Stochastic methods for minimizing finite sum of convex functions

Yangyang Xu

IMA, University of Minnesota

December 4, 2015

Where the materials come from

- M. Schmidt, N. Roux and F. Bach. Minimizing Finite Sums with the Stochastic Average Gradient. arXiv:1309.2388
- L. Xiao and T. Zhang. A Proximal Stochastic Gradient Method with Progressive Variance Reduction. SIOPT14.
- Z. Zhu and Y. Yuan. UniVR: A Universal Variance Reduction Framework for Proximal Stochastic Gradient Method. arXiv:1506.01972.
- S. Shwartz and T. Zhang. Accelerated Proximal Stochastic Dual Coordinate Ascent for Regularized Loss Minimization. MPA15
- G. Lan, Y. Zhou. An optimal randomized incremental gradient method. arXiv:1507.02000.

Stochastic gradient method

Consider the stochastic programming

$$\min_{\mathbf{x}\in X} F(x) = \mathbb{E}_\xi f(\mathbf{x};\xi).$$

Stochastic gradient update (SG):

$$\mathbf{x}^{k+1} = \mathcal{P}_X\left(\mathbf{x}^k - \alpha_k \tilde{\mathbf{g}}^k\right)$$

- $\tilde{\mathbf{g}}^k$ a stochastic gradient, often $\mathbb{E}[\tilde{\mathbf{g}}^k] \in \partial F(\mathbf{x}^k)$
- Originally for stochastic problem where exact gradient not available
- Now also popular for deterministic problem where exact gradient expensive; e.g., $F(x) = \frac{1}{N}\sum_{i=1}^N f_i(\mathbf{x})$ with large N
- Faster than deterministic gradient method to reach not-high accuracy

Stochastic gradient method

- First appears in [Robbins-Monro'51]; now tons of works
- $\mathcal{O}(1/\sqrt{k})$ rate for weakly convex problem and $\mathcal{O}(1/k)$ for strongly convex problem (e.g., [Nemirovski et. al'09])
- For deterministic problem, faster convergence is possible

This talk is about how to accelerate SG method

A simple analysis

Assume $X = \mathbb{R}^n$ and F is L-smooth. Let $\tilde{\boldsymbol{\delta}}^k = \tilde{\mathbf{g}}^k - \nabla F(\mathbf{x}^k)$. Then

$$\begin{aligned} F(\mathbf{x}^{k+1}) \leq & F(\mathbf{x}^k) + \langle \nabla F(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \\ = & F(\mathbf{x}^k) - \alpha_k \langle \nabla F(\mathbf{x}^k), \tilde{\mathbf{g}}^k \rangle + \frac{L\alpha_k^2}{2}\|\tilde{\mathbf{g}}^k\|^2 \\ = & F(\mathbf{x}^k) - \alpha_k \langle \nabla F(\mathbf{x}^k), \tilde{\mathbf{g}}^k \rangle + \frac{L\alpha_k^2}{2}\left(\|\tilde{\boldsymbol{\delta}}^k\|^2 + 2\langle \tilde{\boldsymbol{\delta}}^k, \nabla F(\mathbf{x}^k)\rangle + \|\nabla F(\mathbf{x}^k)\|^2\right) \end{aligned}$$

If $\mathbb{E}\tilde{\boldsymbol{\delta}}^k = \mathbf{0}$ and $\mathbb{E}\|\tilde{\boldsymbol{\delta}}^k\|^2 \leq \sigma_k^2$, then

$$\mathbb{E}F(\mathbf{x}^{k+1}) \leq \mathbb{E}F(\mathbf{x}^k) - \alpha_k(1 - L\alpha_k/2)\|\nabla F(\mathbf{x}^k)\|^2 + \frac{L\alpha_k^2\sigma_k^2}{2}.$$

The variance σ_k affects the convergence and speed

Stochastic average gradient

Consider problem

$$\min_{x \in \mathbb{R}^p} g(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

Stochastic average gradient

$$x^{k+1} = x^k - \frac{\alpha_k}{n} \sum_{i=1}^{n} y_i^k \qquad \text{(SAG)}$$

where at iteration k, a random index i_k is selected uniformly at random and

$$y_i^k = \begin{cases} \nabla f_i(\mathbf{x}^k), & \text{if } i = i_k, \\ y_i^{k-1}, & \text{otherwise} \end{cases}$$

Drawback: require storing $y_1, \ldots, y_n$

M. Schmidt, N. Roux and F. Bach. Minimizing Finite Sums with the Stochastic Average Gradient. arXiv:1309.2388

Convergence of SAG

Assume $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|$, $\forall i$.

