SDCA without Duality, Regularization, and Individual Convexity

Shai Shalev-Shwartz

School of CS and Engineering,
The Hebrew University of Jerusalem

"ICML 2016",
New York, June 2016

Minimizing Average-of-Functions

Question: What is the runtime to find w s.t. $F(w) \leq F(w^*) + \epsilon$ where

$$F(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

Minimizing Average-of-Functions

Question: What is the runtime to find w s.t. $F(w) \le F(w^*) + \epsilon$ where

$$F(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

Classic: **Gradient Descent (GD):**

- Assume: F is λ-strongly convex and L-smooth
- Runtime: $d \cdot \left(n \cdot \frac{L}{\lambda}\right) \cdot \log\left(\frac{1}{\epsilon}\right)$

Minimizing Average-of-Functions

Question: What is the runtime to find w s.t. $F(w) \leq F(w^*) + \epsilon$ where

$$F(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

Classic: **Gradient Descent (GD):**

- Assume: F is λ-strongly convex and L-smooth
- Runtime: $d \cdot \left(n \cdot \frac{L}{\lambda}\right) \cdot \log\left(\frac{1}{\epsilon}\right)$

Modern: **Stochastic Dual Coordinate Ascent (SDCA):**

- Assume: $f_i(w) = \phi_i(w) + \frac{\lambda}{2}\|w\|^2$ and ϕ_i is convex and L smooth
- Runtime: $d \cdot \left(n + \frac{L}{\lambda}\right) \cdot \log\left(\frac{1}{\epsilon}\right)$

Outline

1. SDCA without Duality
   - SDCA $\in$ SGD family
   - SGD with a stochastic oracle must be slow
   - SDCA reduces the variance using a stronger oracle
   - A simple convergence proof
2. Relaxing the Assumptions
   - Without Explicit Regularization
   - Dependence on Average Smoothness
   - Without Individual Convexity

SDCA without Duality

- Objective:

$$F(w) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} \phi_i(w)$$

- At w^* we have $\nabla F(w^*) = 0$:

$$w^* = \frac{1}{\lambda n} \sum_{i=1}^{n} \underbrace{(-\nabla \phi_i(w^*))}_{:=\alpha_i^*}$$

- Primal variable: w
- Pseudo-Dual variables: $\alpha_1, \ldots, \alpha_n$
- Goal: $w^{(t)} \to w^*$ and for every i, $\alpha_i^{(t)} \to \alpha_i^*$

SDCA without Duality

- **Initialize:** $w^{(0)} = \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i^{(0)}$
- **For:** $t = 1, 2, \ldots$
  - Pick $i \in [n]$ at random
  - Primal update: $w^{(t)} = w^{(t-1)} - \eta \left(\nabla \phi_i(w^{(t-1)}) + \alpha_i^{(t-1)} \right)$
  - Dual update: $\alpha_i^{(t)} = (1-\beta)\alpha_i^{(t-1)} + \beta \left(-\nabla \phi_i(w^{(t-1)}) \right)$ (where $\beta = \eta \lambda n$)

SDCA without Duality

- **Initialize:** $w^{(0)} = \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i^{(0)}$
- **For:** $t = 1, 2, \ldots$
  - Pick $i \in [n]$ at random
  - Primal update: $w^{(t)} = w^{(t-1)} - \eta \left(\nabla \phi_i(w^{(t-1)}) + \alpha_i^{(t-1)} \right)$
  - Dual update: $\alpha_i^{(t)} = (1-\beta)\alpha_i^{(t-1)} + \beta \left(-\nabla \phi_i(w^{(t-1)}) \right)$
    (where $\beta = \eta \lambda n$)

