Stochastic Optimization with Importance Sampling

Peilin Zhao† Tong Zhang‡

†Institute for Infocomm Research, A*STAR, Singapore
‡Rutgers University, USA; and Baidu Research, China

Jun 20, 2015
Outline

1. **Background Review**
   - Regularized Loss Minimization
   - Related Work
   - Motivation
   - Challenges and Our Contributions

2. **Stochastic Optimization with Importance Sampling**
   - Iprox-SMD
   - Iprox-SDCA

3. **Experiments**
   - Application
   - Experimental Datasets
   - Settings
   - Experimental Results

4. **Conclusion**
Regularized Loss Minimization

**Primal Objective**

\[
\min_{\bf w \in \mathbb{R}^d} P(\bf w) := \frac{1}{n} \sum_{i=1}^{n} \phi_i(\bf w) + \lambda r(\bf w),
\]

\[
\phi_i(\bf w) = f(\bf w)
\]

**Examples**

- SVM: \( \phi_i(\bf w) = \max(0, 1 - y_i x_i^\top \bf w) \), \( r(\bf w) = \frac{1}{2} \| \bf w \|_2^2 \);
- LASSO: \( \phi_i(\bf w) = (y_i - x_i^\top \bf w)^2 \), \( r(\bf w) = \| \bf w \|_1 \)

**Goal**

Find a \( \epsilon_P \)-sub-optimal \( \bf w \), i.e., \( P(\bf w) - P(\bf w^*) \leq \epsilon_P \)
### Related Work

#### Classical Algorithms
- SGD: Stochastic Gradient Descent
- prox-SGD: Proximal Stochastic Gradient Descent
- prox-SMD: Proximal Stochastic Mirror Descent

#### Recent Algorithms
- SAG: Stochastic Average Gradient
- SVRG: Stochastic Variance Reduction Gradient
- SDCA: Stochastic Dual Coordinate Ascent
- prox-SDCA: Proximal Stochastic Dual Coordinate Ascent
Motivation

The Limitations of Existing Algorithms
- uniform sampling of training data
- the sampled stochastic quantity may suffer from rather high variances
- the convergence rate is negatively affected

Stochastic Optimization with Importance Sampling
- will adopt a data-dependent sampling distribution to minimize the variance,
- the convergence rate will significantly improved under certain conditions
Challenges and Our Contributions

**Key Challenge**
- analyze the relation between a sampling distribution and the variance, or
- analyze the relation between a sampling distribution and the convergence

**Two Algorithms**
- lprox-SMD: prox-SMD with importance sampling
- lprox-SDCA: prox-SDCA with importance sampling
**prox-SMD**

- $i_t$ will be uniformly drawn from $[n] = \{1, 2, \ldots, n\}$
- $w^{t+1} = \arg\min_w \left[ \langle \nabla \phi_i(w^t), w \rangle + \lambda r(w) + \frac{1}{\eta_t} B_{\psi}(w, w^t) \right]$  

Example: when $\psi(w) = \frac{1}{2} \|w\|_2^2$, this algorithm is prox-SGD.

Disadvantage: the randomness introduces variance, caused by $E[\nabla \phi_i(w^t)] = \nabla f(w^t)$, but $\nabla \phi_i(w^t)$ varies with $i$

**Iprox-SMD**

- assign each $i \in \{1, \ldots, n\}$ a probability $p_i^t \geq 0$, s.t. $\sum_{i=1}^n p_i^t = 1$
- sample $i_t$ from $\{1, \ldots, n\}$ based on $p^t = (p_1^t, \ldots, p_n^t)^\top$
- $w^{t+1} = \arg\min_w \left[ \langle \frac{\nabla \phi_i(w^t)}{np_i^t}, w \rangle + \lambda r(w) + \frac{1}{\eta_t} B_{\psi}(w, w^t) \right],$

another prox-SMD, since $E[(np_i^t)^{-1} \nabla \phi_i(w^t) \mid w^t] = \nabla f(w^t)$.  