Theorem 1 *With a constant step size of $\alpha_k = \frac{1}{16L}$, the SAG iterations satisfy for $k \geq 1$:*

$$\mathbb{E}[g(\bar{x}^k)] - g(x^*) \leqslant \frac{32n}{k} C_0,$$

where if we initialize with $y_i^0 = 0$ we have

$$C_0 = g(x^0) - g(x^*) + \frac{4L}{n}\|x^0 - x^*\|^2 + \frac{\sigma^2}{16L},$$

and if we initialize with $y_i^0 = f_i'(x^0) - g'(x^0)$ we have

$$C_0 = \frac{3}{2}\left[g(x^0) - g(x^*)\right] + \frac{4L}{n}\|x^0 - x^*\|^2.$$

Further, if g is μ-strongly convex we have

$$\mathbb{E}[g(x^k)] - g(x^*) \leqslant \left(1 - \min\left\{\frac{\mu}{16L}, \frac{1}{8n}\right\}\right)^k C_0.$$

$O\left(\max\{n, \frac{L}{\mu}\} \log \frac{1}{\epsilon}\right)$ component gradients to reach ϵ-optimal solution

Stochastic variance-reduced gradient method

Consider problem

$$\min_{x\in\mathbb{R}^d} P(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) + R(x)$$

Stochastic variance-reduced gradient method periodically evaluates the full gradient of $F(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)$

L. Xiao and T. Zhang. A Proximal Stochastic Gradient Method with Progressive Variance Reduction. SIOPT14.

Stochastic variance-reduced gradient method

Algorithm: Prox-SVRG$(\tilde{x}_0, \eta, m)$

iterate: for $s = 1, 2, \ldots$

$\tilde{x} = \tilde{x}_{s-1}$

$\tilde{v} = \nabla F(\tilde{x})$

$x_0 = \tilde{x}$

probability $Q = \{q_1, \ldots, q_n\}$ on $\{1, \ldots, n\}$

iterate: for $k = 1, 2, \ldots, m$

pick $i_k \in \{1, \ldots, n\}$ randomly according to Q

$v_k = (\nabla f_{i_k}(x_{k-1}) - \nabla f_{i_k}(\tilde{x}))/(q_{i_k} n) + \tilde{v}$

$x_k = \text{prox}_{\eta R}(x_{k-1} - \eta v_k)$

end

set $\tilde{x}_s = \frac{1}{m} \sum_{k=1}^{m} x_k$

end

L. Xiao and T. Zhang. A Proximal Stochastic Gradient Method with Progressive Variance Reduction. SIOPT14.

Assumptions

Assumption 1. *The function $R(x)$ is lower semi-continuous and convex, and its effective domain,* $\mathrm{dom}(R) := \{x \in \mathbb{R}^d \mid R(x) < +\infty\}$*, is closed. Each $f_i(x)$, for $i = 1, \ldots, n$, is differentiable on an open set that contains* $\mathrm{dom}(R)$*, and their gradients are Lipschitz continuous. That is, there exist $L_i > 0$ such that for all $x, y \in \mathrm{dom}(R)$,*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i \|x - y\|. \tag{3}$$

Assumption 2. *The overall cost function $P(x)$ is strongly convex, i.e., there exist $\mu > 0$ such that for all $x \in \mathrm{dom}(R)$ and $y \in \mathbb{R}^d$,*

$$P(y) \geq P(x) + \xi^T (y - x) + \frac{\mu}{2} \|y - x\|^2, \quad \forall\, \xi \in \partial P(x). \tag{4}$$

Convergence of Prox-SVRG

Variance approaches to *zero*:

$$\mathbb{E}\|v_k - \nabla F(x_{k-1})\|^2 \leq 4L_{\max}\big[P(x_{k-1}) - P(x_\star) + P(\tilde{x}) - P(x_\star)\big]. \tag{12}$$

Therefore, when both x_{k-1} and $\tilde{x}$ converge to $x_\star$, the variance of v_k also converges to zero. As a result, we can use a constant step size and obtain much faster convergence.