- Claim: SDCA is an instance of SGD
- Proof:
  - By induction, $\lambda\, w^{(t-1)} = \frac{1}{n} \sum_{i=1}^{n} \alpha_i^{(t-1)} = \mathbb{E}_i\, \alpha_i^{(t-1)}$
  - Therefore:
    $\nabla F(w^{(t-1)}) = \lambda\, w^{(t-1)} + \mathbb{E}_i\, \nabla \phi_i(w^{(t-1)}) = \mathbb{E}_i[\alpha_i^{(t-1)} + \nabla \phi_i(w^{(t-1)})]$

SGD with a stochastic oracle must be slow

Theorem

Any algorithm for minimizing F that only accesses the objective using oracle that returns a gradient of a random function and has $\log(1/\epsilon)$ rate must perform $\tilde{\Omega}(n^2)$ iterations

SGD with a stochastic oracle must be slow

Theorem

Any algorithm for minimizing F that only accesses the objective using oracle that returns a gradient of a random function and has $\log(1/\epsilon)$ rate must perform $\tilde{\Omega}(n^2)$ iterations

Proof idea:

- Consider two objectives (in both, $\lambda = 1$): for $i \in \{\pm 1\}$

$$F_i(w) = \frac{1}{2n}\left((n-1)\frac{(w-i)^2}{2} + (n+1)\frac{(w+i)^2}{2}\right)$$

- A stochastic gradient oracle returns $w \pm i$ w.p. $\frac{1}{2} \pm \frac{1}{2n}$
- Easy to see that $w_i^* = -i/n$, $F_i(0) = 1/2$, $F_i(w_i^*) = 1/2 - 1/(2n^2)$
- Therefore, solving to accuracy $\epsilon < 1/(2n^2)$ amounts to determining the bias of the coin

Can we improve SGD ?

A stronger oracle:

- The negative result assumes we only see a gradient of a randomly chosen example
- SDCA relies on a slightly stronger oracle: we also see the index of the chosen example
- This suffices to obtain a significantly faster algorithm
- Main idea: variance reduction

Variance Reduction

- SGD update rule: $w^{(t)} = w^{(t-1)} - \eta v$ where $\mathbb{E}[v] = \nabla F(w^{(t-1)})$

Variance Reduction

- SGD update rule: $w^{(t)} = w^{(t-1)} - \eta v$ where $\mathbb{E}[v] = \nabla F(w^{(t-1)})$
- For vanilla SGD: $v = \nabla \phi_i(w^{(t-1)}) + \lambda w^{(t-1)}$

Variance Reduction

- SGD update rule: $w^{(t)} = w^{(t-1)} - \eta v$ where $\mathbb{E}[v] = \nabla F(w^{(t-1)})$
- For vanilla SGD: $v = \nabla \phi_i(w^{(t-1)}) + \lambda w^{(t-1)}$
- For SDCA: $v = \nabla \phi_i(w^{(t-1)}) + \alpha_i^{(t-1)}$

Variance Reduction

- SGD update rule: $w^{(t)} = w^{(t-1)} - \eta v$ where $\mathbb{E}[v] = \nabla F(w^{(t-1)})$
- For vanilla SGD: $v = \nabla \phi_i(w^{(t-1)}) + \lambda w^{(t-1)}$
- For SDCA: $v = \nabla \phi_i(w^{(t-1)}) + \alpha_i^{(t-1)}$
- What is the variance?

$$\begin{aligned}
\mathbb{E}[\|v\|^2] &= \mathbb{E}[\|\nabla \phi_i(w^{(t-1)}) - \nabla \phi_i(w^*) + \nabla \phi_i(w^*) + \alpha_i^{(t-1)}\|^2] \\
&= \mathbb{E}[\|\nabla \phi_i(w^{(t-1)}) - \nabla \phi_i(w^*) + \alpha_i^{(t-1)} - \alpha_i^*\|^2] \\
&\le 2\,\mathbb{E}[\|\nabla \phi_i(w^{(t-1)}) - \nabla \phi_i(w^*)\|^2] + 2\,\mathbb{E}[\|\alpha_i^{(t-1)} - \alpha_i^*\|^2] \\
&\le 2L\,\mathbb{E}[\|w^{(t-1)} - w^*\|^2] + 2\,\mathbb{E}[\|\alpha_i^{(t-1)} - \alpha_i^*\|^2]
\end{aligned}$$

