Lecture 2: Universal Gradient Methods

Yurii Nesterov, CORE/INMA (UCL)

June 18, 2013 (Senege)

5th Traditional School on Control, Information, and Optimization

Outline

Smooth convex functions

- Gradient represents a first-order model of the objective:

$$f(x) + \langle \nabla f(x), h \rangle \leq f(x+h) \leq f(x) + \langle \nabla f(x), h \rangle + o(\|h\|).$$

- For $f \in C^{1,1}$, we can ensure monotonic decrease of the objective:

$$\begin{aligned} x_+ &= \arg\min_{y \in Q} \{ f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{2} L \|y - x\|^2 \}, \\ f(x_+) &\leq \min_{y \in Q} \{ f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{2} L \|y - x\|^2 \}. \end{aligned}$$

- At unconstrained optimum, the gradient vanishes.

  Consequently, in the gradient method $x_+ = x - h \nabla f(x)$, the stepsize $h > 0$ can be constant.

Nonsmooth convex functions

- Subgradient represents a zero-order model of the objective:

$$f(x) + \langle \nabla f(x), h \rangle \leq f(x + h) \leq f(x) + \langle \nabla f(x), h \rangle + O(\|h\|).$$

- For $f \in C^{0,0}$, we cannot ensure monotonicity.
- At unconstrained optimum, the gradient does not vanish.
- The most useful property of subgradient is

$$\langle \nabla f(x), x - x^* \rangle \geq 0,$$

where x^* is the optimal solution.

Optimization methods

Smooth functions ($f \in C^{1,1}$):

- Primal gradient method: $x_{k+1} = \pi_Q(x_k - \frac{1}{L}\nabla f(x_k))$.
- Dual gradient methods: $y_k = \pi_Q(x_k - \frac{1}{L}\nabla f(x_k))$,

$$x_{k+1} = \arg\min_{x \in Q} \left\{ \sum_{i=0}^{k} \langle \nabla f(x_k), x - x_i \rangle + \frac{1}{2} L \|x - x_0\|^2 \right\}.$$

(Both are not optimal.)

Nonsmooth functions ($f \in C^{0,0}$). Primal subgradient schemes:

- $x_{k+1} = \pi_Q(x_k - h_k \nabla f(x_k))$, $h_k > 0$, $h_k \to 0$, $\sum\limits_{k=0}^{\infty} h_k = \infty$.
- $x_{k+1} = \pi_Q\left(x_k - \frac{f(x_k) - f^*}{\|\nabla f(x_k)\|^2} \nabla f(x_k)\right)$.

(Both are optimal.)

Intermediate problem classes

For finite-dimensional linear vector space E, define a norm $\|\cdot\|$.
Then in the dual space E^*, we have $\|g\|_* \stackrel{\text{def}}{=} \max_{\|x\| \leq 1} \langle g, x \rangle$.

Hölder continuity of the gradients: for some $\nu \in [0, 1]$ and all $x, y \in Q$ we have

$$\|\nabla f(x) - \nabla f(y)\|_* \leq M_\nu(f) \|x - y\|^\nu.$$

Notation: $f \in C^{1,\nu}(Q)$.

Main property: $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M_\nu}{1+\nu} \|x - y\|^{1+\nu}$ for all $x, y \in Q$.

Proof: Denote $h = y - x$. Then

$$f(y) - f(x) - \langle \nabla f(x), h \rangle = \int_0^1 \langle \nabla f(x + \tau h) - \nabla f(x), h \rangle d\tau$$

$$\leq \|h\| \int_0^1 \|\nabla f(x + \tau h) - \nabla f(x)\|_* d\tau \leq M_\nu \|h\|^{1+\nu} \int_0^1 \tau^\nu d\tau.$$

Examples

1. $\nu = 1$: functions with Lipschitz-continuous gradients. If $f \in C^2$, and the metric is Euclidean, then

$$\nabla^2 f(x) \preceq M_1(f) I, \quad x \in Q.$$

2. $\nu = 0$: functions with bounded variation of subgradients:

$$\|\nabla f(x) - \nabla f(y)\|_* \leq M_0(f) I, \quad x \in Q.$$

NB: Addition of linear function does not change the constant $M_0(f)$.