Theorem 1. *Suppose Assumptions 1 and 2 hold, and let* $x_\star = \arg\min_x P(x)$ *and* $L_Q = \max_i L_i/(q_i n)$. *In addition, assume that* $0 < \eta < 1/(4L_Q)$ *and* m *is sufficiently large so that*

$$\rho = \frac{1}{\mu\eta(1 - 4L_Q\eta)m} + \frac{4L_Q\eta(m+1)}{(1 - 4L_Q\eta)m} < 1. \tag{14}$$

Then the Prox-SVRG method in Figure 2 has geometric convergence in expectation:

$$\mathbb{E}P(\tilde{x}_s) - P(x_\star) \leq \rho^s[P(\tilde{x}_0) - P(x_\star)].$$

$O\left((n + \frac{L_{avg}}{\mu})\log\frac{1}{\epsilon}\right)$ component gradients to reach ϵ-optimal solution

A universal variance reduction framework

Consider problem

$$\min_{x\in\mathbb{R}^d}\left\{F(x)\stackrel{\text{def}}{=}f(x)+\Psi(x)\stackrel{\text{def}}{=}\frac{1}{n}\sum_{i=1}^{n}f_i(x)+\Psi(x)\right\}\ . \tag{1.1}$$

Motivation:

Although many variance-reduction based methods have been proposed, most of them only apply to Problem (1.1) when the objective function $F(x)$ is strongly convex [3, 6, 15, 16, 19]. However, in many machine learning applications, $F(x)$ is simply *not* strongly convex. This is particularly true for Lasso [18] and ℓ_1-Regularized Logistic Regression [11], two cornerstone problems extensively used for feature selections.

One way to get around this issue is to add a dummy regularizer $\frac{\lambda}{2}\|x\|_2^2$ to $F(x)$, and then to apply any of the above strong-convexity methods. However, the weight of this regularizer, λ, needs to be chosen before the algorithm starts. This adds a lot of difficulty when applying such methods to real life: (1) one needs to tune λ by repeatedly executing the algorithm, and (2) the error of the algorithm does not converge to zero as time goes (in fact, it converges to $O(\lambda)$ so one needs to know the desired accuracy before the algorithm starts). As we shall demonstrate in the experimental section, adding the dummy regularizer hurts the performance of the algorithm as well.

Z. Zhu and Y. Yuan. UniVR: A Universal Variance Reduction Framework for Proximal Stochastic Gradient Method. arXiv:1506.01972.

A universal variance reduction framework

Algorithm 1 `UniVR`(x^ϕ, m_0, S, η)

$\tilde{x}^0 \leftarrow x^\phi$, $x_0^1 \leftarrow x^\phi$
for $s \leftarrow 1$ **to** S **do**
  $\tilde{\mu}_{s-1} \leftarrow \nabla F(\tilde{x}^{s-1})$
  $m_s \leftarrow 2^s \cdot m_0$
  for $t \leftarrow 0$ **to** $m_s - 1$ **do**
    Pick i uniformly at random in $\{1, \cdots, n\}$.
    $\xi \leftarrow \nabla f_i(x_t^s) - \nabla f_i(\tilde{x}^{s-1}) + \tilde{\mu}_{s-1}$
    $x_{t+1}^s = \arg\min_{y \in \mathbb{R}^d} \left\{ \frac{1}{2\eta} \|x_t^s - y\|^2 + \Psi(y) + \langle \xi, y \rangle \right\}$
  end for
  $\tilde{x}^s \leftarrow \frac{1}{m_s} \sum_{t=1}^{m_s} x_t^s$
  $x_0^{s+1} \leftarrow x_{m_s}^s$
end for
return $\tilde{x}^S$.

Z. Zhu and Y. Yuan. UniVR: A Universal Variance Reduction Framework for Proximal Stochastic Gradient Method. arXiv:1506.01972.

Convergence of UniVR

Assumption:

Throughout this paper, we use $\|\cdot\|$ to denote the Euclidean norm. We assume that each $f_i(\cdot)$ is convex, differentiable and L-*smooth* (or has L-Lipschitz continuous gradient):

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^d \ .$$

In addition, we assume that $\Psi(\cdot)$ is convex and lower semicontinuous.

