n^2}{n^2 - 1} \Big(H_k - \frac{2}{n+1} \frac{H_k g_k g_k^T H_k}{\langle g_k, H_k g_k \rangle} \Big)$$

When $n \to \infty$, we obtain:

$$x_{k+1} = x_k, \qquad H_{k+1} = H_k.$$

$\Longrightarrow$ No convergence.

Can we improve the Ellipsoid Method?
(Make it at least as good as the Subgradient Method
while retaining the original guarantee of the Ellipsoid Method.)

Subgradient Ellipsoid Method: General Scheme

1. Choose $x_0 \in \mathbb{R}^n$ and $R > 0$ such that $Q \subseteq B(x_0, R)$.
2. Define functions $\ell_0(x) := 0$, $\omega_0(x) := \frac{1}{2}\|x - x_0\|^2$.
3. Iterate for $k \geq 0$:
   1. Query the oracle to obtain $g_k := \mathcal{G}(x_k)$.
   2. Compute $U_k := \max_{x \in \Omega_k \cap L_k^-} \langle g_k, x_k - x \rangle$, where
   $$\Omega_k := \{x \in \mathbb{R}^n : \omega_k(x) \leq \tfrac{1}{2}R^2\}, \qquad L_k^- := \{x \in \mathbb{R}^n : \ell_k(x) \leq 0\}.$$
   3. Choose coefficients $a_k, b_k \geq 0$ and update functions
   $$\ell_{k+1}(x) := \ell_k(x) + a_k \langle g_k, x - x_k \rangle,$$
   $$\omega_{k+1}(x) := \omega_k(x) + \tfrac{1}{2} b_k (U_k - \langle g_k, x_k - x \rangle) \langle g_k, x - x_k \rangle.$$
   4. Set $x_{k+1} := \operatorname{argmin}_{x \in \mathbb{R}^n} [\ell_{k+1}(x) + \omega_{k+1}(x)]$.

Note: Ω_k is an ellipsoid, L_k^- is a halfspace.

Explicit Formulas

By definitions:

- $\ell_k(x) = \sum_{i=0}^{k-1} a_i \langle g_i, x - x_i \rangle$.
- $\omega_k(x) = \frac{1}{2}\|x - x_0\|^2 + \frac{1}{2}\sum_{i=0}^{k-1} b_i (U_i - \langle g_i, x_i - x \rangle)\langle g_i, x - x_i \rangle$.
- $x_k = \operatorname{argmin}_{x \in \mathbb{R}^n} [\psi_k(x) := \ell_k(x) + \omega_k(x)]$.

Note that:

- ψ_k is a quadratic function with Hessian G_k:

$$G_0 = I, \qquad G_{k+1} = G_k + b_k g_k g_k^T, \qquad k \geq 0.$$

- We can maintain their inverses $H_k := G_k^{-1}$:

$$H_0 = I, \qquad H_{k+1} = H_k - \frac{b_k H_k g_k g_k^T H_k}{1 + b_k \langle g_k, H_k g_k \rangle}, \qquad k \geq 0.$$

- Then:

$$x_{k+1} = x_k - \frac{a_k + \frac{1}{2} b_k U_k}{1 + b_k \langle g_k, H_k g_k \rangle} H_k g_k, \qquad k \geq 0.$$

Cutting Plane Property

Recall that

$$\Omega_k := \{x \in \mathbb{R}^n : \omega_k(x) \le \tfrac{1}{2}R^2\}, \qquad L_k^- := \{x \in \mathbb{R}^n : \ell_k(x) \le 0\},$$

where $\omega_0(x) := \frac{1}{2}\|x - x_0\|^2$, $\ell_0(x) := 0$, and, for any $k \ge 0$,

$$\ell_{k+1}(x) := \ell_k(x) + a_k \langle g_k, x - x_k \rangle,$$

$$\omega_{k+1}(x) := \omega_k(x) + \tfrac{1}{2} b_k (U_k - \langle g_k, x_k - x \rangle) \langle g_k, x - x_k \rangle$$

with $U_k := \max_{x \in \Omega_k \cap L_k^-} \langle g_k, x_k - x \rangle$.