A simple convergence proof

- Potential: $C_t = \frac{1}{2L} A_t + \frac{\lambda}{2} B_t$ with $A_t = \mathbb{E}_j \|\alpha_j^{(t)} - \alpha_j^*\|^2$, $B_t = \|w^{(t)} - w^*\|^2$

A simple convergence proof

- Potential: $C_t = \frac{1}{2L} A_t + \frac{\lambda}{2} B_t$ with $A_t = \mathbb{E}_j \|\alpha_j^{(t)} - \alpha_j^*\|^2$, $B_t = \|w^{(t)} - w^*\|^2$
- Algebraic Manipulations:

$$\mathbb{E}\, A_t - (1-\eta\lambda) A_{t-1} = \eta\lambda\, \mathbb{E}\left(\|\nabla\phi_i(w^{(t-1)}) - \nabla\phi_i(w^*)\|^2 - (1-\beta)\|v\|^2 \right)$$

$$\mathbb{E}\, B_t - B_{t-1} = -2\eta(w^{(t-1)} - w^*)^\top \nabla F(w^{(t-1)}) + \eta^2\, \mathbb{E}\,\|v\|^2$$

A simple convergence proof

- Potential: $C_t = \frac{1}{2L} A_t + \frac{\lambda}{2} B_t$ with $A_t = \mathbb{E}_j \|\alpha_j^{(t)} - \alpha_j^*\|^2$, $B_t = \|w^{(t)} - w^*\|^2$
- Algebraic Manipulations:

$$\mathbb{E}\, A_t - (1-\eta\lambda) A_{t-1} = \eta\lambda\, \mathbb{E}\left(\|\nabla\phi_i(w^{(t-1)}) - \nabla\phi_i(w^*)\|^2 - (1-\beta)\|v\|^2 \right)$$

$$\mathbb{E}\, B_t - B_{t-1} = -2\eta (w^{(t-1)} - w^*)^\top \nabla F(w^{(t-1)}) + \eta^2\, \mathbb{E}\, \|v\|^2$$

- S.c. of F, $((w^{(t-1)} - w^*)^\top \nabla F(w^{(t-1)}) \geq \epsilon_{t-1} + \frac{\lambda}{2} B_{t-1})$, gives

$$\mathbb{E}\, B_t - (1-\eta\lambda) B_{t-1} \;\leq\; -2\eta\epsilon_{t-1} + \eta^2\, \mathbb{E}\, \|v\|^2$$

A simple convergence proof

- Potential: $C_t = \frac{1}{2L} A_t + \frac{\lambda}{2} B_t$ with $A_t = \mathbb{E}_j \|\alpha_j^{(t)} - \alpha_j^*\|^2$, $B_t = \|w^{(t)} - w^*\|^2$
- Algebraic Manipulations:

$$\mathbb{E}\, A_t - (1-\eta\lambda) A_{t-1} = \eta\lambda\, \mathbb{E}\left(\|\nabla \phi_i(w^{(t-1)}) - \nabla \phi_i(w^*)\|^2 - (1-\beta)\|v\|^2 \right)$$

$$\mathbb{E}\, B_t - B_{t-1} = -2\eta (w^{(t-1)} - w^*)^\top \nabla F(w^{(t-1)}) + \eta^2\, \mathbb{E}\, \|v\|^2$$

- S.c. of F, $((w^{(t-1)} - w^*)^\top \nabla F(w^{(t-1)}) \geq \epsilon_{t-1} + \frac{\lambda}{2} B_{t-1})$, gives

$$\mathbb{E}\, B_t - (1-\eta\lambda) B_{t-1} \;\leq\; -2\eta\epsilon_{t-1} + \eta^2\, \mathbb{E}\, \|v\|^2$$