Theorem 4.3. `UniVR`(x^ϕ, m_0, S, η) *satisfies if* m_0 *and* S *are positive integers and* $\eta = 1/7L$*, then*

$$\mathbb{E}[F(\tilde{x}^S) - F(x^*)] \leq O\Big(\frac{F(x^\phi) - F(x^*)}{2^S} + \frac{L\|x^* - x^\phi\|^2}{2^S m_0}\Big) \ . \tag{4.2}$$

In addition, `UniVR` *has a gradient complexity of* $O(S \cdot n + 2^S \cdot m_0)$.

Linear convergence was shown for a modified UniVR.

Accelerated proximal stochastic dual coordinate ascent

Consider problem

$$\min_{w\in\mathbb{R}^d} P(w) \quad \text{where} \quad P(w) = \left[\frac{1}{n}\sum_{i=1}^{n}\phi_i(X_i^\top w) + \lambda g(w)\right]. \tag{1}$$

Dual problem:

$$\max_{\alpha\in\mathbb{R}^{k\times n}} D(\alpha) \quad \text{where} \quad D(\alpha) = \left[\frac{1}{n}\sum_{i=1}^{n} -\phi_i^*(-\alpha_i) - \lambda g^*\left(\frac{1}{\lambda n}\sum_{i=1}^{n} X_i\alpha_i\right)\right], \tag{2}$$

where α_i is the i'th column of the matrix α, which forms a vector in $\mathbb{R}^k$.

- ϕ^* and g^* are conjugate functions
- $\|X_i\| \leq R, \forall i$

S. Shwartz and T. Zhang. Accelerated Proximal Stochastic Dual Coordinate Ascent for Regularized Loss Minimization. MPA15

Proximal stochastic dual coordinate ascent

Procedure Proximal Stochastic Dual Coordinate Ascent:
Prox-SDCA($P, \epsilon, \alpha^{(0)}$)

Goal: Minimize $P(w) = \frac{1}{n}\sum_{i=1}^{n} \phi_i(X_i^\top w) + \lambda g(w)$
Input: Objective P, desired accuracy ϵ, initial dual solution $\alpha^{(0)}$ (default: $\alpha^{(0)} = 0$)
Assumptions:
$\forall i$, ϕ_i is $(1/\gamma)$-smooth w.r.t. $\|\cdot\|_P$ and let $\|\cdot\|_D$ be the dual norm of $\|\cdot\|_P$
g is 1-strongly convex w.r.t. $\|\cdot\|_{P'}$ and let $\|\cdot\|_{D'}$ be the dual norm of $\|\cdot\|_{P'}$
$\forall i$, $\|X_i\|_{D\to D'} \le R$
Initialize $v^{(0)} = \frac{1}{\lambda n}\sum_{i=1}^{n} X_i \alpha_i^{(0)}$, $w^{(0)} = \nabla g^*(0)$
Iterate: for $t = 1, 2, \ldots$
Randomly pick i
Find $\Delta\alpha_i$ using any of the following options
(or any other update that achieves a larger dual objective):
Option I:
$$\Delta\alpha_i = \operatorname*{argmax}_{\Delta\alpha_i} \left[-\phi_i^*(-(\alpha_i^{(t-1)} + \Delta\alpha_i)) - w^{(t-1)^\top} X_i \Delta\alpha_i - \frac{1}{2\lambda n} \|X_i \Delta\alpha_i\|_{D'}^2 \right]$$
Option II:
Let $u = -\nabla\phi_i(X_i^\top w^{(t-1)})$ and $q = u - \alpha_i^{(t-1)}$
Let $s = \operatorname*{argmax}_{s\in[0,1]} \left[-\phi_i^*(-(\alpha_i^{(t-1)} + sq)) - s\, w^{(t-1)^\top} X_i q - \frac{s^2}{2\lambda n} \|X_i q\|_{D'}^2 \right]$
Set $\Delta\alpha_i = sq$
Option III:
Same as Option II but replace the definition of s as follows:
Let $s = \min\left(1, \frac{\phi_i(X_i^\top w^{(t-1)}) + \phi_i^*(-\alpha_i^{(t-1)}) + w^{(t-1)^\top} X_i \alpha_i^{(t-1)} + \frac{\gamma}{2}\|q\|_D^2}{\|q\|_D^2 (\gamma + \frac{1}{\lambda n}\|X_i\|_{D\to D'}^2)}\right)$
Option IV:
Same as Option III but replace $\|X_i\|_{D\to D'}^2$ in the definition of s with R^2
Option V:
Same as Option II but replace the definition of s to be $s = \frac{\lambda n \gamma}{R^2 + \lambda n \gamma}$
$\alpha_i^{(t)} \leftarrow \alpha_i^{(t-1)} + \Delta\alpha_i$ and for $j \ne i$, $\alpha_j^{(t)} \leftarrow \alpha_j^{(t-1)}$
$v^{(t)} \leftarrow v^{(t-1)} + (\lambda n)^{-1} X_i \Delta\alpha_i$
$w^{(t)} \leftarrow \nabla g^*(v^{(t)})$
Stopping condition:
Let $T_0 < t$ (default: $T_0 = t - n - \lceil \frac{R^2}{\lambda\gamma} \rceil$)
Averaging option:
Let $\bar\alpha = \frac{1}{t - T_0}\sum_{i=T_0+1}^{t} \alpha^{(i-1)}$ and $\bar w = \frac{1}{t-T_0}\sum_{i=T_0+1}^{t} w^{(i-1)}$
Random option:
Let $\bar\alpha = \alpha^{(i)}$ and $\bar w = w^{(i)}$ for some random $i \in T_0 + 1, \ldots, t$
Stop if $P(\bar w) - D(\bar\alpha) \le \epsilon$ and output $\bar w, \bar\alpha$, and $P(\bar w) - D(\bar\alpha)$