3. Functions with $\nu \in (0, 1)$ are often obtained by duality.

Uniformly convex functions

Def: Let $f(x) \in C^1$. It is *p-uniformly convex of degree* $p \geq 2$ if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{p}\sigma_p \|y - x\|^p \text{ for all } x, y \in E,$$

where $\sigma_p = \sigma_p(f)$ is the parameter of uniform convexity.

Adding such f to a convex function does not change the parameter. If $p = 2$, then f is *strongly convex*.

Lemma 1. *Let* $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \sigma \|x - y\|^p$, $\forall x, y \in E$. *Then function* f *is p-uniformly convex on* E *with parameters* σ.

Proof.

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle = \int_0^1 \langle f(x + \tau(y - x)) - \nabla f(x), y - x \rangle d\tau$$

$$= \int_0^1 \frac{1}{\tau} \langle f(x + \tau(y - x)) - \nabla f(x), \tau(y - x) \rangle d\tau$$

$$\geq \int_0^1 \sigma \tau^{p-1} \|y - x\|^p d\tau = \frac{1}{p}\sigma \|y - x\|^p.$$

Duality

For $f(x) \in C^1$ define its *Fenchel dual*: $f_*(s) = \sup\limits_{x \in E}[\langle s, x\rangle - f(x)]$.

NB: $\nabla f_*(s) = x_f(s) = \arg\min\limits_{x \in E}[\langle s, x\rangle - f(x)], \quad \nabla f(x_f(s)) = s.$

Lemma 2. If f is p-uniformly convex, then $f_* \in C^{1,\nu}$ with

$$\nu = \frac{1}{p-1}, \quad M_\nu(f_*) = \left(\frac{p}{2\sigma_p}\right)^{\frac{1}{p-1}}.$$

Proof. For two points s_1 and s_2, denote $x_i = x_f(s_i)$. Then

$$f(x_{3-i}) \geq f(x_i) + \langle \nabla f(x_i), x_{3-i} - x_i\rangle + \frac{1}{p}\sigma_p \|x_{3-i} - x_i\|^p, \ i = 1, 2.$$

Adding these inequalities, we get

$$\frac{2}{p}\sigma_p \|x_1 - x_2\|^p \leq \langle s_1 - s_2, x_1 - x_2\rangle \leq \|s_1 - s_2\|_* \|x_1 - x_2\|. \qquad \square$$

Example

1. Consider $f(\tau) = \frac{1}{3}|\tau|^3$, $\tau \in R$. Then $\nabla f(\tau) = \tau|\tau|$. Note that

$$\begin{aligned}(\nabla f(\tau_1) - \nabla f(\tau_2))(\tau_1 - \tau_2) &= |\tau_1|\tau_1| - \tau_2|\tau_2|| \cdot |\tau_1 - \tau_2| \\ &\geq \frac{1}{2}|\tau_1 - \tau_2|^3.\end{aligned}$$

Hence, $f_*(\xi) = \max_\tau \left[\xi\tau - \frac{1}{3}|\tau|^3\right] = \frac{2}{3}|\xi|^{\frac{3}{2}} \in C^{1,1/2}$, and

$$M_{1/2} = \left[\frac{3}{2 \cdot \frac{1}{2}}\right]^{1/2} = \sqrt{3}.$$

2. Consider $F(x) = \frac{1}{3}\sum\limits_{i=1}^n \alpha_i |x^{(i)}|^3$. Then for $\|h\|_\alpha^3 \stackrel{\text{def}}{=} \sum\limits_{i=1}^n \alpha_i |h^{(i)}|^3$

$$\langle \nabla F(x) - \nabla F(y), x - y \rangle \geq \frac{1}{2}\|x - y\|_\alpha^3 \quad (\alpha > 0).$$

Therefore the dual function $F_*(s) = \frac{2}{3}\sum\limits_{i=1}^n \frac{1}{\sqrt{\alpha_i}}|s^{(i)}|^{3/2}$ is in $C^{1,1/2}$

with $M_{1/2} = \sqrt{3}$. Note that $\|s\|_\alpha^* = \left[\sum\limits_{i=1}^n \frac{1}{\sqrt{\alpha_i}}|s^{(i)}|^{3/2}\right]^{2/3}$ (Check!)