**Lemma 1: Relationship between p^t and the convergence rate of Iprox-SMD**

Define \( w^{t+1} \) by the update of Iprox-SMD. Assume that \( \psi(\cdot) \) is \( \sigma \)-strongly convex with respect to a norm \( \| \cdot \| \) (its dual norm is \( \| \cdot \|_* \)), and \( f \) is \( \mu \)-strongly convex and \((1/\gamma)\)-smooth with respect to \( \psi \). If \( r(w) \) is convex and \( \eta_t \in (0, \gamma] \), then \( w^{t+1} \) satisfies the following inequality for any \( t \geq 1 \),

\[
\mathbb{E}[P(w^{t+1}) - P(w^*)] \\
\leq \frac{1}{\eta_t} \mathbb{E}[B_\psi(w^*, w^t) - B_\psi(w^*, w^{t+1})] - \mu \mathbb{E}B_\psi(w^*, w^t) + \frac{\eta_t}{\sigma} \mathbb{V} \left( (np^t_i)^{-1} \nabla \phi_i(w^t) \right),
\]

where \( \mathbb{V}((np^t_i)^{-1} \nabla \phi_i(w^t)) = \mathbb{E}\| (np^t_i)^{-1} \nabla \phi_i(w^t) - \nabla f(w^t) \|_*^2 \), and the expectation is taken with the distribution \( p^t \).

**Optimal distribution**

\[
\min_{p^t \in \Delta^n} \mathbb{V} \left( \frac{\nabla \phi_i(w^t)}{np^t_i} \right) \Leftrightarrow \min_{p^t \in \Delta^n} \frac{1}{n^2} \sum_{i=1}^n \frac{\| \nabla \phi_i(w^t) \|_*^2}{p^t_i} \Rightarrow p^t_i = \frac{\| \nabla \phi_i(w^t) \|_*}{\sum_{j=1}^n \| \nabla \phi_j(w^t) \|_*}. \tag{1}
\]

Disadvantages: need to compute \( \nabla \phi_i(w^t) \) for all \( i \in [n] \).
Relaxation

\[ G_i \geq \|\nabla \phi_i(w^t)\|_*, \quad \forall t \Rightarrow \min_{p^t \in \Delta^n} \frac{1}{n^2} \sum_{i=1}^{n} \frac{\|\nabla \phi_i(w^t)\|^2_\ast}{p^t_i} \leq \min_{p^t \in \Delta^n} \frac{1}{n^2} \sum_{i=1}^{n} \frac{G_i^2}{p^t_i} \tag{2} \]

So, an suboptimal distribution independent of \( t \):

\[ p^t_i = \frac{G_i}{\sum_{j=1}^{n} G_j}, \quad \forall i \in \{1, 2, \ldots, n\}, \]

Two Distributions

- \( \phi_i(w) \) is \( L_i \)-Lipschitz, i.e., \( \|\nabla \phi_i(w)\|_\ast \leq L_i \), then \( p^t_i = \frac{L_i}{\sum_{j=1}^{n} L_j} \)

- \( \phi_i(w) \) is \( (1/\gamma_i) \)-smooth and \( \|w^t\| \leq R \) for any \( t \), then \( \|\nabla \phi_i(w^t)\|_\ast \leq R/\gamma_i \),

\[ p^t_i = \frac{1}{\gamma_i} \frac{1}{\sum_{j=1}^{n} \frac{1}{\gamma_j}} \]
Iprox-SMD

**Input:** $\lambda \geq 0$, the learning rates $\eta_1, \ldots, \eta_T > 0$.

**Initialize:** $w^1 = 0$, $p_i = \frac{L_i}{\sum_j L_j}$ or $p_i = \frac{1/\gamma_i}{\sum_j 1/\gamma_j}$, $\forall i$.

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

Sample $i_t$ from $\{1, \ldots, n\}$ based on $p$;

\[
w^{t+1} = \arg\min_w \left[ \langle (np_{i_t})^{-1} \nabla \phi_{i_t}(w^t), w \rangle + \lambda r(w) + \frac{1}{\eta_t} B_\psi(w, w^t) \right];
\]