S. Shwartz and T. Zhang. Accelerated Proximal Stochastic Dual Coordinate Ascent for Regularized Loss Minimization. MPA15

Convergence result of Prox-SDCA

Assumption: g is 1-strongly convex, and every ϕ_i is $(1/\gamma) - smooth$.

Theorem 1. *Consider Procedure Prox-SDCA as given in Figure 1. Let α^* be an optimal dual solution and let $\epsilon > 0$. For every T such that*

$$T \geq \left(n + \frac{R^2}{\lambda\gamma}\right) \log\left(\left(n + \frac{R^2}{\lambda\gamma}\right) \cdot \frac{D(\alpha^*) - D(\alpha^{(0)})}{\epsilon}\right),$$

we are guaranteed that $\mathbb{E}[P(w^{(T)}) - D(\alpha^{(T)})] \leq \epsilon$. Moreover, for every T such that

$$T \geq \left(n + \left\lceil \frac{R^2}{\lambda\gamma} \right\rceil\right) \cdot \left(1 + \log\left(\frac{D(\alpha^*) - D(\alpha^{(0)})}{\epsilon}\right)\right),$$

let $T_0 = T - n - \lceil \frac{R^2}{\lambda\gamma} \rceil$, then we are guaranteed that $\mathbb{E}[P(\bar{w}) - D(\bar{\alpha})] \leq \epsilon$.

High-probability result was shown by using Theorem 1 and Markov inequality.

$O\left((n + \frac{1}{\lambda\gamma}) \log \frac{1}{\epsilon}\right)$ component gradients to reach ϵ-optimal solution