Application Example: Gas Network

Given:

- Structure of pipe lines.
- Length and diameter of each pipe.
- Positions and required volumes for sources and sinks.

Goal: Compute the flows in the pipes and pressure at the nodes.

Equilibrium principle: the flows minimize the dispersed energy.

$$\min_{f \in R^n} \left\{ \frac{1}{3} \sum_{i=1}^{n} \alpha_i |f_i|^3 : \; Af = d \right\}.$$

Duality:

$$\min_{f \in R^n} \max_{y \in R^m} \left\{ \frac{1}{3} \sum_{i=1}^{n} \alpha_i |f_i|^3 + \langle y, d - Af \rangle \right\}$$

$$= \max_{y \in R^m} \min_{f \in R^n} \left\{ \frac{1}{3} \sum_{i=1}^{n} \alpha_i |f_i|^3 - \langle A^T y, f \rangle + \langle y, d \rangle \right\}$$

$$= \max_{y \in R^m} \left\{ \langle d, y \rangle - \frac{2}{3} \left(\|A^T y\|_\alpha^* \right)^{3/2} \right\}. \quad \text{(Dual objective is in } C^{1,1/2}.\text{)}$$

Structure of Holder constants

Define $M_\nu \equiv M_\nu(f) = \sup\limits_{\substack{x,y\in Q,\\ x\neq y}} \frac{\|\nabla f(x)-\nabla f(y)\|_*}{\|x-y\|^\nu}, \quad \nu \geq 0.$

Since $\ln M_\nu = \sup\limits_{\substack{x,y\in Q,\\ x\neq y}} [\, \ln \|\nabla f(x) - \nabla f(y)\|_* - \nu \ln \|x - y\| \,],$

M_ν is a *log-convex* function of ν.

- For certain $\nu \in [0, 1]$, M_ν can be infinite.
- If M_0 and M_1 are finite, then $M_\nu \leq M_0^{1-\nu} M_1^\nu$, $0 \leq \nu \leq 1$.
- If $M_\nu < \infty$, then $\|\nabla f(x) - \nabla f(y)\|_* \leq M_\nu \|x - y\|^\nu$, $x, y \in Q$.

Therefore,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M_\nu}{1+\nu} \|x - y\|^{1+\nu}, \quad x, y \in Q.$$

Assumption: $\hat{M}(f) \stackrel{\text{def}}{=} \inf\limits_{0\leq\nu\leq 1} M_\nu(f) < +\infty.$

Composite Minimization and Bregman distances

Problem: $\min\limits_{x \in Q} \left[\tilde{f}(x) \stackrel{\text{def}}{=} f(x) + \Psi(x) \right]$, where

- Q is a simple closed convex set,
- Ψ is a *simple* closed convex function (e.g. squared Euclidean norm, l_1-norm, barrier functions, indicator of convex set, etc.).
- f is assumed to be subdifferentiable on Q.