**end for**
Theorem 1: Convergence Rate of Iprox-SMD for Smooth Losses

Assume that $\psi(\cdot)$ is $\sigma$-strongly convex with respect to a norm $\| \cdot \|$, $f$ is $\mu$-strongly convex and $(1/\gamma)$-smooth with respect to $\psi$, $r(w)$ is convex and $\eta_t = \frac{1}{\alpha + \mu t}$ with $\alpha \geq 1/\gamma - \mu$. If we further assume $\phi_i(w)$ is $(1/\gamma_i)$-smooth, $\|w^t\| \leq R$ for any $t$, and the distribution is set as $p_t^i = \frac{R/\gamma_i}{\sum_{j=1}^n R/\gamma_j}$, then the following inequality holds for any $T \geq 1$,

$$
\frac{1}{T} \sum_{t=1}^T \mathbb{E} P(w^{t+1}) - P(w^*) \leq O \left[ \frac{\left( \sum_{i=1}^n R/\gamma_i \right)^2 \ln(\alpha + \mu T)}{\sigma \mu n^2 T} \right].
$$

In addition, if $\mu = 0$, and $\eta_t$ is set as $\sqrt{\sigma B_\psi(w^*, w^1) / (\sqrt{T} \sum_{i=1}^n R/\gamma_i)}$, we have,

$$
\frac{1}{T} \sum_{t=1}^T \mathbb{E} P(w^{t+1}) - P(w^*) \leq 2 \sqrt{\frac{B_\psi(w^*, w^1)}{\sigma} \sum_{i=1}^n R/\gamma_i} \frac{1}{\sqrt{T}}.
$$

Remark. If uniform distribution is adopted, $\frac{\left( \sum_{i=1}^n R/\gamma_i \right)^2}{n^2}$ in the bound should be replaced with $\frac{\sum_{i=1}^n (R/\gamma_i)^2}{n} \cdot \frac{\sum_{i=1}^n (R/\gamma_i)^2}{n} / \left( \frac{\sum_{i=1}^n R/\gamma_i}{n} \right)^2 = \frac{n \sum_{i=1}^n (R/\gamma_i)^2}{(\sum_{i=1}^n R/\gamma_i)^2} \geq 1$ implies the effectiveness of importance sampling, especially when $\frac{(\sum_{i=1}^n R/\gamma_i)^2}{\sum_{i=1}^n (R/\gamma_i)^2} \ll n$. 
Theorem 2: Convergence Rate of Iprox-SMD for Lipschitz Losses

Assume that $\psi(\cdot)$ is $\sigma$-strongly convex with respect to a norm $\| \cdot \|$, $f$ and $r(w)$ are convex, and $\eta_t = \eta$. If $\phi_i(w)$ is $L_i$-Lipschitz, and the distribution is set as $p_i = L_i / \sum_{j=1}^{n} L_j$, $\forall i$, then when $\eta_t$ is set as $\sqrt{2\sigma B_\psi(w^*, w^1)} / (\sum_{i=1}^{n} L_i \sqrt{T})$, the following inequality holds for any $T \geq 1$,

$$\frac{1}{T} \sum_{t=1}^{T} E[\mathcal{P}(w^t) - \mathcal{P}(w^*)] \leq \sqrt{B_\psi(w^*, w^1)} \frac{2}{\sigma} \left( \frac{\sum_{i=1}^{n} L_i}{n} \right) \frac{1}{\sqrt{T}}.$$

Remark: If uniform distribution is adopted, $(\sum_{i=1}^{n} L_i^2)^2$ should be replaced with $\sum_{i=1}^{n} L_i^2$. However, according to the Cauchy-Schwarz inequality,

$$\frac{(\sum_{i=1}^{n} L_i^2)/n}{(\sum_{i=1}^{n} L_i)^2/n^2} \geq 1,$$

implies the effectiveness of importance sampling, especially when $\frac{(\sum_{i=1}^{n} L_i)^2}{\sum_{i=1}^{n} (L_i)^2} \ll n.$
Iprox-SDCA

### Dual Problem

\[
\max_{\theta} D(\theta) := \frac{1}{n} \sum_{i=1}^{n} -\phi_i^*(-\theta_i) - \lambda r^*(\frac{1}{\lambda n} \sum_{i=1}^{n} \theta_i).
\]

