An optimal randomized incremental gradient method

Guanghui (George) Lan
Joint work with Yi Zhou

Department of Industrial and Systems Engineering, University of Florida, USA

UF Risk Management Workshop 2015
November 9, 2015

The problem of interest

Problem: $\Psi^* := \min_{x \in X} \left\{ \Psi(x) := \sum_{i=1}^m f_i(x) + h(x) + \mu\, \omega(x) \right\}$.

- X closed and convex.
- f_i smooth convex: $\|\nabla f_i(x_1) - \nabla f_i(x_2)\|_* \leq L_i \|x_1 - x_2\|$.
- h simple, e.g., l_1 norm.
- ω strongly convex with modulus 1 w.r.t. an arbitrary norm.
- $\mu \geq 0$.
- Subproblem $\operatorname{argmin}_{x \in X} \langle g, x \rangle + h(x) + \mu\, \omega(x)$ is easy.
- Denote $f(x) \equiv \sum_{i=1}^m f_i(x)$ and $L \equiv \sum_{i=1}^m L_i$. f is smooth with Lipschitz constant $L_f \leq L$.

Stochastic subgradient descent for nonsmooth problems

- General stochastic programming (SP): $\min_{x \in X} \mathbb{E}_{\xi}[F(x, \xi)]$.
- Reformulation of the finite sum problem as SP:
  - $\xi \in \{1, \ldots, m\}$, $\text{Prob}\{\xi = i\} = \nu_i$, and $F(x, i) = \nu_i^{-1} f_i(x) + h(x) + \mu\omega(x)$, $i = 1, \ldots, m$.
- Iteration complexity: $\mathcal{O}(1/\epsilon^2)$ or $\mathcal{O}(1/\epsilon)$ ($\mu > 0$).
- Iteration cost: m times cheaper than deterministic first-order methods.
- Save up to a factor of $\mathcal{O}(m)$ subgradient computations.
- For details, see Nemirovski et. al. (09).

Required ∇f_i's in the smooth case

For simplicity, focus on the strongly convex case ($\mu > 0$).
Goal: find a solution $\bar{x} \in X$ s.t. $\|\bar{x} - x^*\| \leq \epsilon \|x^0 - x^*\|$.

- Nesterov's optimal method (Nesterov 83):
$$\mathcal{O}\left\{m\sqrt{\frac{L_f}{\mu}}\log\frac{1}{\epsilon}\right\},$$
- Accelerated stochastic approximation (Lan 12, Ghadimi and Lan 13):
$$\mathcal{O}\left\{\sqrt{\frac{L_f}{\mu}}\log\frac{1}{\epsilon} + \frac{\sigma^2}{\mu\epsilon}\right\}$$

Note: the optimality of the latter bound for general SP does not preclude more efficient algorithms for the finite-sum problem.

Randomized incremental gradient methods

Each iteration requires a randomly selected $\nabla f_i(x)$.

- Stochastic average gradient (SAG) by Schmidt, Roux and Bach 13:
$$\mathcal{O}\left((m + L/\mu)\log\tfrac{1}{\epsilon}\right).$$
- Similar results were obtained in Johnson and Zhang 13, Defazio et al. 14...
- Worse dependence on the L/μ than Nesterov's method.
- Intimidating proofs ...

Coordinate ascent in the dual

$\min \left\{ \sum_{i=1}^{m} \phi_i(a_i^T x) + h(x) \right\}$, h strongly convex w.r.t. l_2 norm.

All these coordinate algorithms achieve $\mathcal{O}\left\{ m + \sqrt{\frac{mL}{\mu}} \log \frac{1}{\epsilon} \right\}$.

- Shalev-Shwartz and Zhang 13, 15 (restarting stochastic dual ascent),
- Lin, Lu and Xiao, 14 (Nesterov, Fercoq and P. Richtárik's), see also Zhang and Xiao 14 (Chambolle and Pock),
- Dang and Lan 14 (non-strongly convex), $\mathcal{O}(1/\epsilon)$ or $\mathcal{O}(1/\sqrt{\epsilon})$.