Prox-function $d(x)$**:** a differentiable strongly convex function:
$d(y) \geq d(x) + \langle \nabla d(x), y - x \rangle + \frac{1}{2}\|x - y\|^2, \quad x, y \in \operatorname{rint} Q.$

Let $d(x)$ attains its minimum on Q at x_0, and $d(x_0) = 0$.
Thus, $d(x) \geq \frac{1}{2}\|x - x_0\|^2, \quad x \in Q.$

Prox-function defines the *Bregman distance*:

$$\xi(x, y) \stackrel{\text{def}}{=} d(y) - d(x) - \langle \nabla d(x), y - x \rangle.$$

Clearly, $\xi(x, x) \equiv 0$, and $\xi(x, y) \geq \frac{1}{2}\|x - y\|^2, \quad x, y \in Q.$

Bregman Mapping

For any $x \in Q$ we can define the *Bregman mapping* $\mathcal{B}_M(x) =$

$$\arg\min_{y \in Q} \left\{ \psi_M(x, y) \stackrel{\text{def}}{=} f(x) + \langle \nabla f(x), y - x \rangle + M\xi(x, y) + \Psi(y) \right\}.$$

Assumption: This point is easily computable.

First-order optimality condition for the auxiliary optimization problem: $\forall y \in Q$

$$\langle \nabla f(x) + M(\nabla d(\mathcal{B}_M(x)) - \nabla d(x)) + \nabla \Psi(\mathcal{B}_M(x)), y - \mathcal{B}_M(x) \rangle \geq 0.$$

Denote $\psi_M^*(x) = \psi_M(x, \mathcal{B}_M(x))$.

(To be compared with $\tilde{f}(\mathcal{B}_M(x))$.)

Main Lemma

Lemma: If $M \geq \left[\frac{1}{\delta}\right]^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}$ with $\delta > 0$, then for $x, y \in Q$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{2} M \|y - x\|^2 + \tfrac{\delta}{2}.$$

Therefore, $\tilde{f}(\mathcal{B}_M(x)) \leq \psi_M^*(x) + \frac{\delta}{2}$.

Proof: For $\tau, s > 0$ we have $\frac{1}{p}\tau^p + \frac{1}{q}s^q \geq \tau s$, with $\frac{1}{p} + \frac{1}{q} = 1$.

Taking $p = \frac{2}{1+\nu}$, $q = \frac{2}{1-\nu}$, and $\tau = t^{1+\nu}$, we get

$$t^{1+\nu} \leq \tfrac{1+\nu}{2s} t^2 + \tfrac{1-\nu}{2} s^{\frac{1+\nu}{1-\nu}}.$$

Denote $\delta = \frac{1-\nu}{1+\nu} M_\nu s^{\frac{1+\nu}{1-\nu}}$. Then $s = \left[\frac{1+\nu}{1-\nu} \cdot \frac{\delta}{M_\nu}\right]^{\frac{1-\nu}{1+\nu}}$. Therefore,

$$\tfrac{M_\nu}{1+\nu} t^{1+\nu} \leq \tfrac{1}{2s} M_\nu t^2 + \tfrac{\delta}{2} = \tfrac{1}{2}\left[\tfrac{1-\nu}{1+\nu} \cdot \tfrac{1}{\delta}\right]^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} t^2 + \tfrac{\delta}{2} \leq \tfrac{1}{2} M t^2 + \tfrac{\delta}{2}.$$

Proof (continued)

Further, denoting $x_+ = \mathcal{B}_M(x)$, we obtain:

$$\begin{aligned} f(x_+) &\leq f(x) + \langle \nabla f(x), x_+ - x \rangle + \frac{M_\nu}{1+\nu} \|x_+ - x\|^{1+\nu} \\ &\leq f(x) + \langle \nabla f(x), x_+ - x \rangle + \frac{M}{2} \|x_+ - x\|^2 + \frac{\delta}{2} \\ &\leq f(x) + \langle \nabla f(x), x_+ - x \rangle + M\xi(x, x_+) + \frac{\delta}{2}. \end{aligned}$$

Therefore, $\tilde{f}(x_+) = f(x_+) + \Psi(x_+) \leq \psi_M^*(x) + \frac{\delta}{2}$. □

Universal Primal Gradient Method (PGM)

Initialization. Choose $L_0 > 0$ and accuracy $\epsilon > 0$.