Accelerated proximal stochastic dual coordinate ascent

`Procedure Accelerated Prox-SDCA`

Goal: Minimize $P(w) = \frac{1}{n}\sum_{i=1}^{n} \phi_i(X_i^\top w) + \lambda g(w)$
Input: Target accuracy ϵ (only used in the stopping condition)
Assumptions:
$\forall i$, ϕ_i is $(1/\gamma)$-smooth w.r.t. $\|\cdot\|_P$ and let $\|\cdot\|_D$ be the dual norm of $\|\cdot\|_P$
g is 1-strongly convex w.r.t. $\|\cdot\|_2$
$\forall i$, $\|X_i\|_{D\to 2} \le R$
$\frac{R^2}{\gamma\lambda} > 10\,n$ (otherwise, solve the problem using vanilla Prox-SDCA)
Define $\kappa = \frac{R^2}{\gamma n} - \lambda$, $\mu = \lambda/2$, $\rho = \mu + \kappa$, $\eta = \sqrt{\mu/\rho}$, $\beta = \frac{1-\eta}{1+\eta}$,
Initialize $y^{(1)} = w^{(1)} = 0,\ \alpha^{(1)} = 0$, $\xi_1 = (1+\eta^{-2})(P(0) - D(0))$
Iterate: for $t = 2, 3, \ldots$
Let $\tilde{P}_t(w) = \frac{1}{n}\sum_{i=1}^{n} \phi_i(X_i^\top w) + \tilde{\lambda}\tilde{g}_t(w)$
where $\tilde{\lambda}\tilde{g}_t(w) = \lambda g(w) + \frac{\kappa}{2}\|w\|_2^2 - \kappa\, w^\top y^{(t-1)}$
Call $(w^{(t)}, \alpha^{(t)}, \epsilon_t) = \text{Prox-SDCA}\left(\tilde{P}_t, \frac{\eta}{2(1+\eta^{-2})}\xi_{t-1}, \alpha^{(t-1)}\right)$
Let $y^{(t)} = w^{(t)} + \beta(w^{(t)} - w^{(t-1)})$
Let $\xi_t = (1-\eta/2)^{t-1}\,\xi_1$
Stopping conditions: break and return $w^{(t)}$ if one of the following conditions hold:
1. $t \ge 1 + \frac{2}{\eta}\log(\xi_1/\epsilon)$
2. $(1+\rho/\mu)\epsilon_t + \frac{\rho\kappa}{2\mu}\|w^{(t)} - y^{(t-1)}\|^2 \le \epsilon$

Convergence result of accelerated Prox-SDCA

Assumption: g is 1-strongly convex, and every ϕ_i is $(1/\gamma) - smooth$.

Theorem 3. *Consider the accelerated Prox-SDCA algorithm given in Figure 2.*

- *Correctness: When the algorithm terminates we have that* $P(w^{(t)}) - P(w^*) \leq \epsilon$.
- *Runtime:*
  - *The number of outer iterations is at most*

$$1 + \frac{2}{\eta} \log(\xi_1/\epsilon) \;\leq\; 1 + \sqrt{\frac{8R^2}{\lambda\gamma n}} \left(\log\left(\frac{2R^2}{\lambda\gamma n}\right) + \log\left(\frac{P(0) - D(0)}{\epsilon}\right)\right) \;.$$

  - *Each outer iteration involves a single call to Prox-SDCA, and the averaged runtime required by each such call is*

$$O\left(d\, n \log\left(\frac{R^2}{\lambda\gamma n}\right)\right) \;.$$

$O\left((n + \min\{\frac{1}{\lambda\gamma}, \sqrt{\frac{n}{\lambda\gamma}}\}) \log\frac{1}{\epsilon}\right)$ component gradients to reach ϵ-optimal solution

Randomized primal-dual gradient

Consider problem

$$\Psi^* := \min_{x \in X} \left\{ \Psi(x) := \sum_{i=1}^m f_i(x) + h(x) + \mu\, \omega(x) \right\}. \tag{1.1}$$

Saddle-point problem:

$$\Psi^* := \min_{x \in X} \left\{ h(x) + \mu\, \omega(x) + \max_{g \in \mathcal{G}} \langle x, g \rangle - J_f(g) \right\}. \tag{2.7}$$

- $\omega(x)$ is 1-strongly convex
- J_f is the conjugate function of f.

G. Lan, Y. Zhou. An optimal randomized incremental gradient method. arXiv:1507.02000.

Primal-dual gradient

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 (x^t, g^t) according to

$$\tilde{x}^t = \alpha_t(x^{t-1} - x^{t-2}) + x^{t-1}. \tag{2.8}$$

$$g^t = \mathcal{M}_{\mathcal{G}}(-\tilde{x}^t, g^{t-1}, \tau_t). \tag{2.9}$$

$$x^t = \mathcal{M}_X(g^t, x^{t-1}, \eta_t). \tag{2.10}$$

end for

where

$$\mathcal{M}_X(g, x^0, \eta) \equiv \mathcal{M}_{X,\omega,h}(g, x^0, \eta) := \arg\min_{x \in X} \left\{ \langle g, x \rangle + h(x) + \mu\, \omega(x) + \eta P(x^0, x) \right\}, \tag{2.3}$$

$$\mathcal{M}_{\mathcal{G}}(-\tilde{x}, g^0, \tau) := \arg\min_{g \in \mathcal{G}} \left\{ \langle -\tilde{x}, g \rangle + J_f(g) + \tau D_f(g^0, g) \right\}, \tag{2.6}$$

and P and D_f are prox-functions.