Some issues:

- Deal with a more special class of problems.
- Require $\operatorname{argmin}\{\langle g, y \rangle + \phi_i^*(y) + \|y\|_*^2\}$, not incremental gradient methods.

Open problems and our research

Problems:

- Can we accelerate the convergence of randomized incremental gradient method?
- What is the best possible performance we can expect?

Our approach:

- Develop the primal-dual gradient (PDG) method and show its inherent relation to Nesterov's method.
- Develop a randomized PDG (RPDG).
- Present a new lower complexity bound.
- Provide game-theoretic interpretation for acceleration.

Reformulation and game/economic interpretation

Let J_f be the conjugate function of f. Consider

$$\Psi^* := \min_{x \in X} \left\{ h(x) + \mu\, \omega(x) + \max_{g \in \mathcal{G}} \langle x, g \rangle - J_f(g) \right\}$$

- The buyer purchases products from the supplier.
- The unit price is given by $g \in \mathbb{R}^n$.
- X, h and ω are constraints and other local cost for the buyer.
- The profit of supplier: revenue ($\langle x, g \rangle$) - local cost $J_f(g)$.

How to achieve equilibrium?

Current order quantity x^0, and product price g^0.

Proximity control functions:

$$\begin{aligned} P(x^0, x) &:= \omega(x) - [\omega(x^0) + \langle \omega'(x^0), x - x^0 \rangle]. \\ D_f(g_i^0, y_i) &:= J_f(g) - [J_f(g^0) + \langle J_f'(g^0), g - g^0 \rangle]. \end{aligned}$$

Dual prox-mapping:

$$\mathcal{M}_{\mathcal{G}}(-\tilde{x}, g^0, \tau) := \arg\min_{g \in \mathcal{G}} \left\{ \langle -\tilde{x}, g \rangle + J_f(g) + \tau D_f(g^0, g) \right\}.$$

$\tilde{x}$ is the given or predicted demand. Maximize the profit, but not too far away from g^0.

Primal prox-mapping:

$$\mathcal{M}_X(g, x^0, \eta) := \arg\min_{x \in X} \left\{ \langle g, x \rangle + h(x) + \mu\omega(x) + \eta P(x^0, x) \right\}.$$

g is the given or predicted price. Minimize the cost, but not too far way from x^0.

The deterministic PDG

Algorithm 1 The primal-dual gradient method

Let $x^0 = x^{-1} \in X$, and the nonnegative parameters $\{\tau_t\}$, $\{\eta_t\}$, and $\{\alpha_t\}$ be given.
Set $g^0 = \nabla f(x^0)$.
for $t = 1, \ldots, k$ **do**
  Update $z^t = (x^t, y^t)$ according to

$$\begin{aligned} \tilde{x}^t &= \alpha_t(x^{t-1} - x^{t-2}) + x^{t-1}. \\ g^t &= \mathcal{M}_{\mathcal{G}}(-\tilde{x}^t, g^{t-1}, \tau_t). \\ x^t &= \mathcal{M}_X(g^t, x^{t-1}, \eta_t). \end{aligned}$$

end for

A game/economic interpretation

- The supplier predicts the buyer's demand based on historical information: $\tilde{x}^t = \alpha_t(x^{t-1} - x^{t-2}) + x^{t-1}$.
- The supplier seeks to maximize predicted profit, but not too far away from g^{t-1}: $g^t = \mathcal{M}_{\mathcal{G}}(-\tilde{x}^t, g^{t-1}, \tau_t)$.
- The buyer tries to minimize the cost, but not too far away from x^{t-1}: $x^t = \mathcal{M}_X(g^t, x^{t-1}, \eta_t)$.

PDG in gradient form

Algorithm 2 PDG method in gradient form

Input: Let $x^0 = x^{-1} \in X$, and the nonnegative parameters $\{\tau_t\}$, $\{\eta_t\}$, and $\{\alpha_t\}$ be given.