Cutting plane property: For all $k \ge 0$, we have

- $x^* \in \Omega_k \cap L_k^-$.
- $\{x \in \Omega_k \cap L_k^- : \langle g_k, x - x_k \rangle \le 0\} \subseteq \Omega_{k+1} \cap L_{k+1}^-$.

Explicit Representation of Ω_k

Lemma. For all $k \geq 0$, we have

$$\Omega_k = \{x \in \mathbb{R}^n : -\ell_k(x) + \tfrac{1}{2}\|x - x_k\|_{G_k}^2 \leq \tfrac{1}{2} R_k^2\},$$

where

$$R_0 := R, \qquad R_{k+1}^2 = R_k^2 + (a_k + \tfrac{1}{2} b_k U_k)^2 \frac{(\|g_k\|_{G_k}^*)^2}{1 + b_k(\|g_k\|_{G_k}^*)^2}, \quad k \geq 0.$$

Consequences:

1. $\Omega_k \cap L_k^- \subseteq \tilde{\Omega}_k := \{x \in \mathbb{R}^n : \|x - x_k\|_{G_k} \leq R_k\}$.
2. $-\ell_k(x) \leq \frac{1}{2} R_k^2$ for all $x \in \Omega_k$.

Sliding gap: For any $k \geq 0$, such that $\Gamma_k := \sum_{i=0}^{k-1} a_i \|g_i\| > 0$, define

$$\Delta_k := \max_{x \in \Omega_k} \frac{1}{\Gamma_k}[-\ell_k(x)] \equiv \max_{x \in \Omega_k} \frac{1}{\Gamma_k} \sum_{i=0}^{k-1} a_i \langle g_i, x_i - x \rangle \leq \frac{R_k^2}{2\Gamma_k}.$$

Note: By appropriately choosing a_k and b_k, we can ensure that $\Delta_k \to 0$.

Gap and Functional Residual

Gap: Given $\lambda \coloneqq (\lambda_0, \dots, \lambda_{k-1}) \geq 0$ with $\Gamma_k(\lambda) \coloneqq \sum_{i=0}^{k-1} \lambda_i \|g_i\| > 0$, set

$$\delta_k(\lambda) \coloneqq \max_{x \in \Omega_0} \frac{1}{\Gamma_k(\lambda)} \sum_{i=0}^{k-1} \lambda_i \langle g_i, x_i - x \rangle.$$

Recall that $\Omega_0 = B(x_0, R) \supseteq Q$.

Main result: If $\delta_k \coloneqq \delta_k(\lambda) < r$ for some λ, then the approximate solution $x_k^* \coloneqq \operatorname{argmin}\{f(x) : x \in \{x_0, \dots, x_{k-1}\} \cap Q\}$ is well-defined, and

$$f(x_k^*) - f^* \leq \frac{\delta_k}{r} MD,$$

Note: We will see that $\delta_k(\lambda) \leq \Delta_k$ for some λ (explained later).

Bounding Sliding Gap

Recall:

$$\Delta_k \leq \frac{R_k^2}{2\Gamma_k}.$$

Let us choose

$$a_k := \frac{\alpha_k R + \frac{1}{2}\theta\gamma R_k}{\|g_k\|_{G_k}^*}, \qquad b_k := \frac{\gamma}{(\|g_k\|_{G_k}^*)^2}, \qquad k \geq 0,$$

where $\alpha_k, \theta, \gamma \geq 0$ (to be chosen later).

Lemma. For any $k \geq 0$ and any $\tau > 0$, we have

$$R_k^2 \leq q^k C_k R^2, \qquad \Gamma_k \geq R\Big(\sum_{i=0}^{k-1} \alpha_i + \frac{1}{2}\theta\sqrt{\gamma n[(1+\gamma)^{k/n} - 1]}\Big),$$

where

$$q := 1 + \frac{c\gamma^2}{2(1+\gamma)}, \quad c := \frac{1}{2}(\tau+1)(\theta+1)^2, \quad C_k := 1 + \frac{\tau+1}{\tau}\sum_{i=0}^{k-1}\alpha_i^2.$$

Subgradient Method

$$\alpha_k > 0, \qquad \theta := 0, \qquad \gamma := 0.$$

Then:

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \Big[\frac{1}{2}\|x - x_0\|^2 + \sum_{i=0}^{k-1} a_i \langle g_i, x - x_i \rangle \Big] = x_k - a_k g_k.$$

In this case, $\Omega_k = \Omega_0 = B(x_0, R)$ (no sliding), and

$$\delta_k = \Delta_k \leq \frac{1 + \sum_{i=0}^{k-1} \alpha_i^2}{2 \sum_{i=0}^{k-1} \alpha_i} R.$$

- **Constant step size**: Fix $k \geq 1$ and set

$$\alpha_i := \frac{1}{\sqrt{k}} \quad (0 \leq i \leq k-1) \qquad \Longrightarrow \qquad \Delta_k \leq \frac{R}{\sqrt{k}}.$$

- **Time-varying step size**:

$$\alpha_k := \frac{1}{\sqrt{k+1}} \qquad \Longrightarrow \qquad \Delta_k \leq \frac{2 + \ln k}{2\sqrt{k}} R.$$

(Standard) Ellipsoid Method

$$\alpha_k := 0, \qquad \theta := 0, \qquad \gamma := \frac{2}{n-1}.$$

In this method:

- $a_k = 0$ for all $k \geq 0$.
- $\ell_k(x) \equiv \sum_{i=0}^{k-1} a_i \langle g_i, x - x_i \rangle = 0$.
- $L_k^- = \mathbb{R}^n$ and $\Omega_k \cap L_k^- = \Omega_k = \{x : \|x - x_k\|_{G_k} \leq R_k\}$.
- $\Gamma_k := \sum_{i=0}^{k-1} a_i \|g_i\| = 0 \implies \Delta_k := \frac{1}{\Gamma_k} \max_{x \in \Omega_k} [-\ell_k(x)]$ undefined.
- Substitute: **average radius**

$$\rho_k := \left[\frac{\operatorname{vol} \Omega_k}{\operatorname{vol} B(0,1)}\right]^{1/n} = R_k [\det G_k]^{-1/(2n)} \leq \exp\bigl(-k/(2n^2)\bigr) R.$$

Can be shown that $f(x_k^*) - f^* \leq \frac{\rho_k}{r} MD$.

Ellipsoid Method with Preliminary Certificate

$$\alpha_k := 0, \qquad \theta := \sqrt{2} - 1 \ (\approx 0.41), \qquad \gamma := \gamma_1(2n) \in \Big[\frac{1}{2n}, \frac{1}{n}\Big].$$

In this method:

- $a_k > 0$, hence the sliding gap Δ_k is well-defined.
- Rate:

$$\Delta_k \leq 6 \exp\big(-k/(8n^2)\big) R.$$

Subgradient Ellipsoid Method

$$\alpha_k := \beta_k \sqrt{\frac{\theta}{\theta+1}}, \quad \theta := \sqrt[3]{2} - 1 \ (\approx 0.26), \quad \gamma := \gamma_1(2n) \in \left[\frac{1}{2n}, \frac{1}{n}\right].$$

Then:

$$\Delta_k \leq \begin{cases} 2(\sum_{i=0}^{k-1} \beta_i)^{-1}(1 + \sum_{i=0}^{k-1} \beta_i^2)R, & \text{if } k \leq n^2, \\ 6\exp\bigl(-k/(8n^2)\bigr)(1 + \sum_{i=0}^{k-1} \beta_i^2)R, & \text{if } k \geq n^2. \end{cases}$$

- **Constant step size**: Fix $k \geq 1$ and set

$$\beta_i := \frac{1}{\sqrt{k}} \ (0 \leq i \leq k-1) \implies \Delta_k \leq \begin{cases} 4R/\sqrt{k}, & \text{if } k \leq n^2, \\ 12R\exp\bigl(-k/(8n^2)\bigr), & \text{if } k \geq n^2. \end{cases}$$

- **Time-varying step size**:

$$\beta_k := \frac{1}{\sqrt{k+1}} \implies \Delta_k \leq \begin{cases} 2(2 + \ln k)R/\sqrt{k}, & \text{if } k \leq n^2, \\ 6(2 + \ln k)R\exp\bigl(-k/(8n^2)\bigr), & \text{if } k \geq n^2. \end{cases}$$

Discussion

Rate of the Subgradient Ellipsoid Method:

$$\Delta_k \leq \begin{cases} 4R/\sqrt{k}, & \text{if } k \leq n^2, \\ 12R \exp\bigl(-k/(8n^2)\bigr), & \text{if } k \geq n^2. \end{cases}$$

Rates of the Subgradient and Ellipsoid methods:

- $\Delta_k^{\mathrm{Subgr}} := R/\sqrt{k}$.
- $\rho_k^{\mathrm{Ell}} := R \exp\bigl(-k/(2n^2)\bigr)$.