G. Lan, Y. Zhou. An optimal randomized incremental gradient method. arXiv:1507.02000.

Convergence primal-dual gradient

Theorem 1 *Let x^* be an optimal solution of (1.1), x^k and $\bar{x}^k$ be defined in (2.10) and (2.23), respectively.*

a) Suppose that $\mu > 0$ and that $\{\tau_t\}$, $\{\eta_t\}$, $\{\alpha_t\}$ and $\{\theta_t\}$ are set to

$$\tau_t = \sqrt{\tfrac{2L_f}{\mu}}, \quad \eta_t = \sqrt{2L_f\mu}, \quad \alpha_t = \alpha \equiv \tfrac{\sqrt{2L_f/\mu}}{1+\sqrt{2L_f/\mu}}, \quad \text{and} \quad \theta_t = \tfrac{1}{\alpha^t}, \ \forall t = 1, \ldots, k. \tag{2.24}$$

Then,

$$P(x^k, x^*) \leq \tfrac{\mu+L_f}{\mu}\alpha^k P(x^0, x^*), \tag{2.25}$$

$$\Psi(\bar{x}^k) - \Psi(x^*) \leq \text{gap}^*(\bar{z}^k) \leq \mu(1-\alpha)^{-1}\left[1 + \tfrac{L_f}{\mu}(2 + \tfrac{L_f}{\mu})\right]\alpha^k P(x^0, x^*), \tag{2.26}$$

$$\Psi(\bar{x}^k) - \Psi(x^*) \leq \text{gap}(\bar{z}^k) \leq \mu(1-\alpha)^{-1}\left[1 + \tfrac{L_f}{\mu}(2 + \tfrac{L_f}{\mu})\right]\alpha^k \max_{x\in X} P(x^0, x). \tag{2.27}$$

b) Suppose that $\{\tau_t\}$, $\{\eta_t\}$, $\{\alpha_t\}$ and $\{\theta_t\}$ are set to

$$\tau_t = \frac{t-1}{2}, \quad \eta_t = \frac{4L_f}{t}, \quad \alpha_t = \frac{t-1}{t} \quad \text{and} \quad \theta_t = t, \ \forall t = 1, \ldots, k. \tag{2.28}$$

Then,

$$\Psi(\bar{x}^k) - \Psi(x^*) \leq \text{gap}^*(\bar{z}^k) \leq \frac{8L_f}{k(k+1)} P(x^0, x^*), \tag{2.29}$$

$$\Psi(\bar{x}^k) - \Psi(x^*) \leq \text{gap}(\bar{z}^k) \leq \frac{8L_f}{k(k+1)} \max_{x\in X} P(x^0, x). \tag{2.30}$$

Randomized primal-dual gradient

The primal problem (1.1) can be equivalently written to

$$\Psi^* := \min_{x\in X}\left\{h(x) + \mu\,\omega(x) + \max_{y\in\mathcal{Y}}\langle x, Uy\rangle - J(y)\right\}, \tag{3.1}$$

where $U = (I, \ldots, I)$, $y = (y_1, \ldots, y_m)$, $J(y) = \sum_i J_i(y_i)$, and J_i is the conjugate of f_i.

Algorithm 3 A randomized primal-dual gradient (RPDG) method

Let $x^0 = x^{-1} \in X$, and the nonnegative parameters $\{\tau_t\}$, $\{\eta_t\}$, 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 $\text{Prob}\{i_t = i\} = p_i$, $i = 1, \ldots, m$.
Update $z^t = (x^t, y^t)$ according to

$$\tilde{x}^t = \alpha_t(x^{t-1} - x^{t-2}) + x^{t-1}. \tag{3.7}$$

$$y_i^t = \begin{cases}\mathcal{M}_{\mathcal{Y}_i}(-\tilde{x}^t, y_i^{t-1}, \tau_t), & i = i_t,\\ y_i^{t-1}, & i \neq i_t.\end{cases} \tag{3.8}$$

