Adaptive Proximal Average based Variance Reducing Stochastic Methods for Optimization with Composite Regularization

Jingchang Liu, Linli Xu, Junliang Guo, Xin Sheng

School of Computer Science and Technology, University of Science and Technology of China, Hefei, China

1. Introduction for basic methods

Traditional Formulation:

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

- $f_i : \mathbb{R}^d \to \mathbb{R}$: the empirical loss of the i-th sample with regard to the parameter x.
- r: the regularization term, which is convex but possibly non-smooth.
- Examples: LASSO, sparse SVM, ℓ_1, ℓ_2-Logistic Regression.

Forward-Backward Splitting:

$$x^{k+1} = \mathrm{prox}_r^{\gamma}(x^k - \gamma\cdot\Box), \tag{2}$$

where $\Box$ can be $\nabla f(x^k)$ in GD, $\nabla f_i(x^k)$ in SGD, or $\nabla f_i(x^k) - \nabla f_i(\tilde{x}) + \nabla f(\tilde{x})$ in variance reducing stochastic gradient descent.

Proximal Operator:

$$\mathrm{prox}_r^{\gamma}(x) = \arg\min_{y\in\mathbb{R}^d}\Big(r(y) + \frac{1}{2\gamma}\|y-x\|^2\Big). \tag{3}$$

One requirement for using proximal operators is that $\mathrm{prox}_r^{\gamma}(x)$ can be calculated effectively.

2. More complex problem

Composite Regularization:

$$\begin{aligned}\min_{x\in\mathbb{R}^d} F(x) &= f(x) + r(x)\\ &= \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \sum_{k=1}^{K} w_k r_k(x),\end{aligned}$$

where $w_k \ge 0$ and $\sum_{k=1}^K w_k = 1$.

Examples:

- Overlapping group lasso: $r(x) = \lambda\sum_{k=1}^{K}\|x_{g_k}\|$.
- Graph-guided fused lasso:

$$r(x) = \sum_{\{i,j\}\in\mathcal{E}} w_{ij}|x_i - x_j|.$$

Difficulty:

$\mathrm{prox}_r^{\gamma}(x)$ is hard to be calculated.

Drawbacks of Existing Methods

- ADMM: requires more space and involves complex implementation and convergence analysis.
- Three operator splitting: involves strong assumption.

3. Related works

Variance Reducing Stochastic Methods (Prox-SVRG, Prox-SAGA)

- Use $v^k = \nabla f_j(x^k) - \nabla f_j(\tilde{x}) + \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\tilde{x})$ to replace $\nabla f_j(x^k)$.

- Prox-SVRG:

Define $\theta = \frac{1}{\gamma\mu(1-4L\gamma)m} + \frac{4L\gamma(m+1)}{(1-4L\gamma)m}$, then

$$\mathbb{E}F(\tilde{x}_s) - F^* \le \theta[F(\tilde{x}_{s-1}) - F^*]. \tag{4}$$

If $0<\gamma<1/(4L)$ and m is large enough such that $\theta<1$, then Prox-SVRG can achieve the linear convergence rate.

- Prox-SAGA:

By the Lyapunov function $T^k = \frac{1}{n}\sum_{i=1}^{n} f_i(x_i^k) - f(x^*) - \frac{1}{n}\sum_{i=1}^{n}\langle\nabla f_i(x^*), x_i^k - x^*\rangle + c\|x^k - x^*\|^2$,

$$\mathbb{E}\|x^k - x^*\|^2 \le \Big(1 - \min\Big\{\frac{1}{4n}, \frac{\mu}{3L}\Big\}\Big)^k T^0. \tag{5}$$

Proximal Average (PA)

- Definition

The proximal average of r is the unique semi-continuous convex function $\hat{r}(x)$ such that

$$\mathrm{prox}_{\hat{r}}^{\gamma}(x) = \sum_{k=1}^{K} w_k\cdot\mathrm{prox}_{r_k}^{\gamma}(x). \tag{6}$$

- Lemma

Assume that each r_k is L_k-Lipschitz continuous, then $0 \le r(x) - \hat{r}(x) \le \frac{\gamma\bar{L}^2}{2}$, where $\bar{L}^2 = \sum_{k=1}^K w_k L_k^2$.

- Conclusion. As the stepsize γ gets smaller, $\hat{r}(x)$ would be closer to $r(x)$.