For $k \geq 0$ **do:**

1. Find the smallest $i_k \geq 0$ such that

$$\tilde{f}\left(\mathcal{B}_{2^{i_k}L_k}(x_k)\right) \leq \psi^*_{2^{i_k}L_k}(x_k) + \tfrac{1}{2}\epsilon.$$

2. Set $x_{k+1} = \mathcal{B}_{2^{i_k}L_k}(x_k)$, and $L_{k+1} = 2^{i_k-1}L_k$.

PGM: convergence

Denote $\gamma(M,\epsilon) \stackrel{\text{def}}{=} \left[\frac{1}{\epsilon}\right]^{\frac{1-\nu}{1+\nu}} M^{\frac{2}{1+\nu}}$, and

$$S_k = \sum_{i=1}^{k+1} \frac{1}{L_k}, \quad \tilde{f}_k^* = \frac{1}{S_k} \sum_{i=0}^{k} \frac{1}{L_{i+1}} \tilde{f}(x_i).$$

Theorem: Let $M_\nu(f) < \infty$ and $L_0 \le \gamma(M_\nu, \epsilon)$.

Then for all $k \ge 0$ we have $L_{k+1} \le \gamma(M_\nu, \epsilon)$. Moreover, for all $y \in Q$

$$\tilde{f}_k^* \le \frac{1}{S_k} \sum_{i=0}^{k} \frac{1}{L_{i+1}} [f(x_i) + \langle \nabla f(x_i), y - x_i \rangle] + \Psi(y) + \frac{\epsilon}{2} + \frac{2}{S_k} \xi(x_0, y).$$

Therefore, $\tilde{f}_k^* - \tilde{f}(x^*) \le \frac{\epsilon}{2} + \frac{2\gamma(M_\nu,\epsilon)}{k+1} \xi(x_0, x^*)$.

Proof, page 1

Let us fix $y \in Q$. Denote $r_k(y) \stackrel{\text{def}}{=} \xi(x_k, y)$. Then (by FOOC)

$$\begin{aligned} r_{k+1}(y) &= d(y) - d(x_{k+1}) - \langle \nabla d(x_{k+1}), y - x_{k+1} \rangle \\ &\leq d(y) - d(x_{k+1}) - \langle \nabla d(x_k), y - x_{k+1} \rangle \\ &\quad + \tfrac{1}{2L_{k+1}} \langle \nabla f(x_k) + \nabla \Psi(x_{k+1}), y - x_{k+1} \rangle. \end{aligned}$$

Note that

$$\begin{aligned} & d(y) - d(x_{k+1}) - \langle \nabla d(x_k), y - x_{k+1} \rangle \\ = \; & d(y) - d(x_k) - \langle \nabla d(x_k), x_{k+1} - x_k \rangle - \xi(x_k, x_{k+1}) \\ -\langle \nabla d(x_k), y - x_{k+1} \rangle \; &= \; r_k(y) - \xi(x_k, x_{k+1}). \end{aligned}$$

Proof, page 2

Thus, $r_{k+1}(y) - r_k(y) \leq$

$$\frac{1}{2L_{k+1}}\langle \nabla f(x_k) + \nabla \Psi(x_{k+1}), y - x_{k+1}\rangle - \xi(x_k, x_{k+1})$$

$$= \frac{1}{2L_{k+1}}\langle \nabla \Psi(x_{k+1}), y - x_{k+1}\rangle - \frac{1}{2L_{k+1}}\Big(\langle \nabla f(x_k), x_{k+1} - x_k\rangle + 2L_{k+1}\xi(x_k, x_{k+1})\Big) + \frac{1}{2L_{k+1}}\langle \nabla f(x_k), y - x_k\rangle$$

$$\leq \frac{1}{2L_{k+1}}\left(\Psi(y) - \Psi(x_{k+1}) + f(x_k) - f(x_{k+1}) + \frac{\epsilon}{2} + \langle \nabla f(x_k), y - x_k\rangle\right).$$

Hence, $\frac{1}{2L_{k+1}}\tilde{f}(x_{k+1}) + r_{k+1}(y)$

$$\leq \frac{1}{2L_{k+1}}\left(f(x_k) + \langle \nabla f(x_k), y - x_k\rangle + \Psi(y) + \frac{\epsilon}{2}\right) + r_k(y).$$