- \( w = \nabla r^*(v(\theta)) \), \( v(\theta) = \frac{1}{\lambda n} \sum_{i=1}^{n} \theta_i \).
- \( r(w) \) is 1-strongly convex, i.e., \( r(w + \Delta w) \geq r(w) + \nabla r(w)^\top \Delta w + \frac{1}{2} \|\Delta w\|_P^2 \).
- \( r^*(w) \) is 1-smooth, \( r^*(v + \Delta v) \leq h(v; \Delta v) := r^*(v) + \nabla r^*(v)^\top \Delta v + \frac{1}{2} \|\Delta v\|_{D'}^2 \).

### prox-SDCA

picks \( i \in \{1, \ldots, n\} \) uniformly at random, then:

\[
\theta_i^t = \theta_i^{t-1} + \Delta \theta_i^{t-1}, \quad \max_{\Delta \theta_i} \left[ -\frac{1}{n} \phi_i^*(-(\theta_i^{t-1} + \Delta \theta_i)) - \lambda h(v^{t-1}; \frac{1}{\lambda n} \Delta \theta_i) \right],
\]

where \( v^{t-1} = \frac{1}{\lambda n} \sum_{i=1}^{n} \theta_i^{t-1} \), equivalent to maximizing a lower bound of

\[
\max_{\Delta \theta_i} \left[ -\frac{1}{n} \phi_i^*(-(\theta_i^{t-1} + \Delta \theta_i)) - \lambda r^*(v^{t-1} + \frac{1}{\lambda n} \Delta \theta_i) \right].
\]
Iprox-SDCA:

- randomly pick $i$ according to probability $p_i$, where $p \in \mathbb{R}_+^n$, $\sum_i p_i = 1$.
- $\theta_i$ is updated as traditional prox-SDCA.

Lemma 2: relationship between $p$ and the convergence rate of Iprox-SDCA

Given a distribution $p$, if assume $\phi_i$ is $(1/\gamma_i)$-smooth with norm $\| \cdot \|_P$, then for any iteration $t$ and any $s$ such that $s_i = s/(p_i n) \in [0, 1]$, $\forall i$, we have

$$
\mathbb{E}[D(\theta^t) - D(\theta^{t-1})] \geq \frac{s}{n} \mathbb{E}[P(w^{t-1}) - D(\theta^{t-1})] - \frac{sG^t}{2\lambda n^2},
$$

where $G^t = \frac{1}{n} \sum_{i=1}^n (s_i R^2 - \gamma_i (1 - s_i) \lambda n) \mathbb{E}\|u_i^{t-1} - \theta_i^{t-1}\|_D^2$, $R = \sup_{u \neq 0} \|u\|_{D'} / \|u\|_D$, and $-u_i^{t-1} \in \partial \phi_i(w^{t-1})$.

Optimal Distribution

$$
\max_{s/(p_i n) \in [0, 1], p \in \Delta^n} \frac{s}{n} \mathbb{E}[P(w^{t-1}) - D(\theta^{t-1})] - \frac{s}{n^2} \frac{G^t}{2\lambda}.
$$

where $\Delta^n$ is the $n$-dimensional simplex.
**Relaxation: when \( \phi_i \) is \( 1/\gamma_i \)-smooth**

\[
\max_{s/n \in [0,1], p \in \Delta^n} \frac{s}{n} \mathbb{E}[P(w^{t-1}) - D(\theta^{t-1})] - \frac{s}{n^2} \frac{G^t}{2}\lambda \\
\geq \max_{s/n \in [0, \frac{\lambda n \gamma_i}{R^2 + \lambda n \gamma_i}], p \in \Delta^n} \frac{s}{n} \mathbb{E}[P(w^{t-1}) - D(\theta^{t-1})] - \frac{s}{n^2} \frac{G^t}{2}\lambda \\
\geq \max_{s/n \in [0, \frac{\lambda n \gamma_i}{R^2 + \lambda n \gamma_i}], p \in \Delta^n} \frac{s}{n} \mathbb{E}[P(w^{t-1}) - D(\theta^{t-1})].
\]

where \( G^t = \frac{1}{n} \sum_{i=1}^{n} (s_i R^2 - \gamma_i (1 - s_i) \lambda n) \mathbb{E} \| u_i^{t-1} - \theta_i^{t-1} \|^2_D \leq 0 \) is used for the second inequality.