Set $\underline{x}^0 = x^0$.

for $t = 1, 2, \ldots, k$ **do**

$$\begin{aligned} \tilde{x}^t &= \alpha_t(x^{t-1} - x^{t-2}) + x^{t-1}. \\ \underline{x}^t &= \left(\tilde{x}^t + \tau_t \underline{x}^{t-1}\right)/(1+\tau_t). \\ g^t &= \nabla f(\underline{x}^t). \\ x^t &= \mathcal{M}_X(g^t, x^{t-1}, \eta_t). \end{aligned}$$

end for

Idea: set $J_f'(g^{t-1}) = \underline{x}^{t-1}$.

Relation to Nesterov's method

A variant of Nesterov's method:

$$\begin{aligned} \underline{x}^t &= (1-\theta_t)\bar{x}^{t-1} + \theta_t x^{t-1}. \\ x^t &= M_X(\textstyle\sum_{i=1}^m \nabla f_i(\underline{x}^t), x^{t-1}, \eta_t). \\ \bar{x}^t &= (1-\theta_t)\bar{x}^{t-1} + \theta_t x^t. \end{aligned}$$

Note that

$$\underline{x}^t = (1-\theta_t)\underline{x}^{t-1} + (1-\theta_t)\theta_{t-1}(x^{t-1} - x^{t-2}) + \theta_t x^{t-1}.$$

Equivalent to PDG with $\tau_t = (1-\theta_t)/\theta_t$ and $\alpha_t = \theta_{t-1}(1-\theta_t)/\theta_t$.

Nesterov's acceleration: looking-ahead dual players.
Gradient descent: myopic dual players ($\alpha_t = \tau_t = 0$ in PDG).

Convergence of PDG (or Nesterov's variant)

Theorem

Define $\bar{x}^k := (\sum_{t=1}^k \theta_t)^{-1} \sum_{t=1}^k (\theta_t x^t)$. *Suppose that*

$$\tau_t = \sqrt{\frac{2L_f}{\mu}}, \quad \eta_t = \sqrt{2L_f \mu}, \quad \alpha_t = \alpha \equiv \frac{\sqrt{2L_f/\mu}}{1+\sqrt{2L_f/\mu}}, \quad \text{and} \quad \theta_t = \frac{1}{\alpha^t}.$$

Then,

$$\begin{aligned} P(x^k, x^*) &\leq \frac{\mu+L_f}{\mu} \alpha^k P(x^0, x^*). \\ \Psi(\bar{x}^k) - \Psi(x^*) &\leq \mu(1-\alpha)^{-1} \left[1 + \frac{L_f}{\mu}(2 + \frac{L_f}{\mu})\right] \alpha^k P(x^0, x^*). \end{aligned}$$

Theorem

If $\tau_t = \frac{t-1}{2}$, $\eta_t = \frac{4L_f}{t}$, $\alpha_t = \frac{t-1}{t}$, *and* $\theta_t = t$, *then*

$$\Psi(\bar{x}^k) - \Psi(x^*) \leq \frac{8L_f}{k(k+1)} P(x^0, x^*).$$

A multi-dual-player reformulation

- Let $J_i : \mathcal{Y}_i \to \mathbb{R}$ be the conjugate functions of f_i and $\mathcal{Y}_i$, $i = 1, \ldots, m$, denote the dual spaces.
  $\min_{x \in X} \left\{ h(x) + \mu\, \omega(x) + \max_{y_i \in \mathcal{Y}_i} \langle x, \sum_i y_i \rangle - \sum_i J(y) \right\}$,
- Define their new dual prox-functions and dual prox-mappings as

$$\begin{aligned} D_i(y_i^0, y_i) &:= J_i(y_i) - [J_i(y_i^0) + \langle J_i'(y_i^0), y_i - y_i^0 \rangle], \\ \mathcal{M}_{\mathcal{Y}_i}(-\tilde{x}, y_i^0, \tau) &:= \arg\min_{y_i \in \mathcal{Y}_i} \left\{ \langle -\tilde{x}, y \rangle + J_i(y_i) + \tau D_i(y_i^0, y_i) \right\}. \end{aligned}$$