4. Our methods

Alternative:

$$\min_{x\in\mathbb{R}^d}\hat{F}(x) = f(x) + \hat{r}(x), \tag{7}$$

in which r is replaced by its proximal average $\hat{r}$. Then the iteration becomes

$$\begin{aligned}x^{k+1} &= \mathrm{prox}_{\hat{r}}^{\gamma}(x^k - \gamma v^k)\\ &= \sum_{k=1}^{K} w_k\cdot\mathrm{prox}_{r_k}^{\gamma}(x^k - \gamma v^k).\end{aligned}$$

We need to decrease γ adaptively.

APA-SVRG

- ADA-SVRG Algorithm

```
Initialize: An initial number of inner loops m_0 > 0, decay rate 0 < ρ < 1, and an initial point x̃_0.
for s = 1, 2, ···, do
    x^0 = x̃_{s-1}, ṽ = Σ_{i=1}^n f_i(x̃_{s-1})/n;
    m_s = m_0 · ρ^{-s};
    γ_s = min{1/4L, ρ^s};
    for l = 1, 2, ···, m_s do
        Randomly pick j from {1, 2, ..., n};
        v^l = ∇f_j(x^{l-1}) − ∇f_j(x̃_{s-1}) + ṽ;
        x^l = Σ_{k=1}^K w_k · prox_{r_k}^{γ_s}(x^{l-1} − γ_s v^l);
    end for
    x̃_s = Σ_{l=1}^{m_s} x^l / n.
end for
```

- Theorem

Theorem 1 (APA-SVRG)**.** Suppose that L-smoothness, μ-strong convexity and L_k-Lipschitz continuous regularisers assumptions hold. Then for the update in APA-SVRG, it holds that

$$\begin{aligned}&\mathbb{E}F(\tilde{x}_s) - F^*\\ \le\ &\theta^s(\hat{F}_0(\tilde{x}_0) - F^*) + \frac{\gamma_0}{2}\bar{L}^2\frac{\theta}{\theta-\rho}(\theta^s - \rho^s).\end{aligned}$$

- Remarks

· When $\rho = 1$, i.e. the stepsize is fixed, $\mathbb{E}F(\tilde{x}_{s+1})$ will not converge to the minimum value.
· When $0<\rho<1$, $F(\tilde{x}_{s+1}) - F^*$ approaches 0 at the exponential rate.

- Complexity

Corollary 1. To achieve the ϵ-accurate solution, the overall iteration complexity of APA-SVRG is $\sum_{s=0}^{S}\mathcal{O}(n+2m_s) = \mathcal{O}(nS + \sum_{s=0}^{S} m_s) = \mathcal{O}(n\log\frac{1}{\epsilon} + m_0\frac{1}{\epsilon})$.

APA-SAGA

- APA-SAGA Algorithm

```
Initialize: An initial number of inner loops m_0 > 0, decay rate 0 < ρ < 1, an initial point x^0, and g_i^0 = ∇f(x^0), i = 1, 2, ..., n.
for s = 1, 2, ···, do
    m = m_0 · ρ^{-s};
    γ_s = (1/3L) · ρ^s;
    x^0 = x_s;
    for l = 1, ···, m do
        Randomly pick j from {1, 2, ..., n};
        v^l = ∇f_j(x^{l-1}) − g_j^l + Σ_{i=1}^n g_i^l / n;
        x^l = Σ_{k=1}^K w_k · prox_{r_k}^{γ_s}(x^{l-1} − γ_s v^l);
        Update g_i^l, i = 1, 2, ..., n:
            g_i^l = { ∇f_j(x^{l-1}),  if i = j,
                      g_i^{l-1},      otherwise.
    end for
    x_s = x^m.
end for
```