Note: $\Delta_k^{\mathrm{Subgr}} \leq \rho_k^{\mathrm{Ell}} \iff k \leq K_0$, where $n^2 \ln(2n) \leq K_0 \leq 3n^2 \ln(2n)$.

Conslusion: $\Delta_k \lesssim \min\{\Delta_k^{\mathrm{Subgr}}, \rho_k^{\mathrm{Ell}}\}$.

From Sliding Gap to (Usual) Gap

- **Sliding gap**:

$$\Delta_k := \max_{x \in \Omega_k} \frac{1}{\Gamma_k} \sum_{i=0}^{k-1} a_i \langle g_i, x_i - x \rangle, \qquad \Gamma_k := \sum_{i=0}^{k-1} a_i \|g_i\|.$$

- **Gap** (for a certificate $\lambda := (\lambda_0, \ldots, \lambda_{k-1}) \geq 0$):

$$\delta_k(\lambda) := \max_{x \in \Omega_0} \frac{1}{\Gamma_k(\lambda)} \sum_{i=0}^{k-1} \lambda_i \langle g_i, x_i - x \rangle, \qquad \Gamma_k(\lambda) := \sum_{i=0}^{k-1} \lambda_i \|g_i\|.$$

Main result: There exists $\mu := (\mu_0, \ldots, \mu_{k-1}) \geq 0$ such that

$$\delta_k(a + \mu) \leq \Delta_k.$$

Note: μ can be efficiently computed (next slide).

Computing Accuracy Certificate

Recall the **cutting plane property**: for $Q_k := \Omega_k \cap L_k^-$, we have

$$\hat{Q}_k := \{x \in Q_k : \langle g_k, x - x_k \rangle \leq 0\} \subseteq Q_{k+1}.$$

Dual multiplier: For any $s \in \mathbb{R}^n$, we can find $\mu_i := \mu_i(s) \geq 0$ such that

$$\max_{x \in \hat{Q}_i} \langle s, x \rangle = \max_{x \in Q_i} [\langle s, x \rangle + \mu_i \langle g_i, x_i - x \rangle].$$

Note: μ_i can be efficiently computed in $O(n^2)$ operations.

Augmentation Algorithm [$\sim$ Nemirovski, Onn, Rothblum, 2010]

1. Set $s_k := -\sum_{i=0}^{k-1} a_i g_i$.
2. Iterate for $i = k-1, \ldots, 0$:
    1. Compute $\mu_i := \mu_i(s)$.
    2. Set $s_i := s_{i+1} - \mu_i g_i$.

Total cost: $O(kn^2)$.

Conclusions

- The standard Ellipsoid Method has an "incorrect" dependency on n.
- We have proposed a new version which is more robust w.r.t. n.
- It can be seen as a combination of:
  - "Dimension-dependent" Ellipsoid Method.
  - "Dimension-independent" Subgradient Method.
- Can be extended to more general problems with "convex structure" (primal-dual problems, saddle-point problems, variational inequalities).

Open questions:

- Get rid of extra $\ln k$ for time-varying step sizes?

$$\text{(Dual Averaging?)} \quad x_k = \operatorname*{argmin}_x \Big[\frac{\beta_k}{2} \|x - x_0\|^2 + \sum_{i=0}^{k-1} a_i \langle g_i, x - x_i \rangle \Big].$$

- Continuous (monotone) convergence rate estimate?
- Other combinations of methods?

Paper

A. Rodomanov and Y. Nesterov.
Subgradient Ellipsoid Method for Nonsmooth Convex Problems.
ArXiv:2106.13340.

Thank you!