**Distribution for smooth losses**

\[
\max_{s/n \in [0, \frac{\lambda n \gamma_i}{R^2 + \lambda n \gamma_i}], p \in \Delta^n} s \Rightarrow p_i = \frac{1 + \frac{R^2}{\lambda n \gamma_i}}{n + \sum_{j=1}^{n} \frac{R^2}{\lambda n \gamma_j}}. \quad (5)
\]
Relaxation: when $\phi_i$ is $L_i$-Lipschitz

Combining lemma 2 with

$$P(w^{t-1}) - D(\theta^{t-1}) \geq D(\theta^*) - D(\theta^{t-1}) := \epsilon^{t-1},$$

gives

$$\mathbb{E}[\epsilon^t] \leq (1 - \frac{s}{n})\mathbb{E}[\epsilon^{t-1}] + \frac{s}{2\lambda n^2} \frac{4R^2s}{n^2} \sum_{i=1}^{n} \frac{1}{p_i} L_i^2. \quad (6)$$

Distribution for Lipschitz losses

$$\min_{p \in \Delta^n} \sum_{i=1}^{n} \frac{1}{p_i} L_i^2 \Rightarrow p_i = L_i / \sum_{j=1}^{n} L_j.$$
Iprox-SDCA

**Input:** $\lambda > 0$, $R = \sup_{\mathbf{u} \neq 0} \|\mathbf{u}\|_{D'} / \|\mathbf{u}\|_{D}$, norms $\| \cdot \|_{D}$, $\| \cdot \|_{D'}$, $\gamma_1, \ldots, \gamma_n > 0$, or $L_1, \ldots, L_n \geq 0$.

**Initialize:** $\theta_0 = 0$, $w^0 = \nabla r^*(0)$, $p_i = \frac{1 + \frac{R^2}{\lambda n \gamma_i}}{n + \sum_{j=1}^{n} \frac{R^2}{\lambda n j}}$, or $p_i = \frac{L_i}{\sum_{j=1}^{n} L_j}$, $\forall i \in \{1, \ldots, n\}$.

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

Sample $i_t$ from $\{1, \ldots, n\}$ based on $p$;

$$\Delta \theta^{t-1}_{i_t} = \arg \max_{\Delta \theta_{i_t}} \left[ -\phi^*_{i_t}(- (\theta^{t-1}_{i_t} + \Delta \theta_{i_t} ) ) - (w^{t-1})^\top \Delta \theta_{i_t} - \frac{1}{2 \lambda n} \| \Delta \theta_{i_t} \|_{D'}^2 \right];$$

$$\theta^t_{i_t} = \theta^{t-1}_{i_t} + \Delta \theta^{t-1}_{i_t};$$

$$v^t = v^{t-1} + \frac{1}{\lambda n} \Delta \theta^{t-1}_{i_t};$$

$$w^t = \nabla r^*(v^t);$$

**end for**
Theorem 3: convergence rate of Iprox-SDCA for smooth losses

Assume $\phi_i$ is $(1/\gamma_i)$-smooth $\forall i \in \{1, \ldots, n\}$ and set $p_i = (1 + \frac{R^2}{\lambda n \gamma_i})/(n + \sum_{j=1}^{n} \frac{R^2}{\lambda n \gamma_j})$, for all $i \in \{1, \ldots, n\}$. To obtain an expected duality gap of $\mathbb{E}[P(w^t) - D(\theta^T)] \leq \epsilon_P$ for the proposed Proximal SDCA with importance sampling, it suffices to have a total number of iterations of

$$T \geq (n + \sum_{i=1}^{n} \frac{R^2}{\lambda n \gamma_i}) \log \left( (n + \sum_{i=1}^{n} \frac{R^2}{\lambda n \gamma_i}) \frac{1}{\epsilon_P} \right).$$