The RPDG method

Algorithm 3 The RPDG method

Let $x^0 = x^{-1} \in X$, and $\{\tau_t\}$, $\{\eta_t\}$, and $\{\alpha_t\}$ be given.
Set $y_i^0 = \nabla f_i(x^0)$, $i = 1, \ldots, m$.
for $t = 1, \ldots, k$ **do**
Choose i_t according to $\mathrm{Prob}\{i_t = i\} = p_i$, $i = 1, \ldots, m$.

$$
\begin{aligned}
\tilde{x}^t &= \alpha_t(x^{t-1} - x^{t-2}) + x^{t-1}. \\
y_i^t &= \begin{cases} \mathcal{M}_{\mathcal{Y}_i}(-\tilde{x}^t, y_i^{t-1}, \tau_t), & i = i_t, \\ y_i^{t-1}, & i \neq i_t. \end{cases} \\
\tilde{y}_i^t &= \begin{cases} p_i^{-1}(y_i^t - y_i^{t-1}) + y_i^{t-1}, & i = i_t, \\ y_i^{t-1}, & i \neq i_t. \end{cases} \\
x^t &= \mathcal{M}_X(\textstyle\sum_{i=1}^m \tilde{y}_i^t, x^{t-1}, \eta_t).
\end{aligned}
$$

end for

RPDG in gradient form

Algorithm 4 RPDG

for $t = 1, \ldots, k$ **do**

Choose i_t according to $\mathrm{Prob}\{i_t = i\} = p_i$, $i = 1, \ldots, m$.

$$
\begin{aligned}
\tilde{x}^t &= \alpha_t(x^{t-1} - x^{t-2}) + x^{t-1}. \\
\underline{x}_i^t &= \begin{cases} (1+\tau_t)^{-1}\left(\tilde{x}^t + \tau_t \underline{x}_i^{t-1}\right), & i = i_t, \\ \underline{x}_i^{t-1}, & i \neq i_t. \end{cases} \\
y_i^t &= \begin{cases} \nabla f_i(\underline{x}_i^t), & i = i_t, \\ y_i^{t-1}, & i \neq i_t. \end{cases} \\
x^t &= \mathcal{M}_X(g^{t-1} + (p_{i_t}^{-1} - 1)(y_{i_t}^t - y_{i_t}^{t-1}), x^{t-1}, \eta_t). \\
g^t &= g^{t-1} + y_{i_t}^t - y_{i_t}^{t-1}.
\end{aligned}
$$

end for

Game-theoretic interpretation for RPDG

- The suppliers predict the buyer's demand as before.
- Only one randomly selcted supplier will change his/her price, arriving at y^t.
- The buyer would have used y^t as the price, but the algorithm converges slowly (a worse depedence on m) (Dang and Lan 14).
- Add a dual prediction (estimation) step, i.e., $\tilde{y}^t$ s.t. $\mathbb{E}_t[\tilde{y}_i^t] = \hat{y}_i^t$, where $\hat{y}_i^t := \mathcal{M}_{\mathcal{Y}_i}(-\tilde{x}^t, y_i^{t-1}, \tau_i^t)$.
- The buyer uses $\tilde{y}^t$ to determine the order quantity.