- Summing with weights $(\frac{1}{2L}, \frac{\lambda}{2})$ cancels the $\mathbb{E}\,\|v\|^2$ term and gives

$$C_t - (1-\eta\lambda) C_{t-1} \;\leq\; \eta\lambda \left(\frac{1}{2L} \mathop{\mathbb{E}}_i \|\nabla \phi_i(w^{(t-1)}) - \nabla \phi_i(w^*)\|^2 - \epsilon_{t-1} \right)$$

A simple convergence proof

- Potential: $C_t = \frac{1}{2L} A_t + \frac{\lambda}{2} B_t$ with $A_t = \mathbb{E}_j \|\alpha_j^{(t)} - \alpha_j^*\|^2$, $B_t = \|w^{(t)} - w^*\|^2$
- Algebraic Manipulations:

$$\mathbb{E}\, A_t - (1-\eta\lambda) A_{t-1} = \eta\lambda\, \mathbb{E}\left(\|\nabla\phi_i(w^{(t-1)}) - \nabla\phi_i(w^*)\|^2 - (1-\beta)\|v\|^2 \right)$$

$$\mathbb{E}\, B_t - B_{t-1} = -2\eta (w^{(t-1)} - w^*)^\top \nabla F(w^{(t-1)}) + \eta^2\, \mathbb{E}\, \|v\|^2$$

- S.c. of F, $((w^{(t-1)} - w^*)^\top \nabla F(w^{(t-1)}) \ge \epsilon_{t-1} + \frac{\lambda}{2} B_{t-1})$, gives

$$\mathbb{E}\, B_t - (1-\eta\lambda) B_{t-1} \;\le\; -2\eta\epsilon_{t-1} + \eta^2\, \mathbb{E}\, \|v\|^2$$

- Summing with weights $(\frac{1}{2L}, \frac{\lambda}{2})$ cancels the $\mathbb{E}\, \|v\|^2$ term and gives

$$C_t - (1-\eta\lambda) C_{t-1} \;\le\; \eta\lambda \left(\frac{1}{2L} \mathop{\mathbb{E}}_i \|\nabla\phi_i(w^{(t-1)}) - \nabla\phi_i(w^*)\|^2 - \epsilon_{t-1} \right)$$

- ϕ_i is L-smooth and convex $\Rightarrow$ $\mathbb{E}_i \|\nabla\phi_i(w^{(t-1)}) - \nabla\phi_i(w^*)\|^2 \le 2L\epsilon_{t-1}$

A simple convergence proof

- Potential: $C_t = \frac{1}{2L} A_t + \frac{\lambda}{2} B_t$ with $A_t = \mathbb{E}_j \, \|\alpha_j^{(t)} - \alpha_j^*\|^2$, $B_t = \|w^{(t)} - w^*\|^2$
- Algebraic Manipulations:

$$\mathbb{E}\, A_t - (1-\eta\lambda) A_{t-1} = \eta\lambda \, \mathbb{E} \left(\|\nabla\phi_i(w^{(t-1)}) - \nabla\phi_i(w^*)\|^2 - (1-\beta)\|v\|^2 \right)$$

$$\mathbb{E}\, B_t - B_{t-1} = -2\eta (w^{(t-1)} - w^*)^\top \nabla F(w^{(t-1)}) + \eta^2 \, \mathbb{E}\, \|v\|^2$$

- S.c. of F, $((w^{(t-1)} - w^*)^\top \nabla F(w^{(t-1)}) \geq \epsilon_{t-1} + \frac{\lambda}{2} B_{t-1})$, gives

$$\mathbb{E}\, B_t - (1-\eta\lambda) B_{t-1} \;\leq\; -2\eta\epsilon_{t-1} + \eta^2 \, \mathbb{E}\, \|v\|^2$$