- Complexity

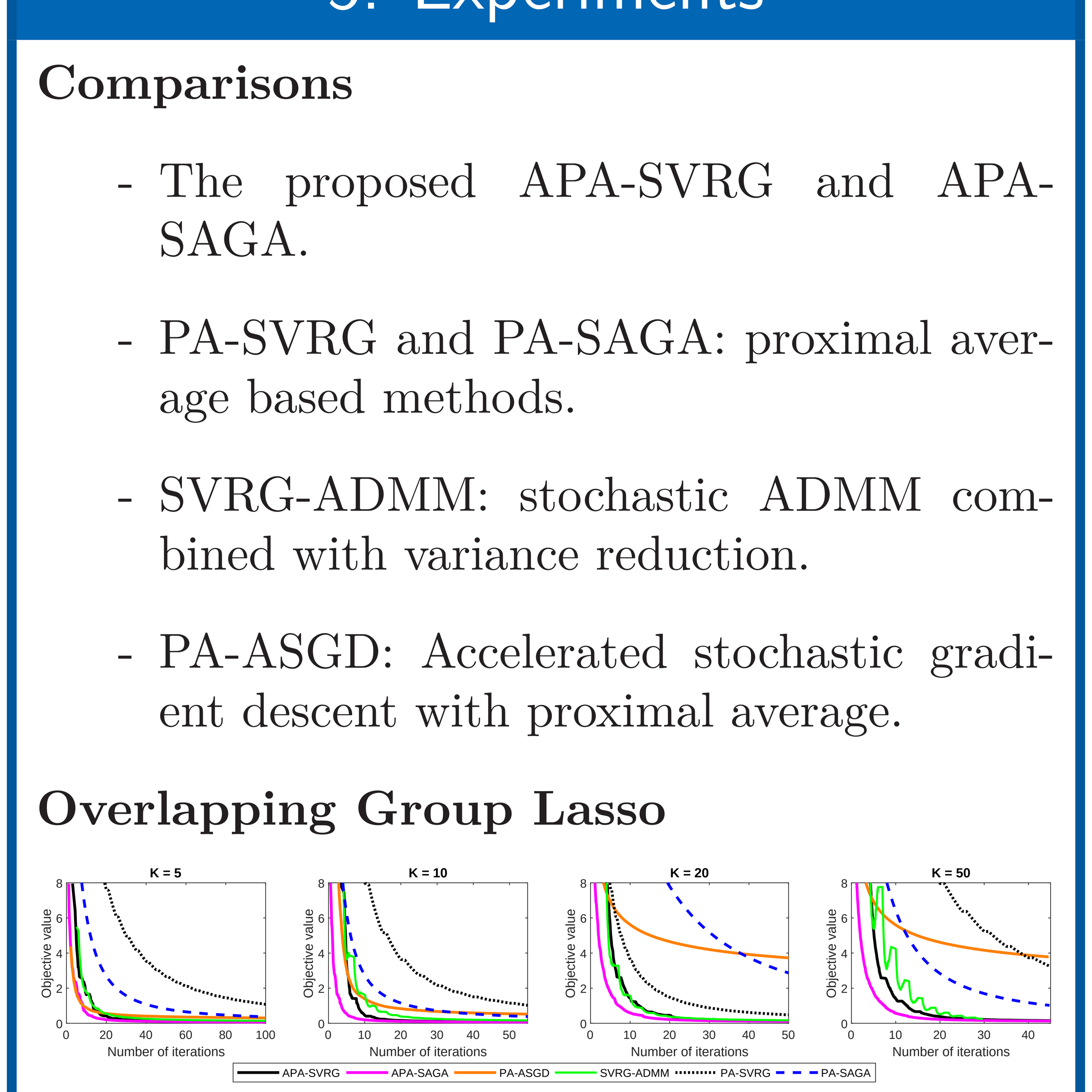
Corollary 2. To achieve the ϵ-accurate solution, the overall iteration complexity of APA-SAGA is $\mathcal{O}(n\log\frac{1}{\epsilon} + m_0\frac{1}{\epsilon})$.

5. Experiments

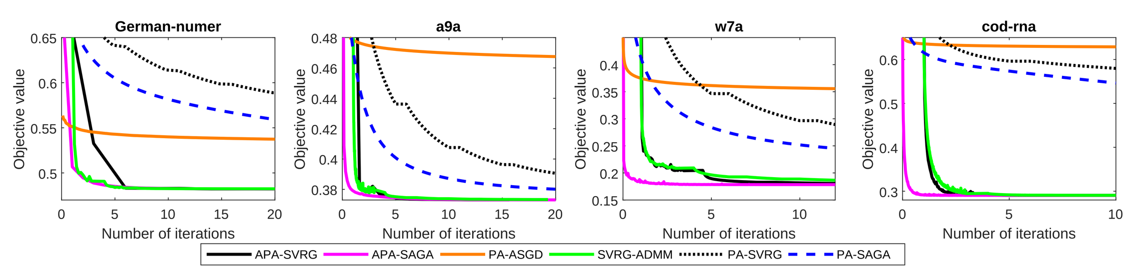
Comparisons

- The proposed APA-SVRG and APA-SAGA.
- PA-SVRG and PA-SAGA: proximal average based methods.
- SVRG-ADMM: stochastic ADMM combined with variance reduction.
- PA-ASGD: Accelerated stochastic gradient descent with proximal average.

Overlapping Group Lasso



Graph-Guided Logistic Regression



thankyou