Rate of Convergence

Proposition

Let $C = \frac{8L}{\mu}$. *and*

$$\begin{aligned} p_i &= \text{Prob}\{i_t = i\} = \tfrac{1}{2m} + \tfrac{L_i}{2L}, i = 1, \ldots, m, \\ \tau_t &= \frac{\sqrt{(m-1)^2+4mC}-(m-1)}{2m}, \\ \eta_t &= \frac{\mu\sqrt{(m-1)^2+4mC}+\mu(m-1)}{2}, \\ \alpha_t &= \alpha := 1 - \frac{1}{(m+1)+\sqrt{(m-1)^2+4mC}}. \end{aligned}$$

Then

$$\begin{aligned} \mathbb{E}[P(x^k, x^*)] &\leq (1 + \tfrac{3L_f}{\mu})\alpha^k P(x^0, x^*), \\ \mathbb{E}[\Psi(\bar{x}^k)] - \Psi^* &\leq \alpha^{k/2}(1-\alpha)^{-1}\left[\mu + 2L_f + \frac{L_f^2}{\mu}\right] P(x^0, x^*). \end{aligned}$$

The iteration complexity of RPGD

- To find a point $\bar{x} \in X$ s.t. $\mathbb{E}[P(\bar{x}, x^*)] \leq \epsilon$:
  $\mathcal{O}\left\{(m + \sqrt{\frac{mL}{\mu}}) \log\left[\frac{P(x^0, x^*)}{\epsilon}\right]\right\}$.
- To find a point $\bar{x} \in X$ s.t. $\text{Prob}\{P(\bar{x}, x^*) \leq \epsilon\} \geq 1 - \lambda$ for some $\lambda \in (0, 1)$:
  $\mathcal{O}\left\{(m + \sqrt{\frac{mL}{\mu}}) \log\left[\frac{P(x^0, x^*)}{\lambda\epsilon}\right]\right\}$.
- A factor of $\mathcal{O}\left\{\min\{\sqrt{\frac{L}{\mu}}, \sqrt{m}\}\right\}$ savings on gradient computation (or price changes), if $L \approx L_f$, at the price of more order transactions.

Lower complexity bound

$\min_{x_i \in \mathbb{R}^{\tilde{n}}, i=1,\ldots,m} \left\{ \Psi(x) := \sum_{i=1}^{m} \left[f_i(x_i) + \frac{\mu}{2}\|x_i\|_2^2 \right] \right\}$.

$f_i(x_i) = \frac{\mu(\mathcal{Q}-1)}{4} \left[\frac{1}{2}\langle A x_i, x_i \rangle - \langle e_1, x_i \rangle \right]$. $\tilde{n} \equiv n/m$,

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\ 0 & 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & \kappa \end{pmatrix}, \kappa = \frac{\sqrt{\mathcal{Q}}+3}{\sqrt{\mathcal{Q}}+1}.$$

Theorem

Denote $q := (\sqrt{\mathcal{Q}} - 1)(\sqrt{\mathcal{Q}} + 1)$. *Then the iterates* $\{x^k\}$ *generated by a randomized incremental gradient method must satisfy* $\frac{\mathbb{E}[\|x^k - x^*\|_2^2]}{\|x^0 - x^*\|_2^2} \geq \frac{1}{2} \exp\left(- \frac{4k\sqrt{\mathcal{Q}}}{m(\sqrt{\mathcal{Q}}+1)^2 - 4\sqrt{\mathcal{Q}}} \right)$ *for any* $n \geq \underline{n}(m,k) \equiv [m \log\left[\left(1 - (1-q^2)/m\right)^k /2 \right]]/(2 \log q)$.

Complexity

Corollary

The number of gradient evaluations performed by any randomized incremental gradient methods for finding a solution $\bar{x} \in X$ *s.t.* $\mathbb{E}[\|\bar{x} - x^*\|_2^2] \leq \epsilon$ *cannot be smaller than* $\Omega\left\{\left(\sqrt{m\mathcal{C}} + m\right) \log \frac{\|x^0 - x^*\|_2^2}{\epsilon}\right\}$ *if* n *is sufficiently large.*

Other results in the paper

- Generalization to problems without strong convexity.
- Lower complexity bound for randomized coordinate descent methods.

Summary

- Present a primal-dual gradient method which exhibits optimal black-box complexity.
- Present a randomized primal-dual gradient method.
- Introduce a lower complexity bound for randomized incremental gradient methods.
- Introduce a game-theoretic interpretation for first-order methods.