$$\tilde{y}_i^t = \begin{cases}p_i^{-1}(y_i^t - y_i^{t-1}) + y_i^{t-1}, & i = i_t,\\ y_i^{t-1}, & i \neq i_t.\end{cases}. \tag{3.9}$$

$$x^t = \mathcal{M}_X(\textstyle\sum_{i=1}^m \tilde{y}_i^t, x^{t-1}, \eta_t). \tag{3.10}$$

end for

$$\mathcal{M}_{\mathcal{Y}_i}(-\tilde{x}, y_i^0, \tau) := \arg\min_{y_i\in\mathcal{Y}_i}\left\{\langle -\tilde{x}, y\rangle + J_i(y_i) + \tau D_i(y_i^0, y_i)\right\}, \tag{3.5}$$

Convergence result of randomized primal-dual gradient

Theorem 2 *Suppose that* $\{\tau_t\}$, $\{\eta_t\}$, *and* $\{\alpha_t\}$ *in the RPDG method are set to*

$$\tau_t = \tau, \quad \eta_t = \eta, \quad \text{and} \quad \alpha_t = \alpha, \tag{3.19}$$

for any $t \geq 1$ *such that*

$$(1-\alpha)(1+\tau) \leq p_i, i = 1, \ldots, m, \tag{3.20}$$

$$\eta \leq \alpha(\mu + \eta), \tag{3.21}$$

$$\eta\tau p_i \geq 4L_i, i = 1, \ldots, m, \tag{3.22}$$

for some $\alpha \in (0, 1)$. *Then, for any* $k \geq 1$, *we have*

$$\mathbb{E}[P(x^k, x^*)] \leq \left(1 + \tfrac{L_f \alpha}{(1-\alpha)\eta}\right) \alpha^k P(x^0, x^*), \tag{3.23}$$

$$\mathbb{E}[\Psi(\bar{x}^k) - \Psi(x^*)] \leq \alpha^{k/2} \left(\alpha^{-1}\eta + \tfrac{3-2\alpha}{1-\alpha} L_f + \tfrac{2L_f^2 \alpha}{(1-\alpha)\eta}\right) P(x^0, x^*), \tag{3.24}$$

where $\bar{x}^k = (\sum_{t=1}^k \theta_t)^{-1} \sum_{t=1}^k (\theta_t x^t)$ *with* $\{\theta_t\}$ *defined as in (2.24), and* x^* *denotes the optimal solution of problem (1.1), and the expectation is taken w.r.t.* $i_1, \ldots, i_k$.

$O\left((m + \sqrt{\frac{mL}{\mu}}) \log \frac{1}{\epsilon}\right)$ component gradients to reach ϵ-optimal solution

This bound is optimal!

Summary of complexity result

Consider the problem $\min_{\mathbf{x}} \Phi(\mathbf{x}) := \sum_{i=1}^{N} f_i(\mathbf{x}) + R(\mathbf{x})$.

Assume $F(\mathbf{x}) = \sum_{i=1}^{N} f_i(\mathbf{x})$ is L-smooth and Φ is μ-strongly convex. Then to make an ϵ-optimal solution, we have the following complexity results based on *evaluation of component gradients*.

Methods	Complexity
Accelerated prox-grad [Nesterov]	$O\left(N\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$
SAG [Schmidt-Roux-Bach'13]	$O\left(\max\{N, \frac{L}{\mu}\}\log\frac{1}{\epsilon}\right)$
Prox-SVRG [Xiao-Zhang'14]	$O\left((N+\frac{L}{\mu})\log\frac{1}{\epsilon}\right)$
Prox-SDCA [Shwartz-Zhang'15]	$O\left((N+\frac{L}{\mu})\log\frac{1}{\epsilon}\right)$
Accelerated prox-SDCA [Shwartz-Zhang'15]	$O\left((N+\min\{\frac{L}{\mu}, \sqrt{\frac{NL}{\mu}}\})\log\frac{1}{\epsilon}\right)$
RPDG [Lan-Zhou'15]	$O\left((N+\sqrt{\frac{NL}{\mu}})\log\frac{1}{\epsilon}\right)$