Summing up these inequalities, we obtain

$$\tilde{f}_k^* \leq \frac{1}{S_k}\sum_{i=0}^{k}\frac{1}{L_{i+1}}\left[f(x_i) + \langle \nabla f(x_i), y - x_i\rangle\right] + \Psi(y) + \frac{\epsilon}{2} + \frac{2}{S_k}r_0(y). \square$$

Consequences

Complexity: $\frac{\epsilon}{2} + \frac{2\gamma(M_\nu,\epsilon)}{k+1}\xi(x_0,x^*) \le \epsilon$ with $\gamma(M,\epsilon) = \left[\frac{1}{\epsilon}\right]^{\frac{1-\nu}{1+\nu}} M^{\frac{2}{1+\nu}}$. Hence, we need

$$4\xi(x_0,x^*) \inf_{0\le\nu\le 1} \left(\frac{M_\nu}{\epsilon}\right)^{\frac{2}{1+\nu}} \text{ iterations.}$$

Stopping criterion.

Assume we have a bound $\xi(x_0,x^*) \le D$.

Denote $\ell_k^p(y) \stackrel{\text{def}}{=} \frac{1}{S_k}\sum_{i=0}^{k} \frac{1}{L_{i+1}}[f(x_i) + \langle \nabla f(x_i), y - x_i\rangle]$, and define

$$\hat{f}_k = \min_{y\in Q}\left\{\ell_k^p(y) + \Psi(y) : \ \xi(x_0,y) \le D\right\}.$$

Then $\tilde{f}_k^* - \tilde{f}(x^*) \le \tilde{f}_k^* - \hat{f}_k \le \frac{2\gamma(M_\nu,\epsilon)}{k+1} D$.

Thus, we have implementable stopping criterion $\tilde{f}_k^* - \hat{f}_k \le \epsilon$.

Number of calls of oracle

Denote $N(k)$, the total number of computations of function values in PGM after k iterations. Note that

$$L_{k+1} = \tfrac{1}{2} 2^{i_k} L_k.$$

Therefore, $i_k - 1 = \log_2 \frac{L_{k+1}}{L_k}$. Hence, for any $\nu \in [0, 1]$, we have

$$\begin{aligned} N(k) &= \sum_{j=0}^{k} (i_j + 1) = 2(k+1) + \log_2 L_{k+1} - \log_2 L_0 \\ &\leq 2(k+1) + \tfrac{1-\nu}{1+\nu} \log_2 \tfrac{1}{\epsilon} + \tfrac{2}{1+\nu} \log_2 M_\nu - \log_2 L_0. \end{aligned}$$

Finally, we come to the following upper bound:

$$N(k) \leq 2(k+1) - \log_2 L_0 + \inf_{0 \leq \nu \leq 1} \left[\tfrac{1-\nu}{1+\nu} \log_2 \tfrac{1}{\epsilon} + \tfrac{2}{1+\nu} \log_2 M_\nu \right].$$

Thus in average, PGM needs two computations of function values per iteration.

Universal Dual Gradient Method (DGM)

Initialization. Choose $L_0 > 0$. Define $\phi_0(x) = \xi(x_0, x)$.

For $k \geq 0$ **do:**

1. Find the smallest $i_k \geq 0$ such that for point
$x_{k,i_k} = \arg\min_{x \in Q} \left\{ \phi_k(x) + \frac{1}{2^{i_k} L_k} [f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \Psi(x)] \right\}$
we have $\tilde{f}\left(\mathcal{B}_{2^{i_k} L_k}(x_{k,i_k}) \right) \leq \psi^*_{2^{i_k} L_k}(x_{k,i_k}) + \frac{\epsilon}{2}$.