Remark: If employ uniform sampling,

$$T \geq (n + \frac{R^2}{\lambda \gamma_{\min}}) \log \left( (n + \frac{R^2}{\lambda \gamma_{\min}}) \frac{1}{\epsilon_P} \right)$$

where $\gamma_{\min} = \min\{\gamma_1, \ldots, \gamma_n\}$. Since

$$\frac{n + \frac{R^2}{\lambda \gamma_{\min}}}{n + \sum_{i=1}^{n} \frac{R^2}{\lambda \gamma_i n}} = \frac{n \lambda \gamma_{\min} + \frac{R^2}{n} \sum_{i=1}^{n} \frac{\gamma_{\min}}{\gamma_i}}{n \lambda \gamma_{\min} + \frac{R^2}{n} \sum_{i=1}^{n} \frac{\gamma_{\min}}{\gamma_i}} \geq 1,$$

the bound for importance sampling is always better, especially when $\sum_{i=1}^{n} \frac{\gamma_{\min}}{\gamma_i} \ll n$. 
Theorem 4: convergence rate of Iprox-SDCA for Lipschitz losses

Consider the proposed proximal SDCA with importance sampling. Assume that \( \phi_i \) is \( L_i \)-Lipschitz and set \( p_i = L_i / \sum_{j=1}^{n} L_j \), \( \forall i \in \{1, \ldots, n\} \). To obtain an expected duality gap of \( \mathbb{E}[P(\tilde{w}) - D(\tilde{\theta})] \leq \epsilon_P \) where \( \tilde{w} = \frac{1}{T-T_0} \sum_{t=T_0+1}^{T} w^{t-1} \) and \( \tilde{\theta} = \frac{1}{T-T_0} \sum_{t=T_0+1}^{T} \theta^{t-1} \), it suffices to have a total number of iterations of

\[
T \geq T_0 + n/\rho + \frac{4R^2(\sum_{i=1}^{n} L_i)^2}{n^2 \lambda \epsilon_P} \geq \omega + n/\rho + \frac{20R^2(\sum_{i=1}^{n} L_i)^2}{n^2 \lambda \epsilon_P},
\]

where \( \omega = \max(0, \left\lceil \frac{n}{\rho} \log(\frac{\lambda n}{\rho 2R^2(\sum_{i=1}^{n} L_i)^2/n^2}) \right\rceil) \), and \( \rho = \frac{\sum_{i=1}^{n} L_i}{n L_{\min}} \). Moreover, when \( t \geq T_0 \), we have dual sub-optimality bound of \( \mathbb{E}[D(\theta^*) - D(\theta^t)] \leq \epsilon_P / 2 \).

**Remark:** If adopt uniform sampling

\[
T \geq \max(0, 2\left\lceil n \log(\frac{\lambda n}{2R^2 L_{\max}^2}) \right\rceil) - n + \frac{20R^2(L_{\max})^2}{\lambda \epsilon_P}
\]

where \( L_{\max} = \max\{L_1, \ldots, L_n\} \). However, the ratio of the leading terms is

\[
\frac{(L_{\max})^2}{(\sum_{i=1}^{n} L_i)^2/n^2} = \left( \frac{n}{\sum_{i=1}^{n} L_i/L_{\max}} \right)^2 \geq 1,
\]

which again implies the effectiveness of importance sampling.
squared hinge loss based SVM with \( \ell_2 \) regularization

\[
\min_w \frac{1}{n} \sum_{i=1}^{n} \left( [1 - y_i w^\top x_i]_+ \right)^2 + \frac{\lambda}{2} \|w\|_2^2
\]

**Iprox-SGD:** \( \{w \in \mathbb{R}^d \|w\|_2 \leq 1/\sqrt{\lambda}\} \)

- \( \phi_i(w) = ([1 - y_i w^\top x_i]_+)^2 + \frac{\lambda}{2} \|w\|_2^2, \quad r(w) = 0 \)
- \( \nabla \phi_i(w) = -2[1 - y_i w^\top x_i]_+ y_i x_i + \lambda w \)
- \( p_i = \frac{2(1+\|x_i\|_2/\sqrt{\lambda})\|x_i\|_2 + \sqrt{\lambda}}{\sum_{j=1}^{n} [2(1+\|x_j\|_2/\sqrt{\lambda})\|x_j\|_2 + \sqrt{\lambda}]} \)