- Summing with weights $(\frac{1}{2L}, \frac{\lambda}{2})$ cancels the $\mathbb{E}\, \|v\|^2$ term and gives

$$C_t - (1-\eta\lambda) C_{t-1} \;\leq\; \eta\lambda \left(\frac{1}{2L} \mathop{\mathbb{E}}_i \|\nabla\phi_i(w^{(t-1)}) - \nabla\phi_i(w^*)\|^2 - \epsilon_{t-1} \right)$$

- ϕ_i is L-smooth and convex $\Rightarrow \;\; \mathbb{E}_i \, \|\nabla\phi_i(w^{(t-1)}) - \nabla\phi_i(w^*)\|^2 \leq 2L\epsilon_{t-1}$
- Therefore: $\mathbb{E}\, C_t \leq (1-\eta\lambda)\, \mathbb{E}\, C_{t-1} \leq (1-\eta\lambda)^t \, C_0$

Outline

1. SDCA without Duality
   - SDCA $\in$ SGD family
   - SGD with a stochastic oracle must be slow
   - SDCA reduces the variance using a stronger oracle
   - A simple convergence proof
2. Relaxing the Assumptions
   - Without Explicit Regularization
   - Dependence on Average Smoothness
   - Without Individual Convexity

GD vs SDCA

Classic: **Gradient Descent (GD):**

- Assume: F is λ-strongly convex and L-smooth
- Runtime: $d \cdot \left(n \cdot \frac{L}{\lambda}\right) \cdot \log\left(\frac{1}{\epsilon}\right)$

Modern: **Stochastic Dual Coordinate Ascent (SDCA):**

- Assume: $f_i(w) = \phi_i(w) + \frac{\lambda}{2}\|w\|^2$ and ϕ_i is convex and L smooth
- Runtime: $d \cdot \left(n + \frac{L}{\lambda}\right) \cdot \log\left(\frac{1}{\epsilon}\right)$

SDCA Without Explicit Regularization

Original objective:

$$F(w) = \frac{1}{n} \sum_{i=1}^{n} \phi_i(w)$$

Rewrite the objective as

$$F(w) = \frac{1}{n+1} \sum_{i=1}^{n+1} \phi_i(w) + \frac{\lambda}{2} \|w\|^2$$

where

- For $i \leq n$, $\phi_i(w) = \frac{n+1}{n} f_i(w)$
- $\phi_{n+1}(w) = \frac{-\lambda(n+1)}{2} \|w\|^2$

Dependence on Average Smoothness

- Assume that ϕ_i is L_i-smooth
- Let $\bar{L} = \mathbb{E}_i L_i$
- Sample $i \sim q$ where $q_i = \frac{L_i + \bar{L}}{2n\bar{L}}$
- Convergence rate now depends on $\bar{L}$ instead of on $\max_i L_i$

SDCA without Individual Convexity

- Remove the assumption that ϕ_i is convex and only assume that F is λ-strongly convex
- The bound $\mathbb{E}_i \|\nabla\phi_i(w^{(t-1)}) - \nabla\phi_i(w^*)\|^2 \leq 2L\epsilon_{t-1}$ no longer holds
- Instead, $\mathbb{E}_i \|\nabla\phi_i(w^{(t-1)}) - \nabla\phi_i(w^*)\|^2 \leq L^2\|(w^{(t-1)} - w^*\|^2$
- Yields a runtime of $d \cdot \left(n + \left(\frac{L}{\lambda}\right)^2\right) \cdot \log\left(\frac{1}{\epsilon}\right)$
- Using acceleration gives runtime of $\tilde{O}\left(d \cdot (n + n^{3/4}\sqrt{\bar{L}/\lambda})\right)$
- Compare to the convex case: $\tilde{O}\left(d \cdot (n + n^{1/2}\sqrt{\bar{L}/\lambda})\right)$

Summary

- SDCA without duality as a variance reduced SGD
- Simpler proof
- Relaxing the assumptions on individual functions
- Open: is the extra $n^{1/4}$ factor necessary ?