2. Set $x_{k+1} = x_{k,i_k}$, $L_{k+1} = 2^{i_k - 1} L_k$, and
$\phi_{k+1}(x) = \phi_k(x) + \frac{1}{2L_{k+1}} [f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \Psi(x)]$.

Convergence of DGM

Theorem. For all $k \geq 0$ and $\nu \in [0, 1]$ we have

$$\tilde{f}_k^* - \tilde{f}(x^*) \leq \frac{\epsilon}{2} + \frac{2\gamma(M_\nu, \epsilon)}{k+1} \xi(x_0, x^*).$$

Complexity: $4\xi(x_0, x^*) \inf\limits_{0 \leq \nu \leq 1} \left(\frac{M_\nu}{\epsilon}\right)^{\frac{2}{1+\nu}}$ iterations.

Average # of calls: 4 per iteration.

NB: for $\nu \in (0, 1]$ the complexity is not optimal!

Universal Fast Gradient Method (FGM)

Choose $L_0 > 0$. Define $\phi_0(x) = \xi(x_0, x)$, $y_0 = x_0$, $A_0 = 0$.

For $k \geq 0$ **do:**

1. Find $v_k = \arg\min_{x \in Q} \phi_k(x)$.

2. Find the smallest $i_k \geq 0$ such that a_{k+1,i_k}, computed from equation $a^2_{k+1,i_k} = \frac{1}{2^{i_k} L_k}(A_k + a_{k+1,i_k})$ and used in the definitions

$$A_{k+1,i_k} = A_k + a_{k+1,i_k}, \quad \tau_{k,i_k} = \frac{a_{k+1,i_k}}{A_{k+1,i_k}}, \quad x_{k+1,i_k} = \tau_{k,i_k} v_k + (1 - \tau_{k,i_k}) y_k,$$

$$\hat{x}_{k+1,i_k} = \arg\min_{y \in Q} \{\xi(v_k, y) + a_{k+1,i_k}[\langle \nabla f(x_{k+1,i_k}), y \rangle + \Psi(y)]\},$$

$y_{k+1,i_k} = \tau_{k,i_k} \hat{x}_{k+1,i_k} + (1 - \tau_{k,i_k}) y_k$, ensures the following relation:

$$\begin{aligned} f(y_{k+1,i_k}) \quad \leq \quad & f(x_{k+1,i_k}) + \langle \nabla f(x_{k+1,i_k}), y_{k+1,i_k} - x_{k+1,i_k} \rangle \\ & + 2^{i_k - 1} L_k \| y_{k+1,i_k} - x_{k+1,i_k} \|^2 + \tfrac{\epsilon}{2} \tau_{k,i_k}. \end{aligned}$$

3. Set $x_{k+1} = x_{k+1,i_k}$, $y_{k+1} = y_{k+1,i_k}$, $a_{k+1} = a_{k+1,i_k}$, $\tau_k = \tau_{k,i_k}$.

Define $A_{k+1} = A_k + a_{k+1}$, $L_{k+1} = 2^{i_k - 1} L_k$, and

$$\phi_{k+1}(x) = \phi_k(x) + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1} \rangle + \Psi(x)].$$

Convergence of FGM

Theorem. For all $k \geq 0$ we have

$$A_k \left(\tilde{f}(y_k) - \tfrac{\epsilon}{2} \right) \leq \phi_k^* \stackrel{\text{def}}{=} \min_{x \in Q} \phi_k(x),$$

where $A_k \geq \left[\frac{1}{2^{2+4\nu} M_\nu^2} \, \epsilon^{1-\nu} \, k^{1+3\nu} \right]^{\frac{1}{1+\nu}}$.

Consequently, for all $k \geq 1$ we have

$$\tilde{f}(y_k) - \tilde{f}(x^*) \leq \left[\frac{2^{2+4\nu} M_\nu^2}{\epsilon^{1-\nu} \, k^{1+3\nu}} \right]^{\frac{1}{1+\nu}} \xi(x_0, x^*) + \tfrac{\epsilon}{2}.$$

Complexity: $k \leq \inf\limits_{0 \leq \nu \leq 1} \left[\left(\frac{2^{\frac{3+5\nu}{2}} M_\nu}{\epsilon} \right)^{\frac{2}{1+3\nu}} \xi(x_0, x^*)^{\frac{1+\nu}{1+3\nu}} \right]$.