**Iprox-SDCA:**

- \( \phi_i(w) = ([1 - y_i w^\top x_i]_+)^2, \quad r(w) = \frac{1}{2} \|w\|_2^2 \)
- \( p_i = (1 + \frac{2\|x_i\|_2^2}{\lambda n})/(n + \sum_{j=1}^{n} \frac{2\|x_j\|_2^2}{\lambda n}), \quad \text{since } \phi_i \text{ is } (2\|x_i\|_2^2)-\text{smooth w.r.t. } \| \cdot \|_2 \).
- \( \Delta \theta_i = \max \left( \frac{1-y_i w^\top x_i - \alpha_i/2}{1/2 + \|x_i\|_2^2/\lambda n}, -\alpha_i \right) y_i x_i. \)
Experimental Datasets

**Table**: Datasets used in the experiments.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Dataset Size</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>ijcnn1</td>
<td>49990</td>
<td>22</td>
</tr>
<tr>
<td>kdd2010(algebra)</td>
<td>8407752</td>
<td>20216830</td>
</tr>
<tr>
<td>w8a</td>
<td>49749</td>
<td>300</td>
</tr>
</tbody>
</table>

All the datasets are downloaded from LIBSVM website \(^1\).

**Table**: Theoretical Constant Ratios for The Datasets.

<table>
<thead>
<tr>
<th>Constant Ratio</th>
<th>ijcnn1</th>
<th>kdd2010</th>
<th>w8a</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{n \sum_{i=1}^{n} (G_i)^2}{(\sum_{i=1}^{n} G_i)^2} )</td>
<td>1.0643</td>
<td>1.4667</td>
<td>1.9236</td>
</tr>
<tr>
<td>( \frac{n \lambda \gamma_{min} + R^2}{n \lambda \gamma_{min} + \frac{R^2}{n} \sum_{i=1}^{n} \frac{\gamma_{min}}{\gamma_i}} )</td>
<td>1.1262</td>
<td>1.1404</td>
<td>1.3467</td>
</tr>
</tbody>
</table>

\(^1\)http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

- for SGD: importance sampling is effective on kdd2010 and w8a, except ijcnn1
- for SDCA: importance sampling accelerates SDCA for all the datasets
### Settings

- Experiments are conducted by fixing five different random seeds for each dataset.
- The reported results are averaged over these five runs.
- Uniform sampling is adopted at the first epoch for Iprox-SGD and Iprox-SDCA.

### Measures

- Measuring the primal objective value \( P(\mathbf{w}^t) \) for SGD.
- Measuring the duality gap \( P(\mathbf{w}^t) - D(\theta^t) \) for SDCA.
- Evaluated the test error rates.
- Report the variances of the stochastic gradients.
Experimental Results

**Figure:** Comparison between Pegasos with Iprox-SGD on “ijcnn”, “kdd”, “w8a”.

- on “kdd”, “w8a”, Iprox-SGD achieves the fastest convergence rates and significantly smaller test error rates
- on “kdd”, “w8a”, Iprox-SGD enjoys much smaller variances
- on “ijcnn”, Iprox-SGD degenerates into the traditional prox-SGD
**Experimental Results**

![Graphs showing comparison between SDCA and Iprox-SDCA](image)

**Figure:** Comparison between SDCA with Iprox-SDCA on “ijcnn”, “kdd”, “w8a”.

- Iprox-SDCA converges faster than SDCA
- Test error rates of Iprox-SDCA is comparable with SDCA
- Iprox-SDCA enjoys slightly smaller variances, however the improvement is not large enough to significantly reduce the test error
Conclusion

- studies stochastic optimization with importance sampling
- proposed two specific algorithms:
  (i) Iprox-SMD: prox-SMD with importance sampling
  (ii) Iprox-SDCA: prox-SDCA with importance sampling
- analyze the theoretical convergence rates
- validate their empirical efficacy
Thank you!

Any question?