It is optimal! (Note quasi-convexity in ν.)

Calls per iteration: four.

Numerical experiments

1. Matrix game: $\min\limits_{x\in\Delta_n}\max\limits_{y\in\Delta_m}\langle x, Ay\rangle$

$$= \min_{x\in\Delta_n}\left\{\psi_p(x) \stackrel{\rm def}{=} \max_{1\le j\le m}\langle x, Ae_j\rangle\right\} = \max_{y\in\Delta_m}\left\{\psi_d(y) \stackrel{\rm def}{=} \min_{1\le i\le n}\langle e_i, Ay\rangle\right\}.$$

It can be posed as a minimization problem

$$\min_{x\in\Delta_n, y\in\Delta_m}\{\psi_{pd}(x,y) = \psi_p(x) - \psi_d(y)\}$$

with optimal value zero. We generate $A_{i,j} \in [-1, 1]$ randomly.

For $\mathcal{F} = \{z = (x, y) : \ x \in \Delta_n,\ y \in \Delta_m\}$, a natural prox-function is the *entropy*:

$$\eta(z) = \sum_{i=1}^{n} z^{(i)} \ln z^{(i)}.$$

It is strongly convex in ℓ_1-norm (good for measuring simplexes).

Entropy Setup ($n = 896,\ m = 128$)

Eps	FGM$_{Entropy}$			PGM$_{Entropy}$		
2^{-5}	516	$6.0E{-}2$	$1.3E2$	722	$8.2E{-}2$	8.0
2^{-6}	1127	$2.9E{-}2$	$2.6E2$	2065	$5.2E{-}2$	$1.6E1$
2^{-7}	1937	$1.6E{-}2$	$2.0E2$	5675	$3.4E{-}2$	$3.2E1$
2^{-8}	4684	$7.9E{-}3$	$2.0E3$	15731	$2.3E{-}2$	$6.4E1$
2^{-9}	8129	$3.8E{-}3$	$8.2E3$	44829	$1.5E{-}2$	$1.3E2$
2^{-10}	17556	$2.1E{-}3$	$4.1E3$	122959	$1.0E{-}2$	$2.6E2$

FGM: $O\left(\frac{1}{\epsilon}\right)$.

PGM: $O\left(\frac{1}{\epsilon^{1.57}}\right)$.

Continuous Steiner problem ($n = 256$, $m = 512$)

$$\min_{x \in Q} f(x) \stackrel{\text{def}}{=} \sum_{i=1}^{m} \|x - a_i\|. \qquad \text{(Euclidean norms)}$$

Eps	FGM$_{Euclid}$			PGM$_{Euclid}$		
2^{-5}	205	$3.1E{-}2$	$2.6E2$	9925	$3.1E{-}2$	$2.6E2$
2^{-6}	307	$1.5E{-}2$	$5.1E2$	19895	$1.5E{-}2$	$5.1E2$
2^{-7}	277	$6.8E{-}3$	$2.6E2$	39803	$7.8E{-}3$	$2.6E2$
2^{-8}	611	$3.9E{-}3$	$5.1E2$	77138	$3.9E{-}3$	$5.1E2$
2^{-9}	827	$1.9E{-}3$	$5.1E2$	155038	$2.0E{-}3$	$2.6E2$
2^{-10}	1226	$9.8E{-}4$	$2.6E2$		out of time	
2^{-11}	1655	$4.8E{-}4$	$2.6E2$			
2^{-12}	2385	$2.4E{-}4$	$5.1E2$			
2^{-13}	3388	$1.2E{-}4$	$5.1E2$			

FGM: $O(\frac{1}{\epsilon^{1/2}})$, **PGM:** $O(\frac{1}{\epsilon})$.