Stochastic Optimization

Lijun Zhang

Nanjing University, China

May 26, 2017

Outline

LAMDA Learning And Mining from DatA

Outline

Definitions

- Stochastic Optimization [Nemirovski et al., 2009]

$$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{\xi}\left[f(\mathbf{w},\xi)\right] = \int_{\Xi} f(\mathbf{w},\xi)dP(\xi)$$

  - ξ is a random variable
  - The challenge: evaluation of the expectation/integration

Definitions

- Stochastic Optimization [Nemirovski et al., 2009]

$$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{\xi}\left[f(\mathbf{w},\xi)\right] = \int_{\Xi} f(\mathbf{w},\xi)dP(\xi)$$

  - ξ is a random variable
  - The challenge: evaluation of the expectation/integration

- Supervised Learning [Vapnik, 1998]

$$\min_{h\in\mathcal{H}} F(h) = \mathrm{E}_{(\mathbf{x},y)}\left[\ell(h(\mathbf{x}),y)\right]$$

  - $(\mathbf{x},y)$ is a random instance-label pair
  - $\mathcal{H} = \{h : \mathcal{X} \mapsto \mathbb{R}\}$ is a hypothesis class
  - $\ell(\cdot,\cdot)$ is certain loss, e.g., hinge loss
    $\ell(u,v) = \max(0, 1-uv)$

Optimization Methods

- Sample Average Approximation (SAA)
- Empirical Risk Minimization (RRM)
  - $$\min_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w},\xi_i)$$
    where $\xi_1,\ldots,\xi_n$ are i.i.d.
  - $$\min_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(\mathbf{x}_i),y_i))$$
    where $(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)$ are i.i.d.

Optimization Methods

- Sample Average Approximation (SAA)
- Empirical Risk Minimization (RRM)
  - $$\min_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w}, \xi_i)$$

    where $\xi_1, \ldots, \xi_n$ are i.i.d.
  - $$\min_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i))$$

    where $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are i.i.d.
- Stochastic Approximation (SA)
  - Optimization via noisy observations of $F(\cdot)$
  - Zero-order SA [Nesterov, 2011]
  - First-order SA, e.g., SGD [Kushner and Yin, 2003]

Performance

- Optimization Error / Excess Risk

$$F(\bar{\mathbf{w}}) - \min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \begin{cases} O\left(\frac{1}{\sqrt{n}}\right), O\left(\frac{1}{n}\right), O\left(\frac{1}{n^2}\right) \\ O\left(\frac{1}{\sqrt{T}}\right), O\left(\frac{1}{T}\right) \end{cases}$$

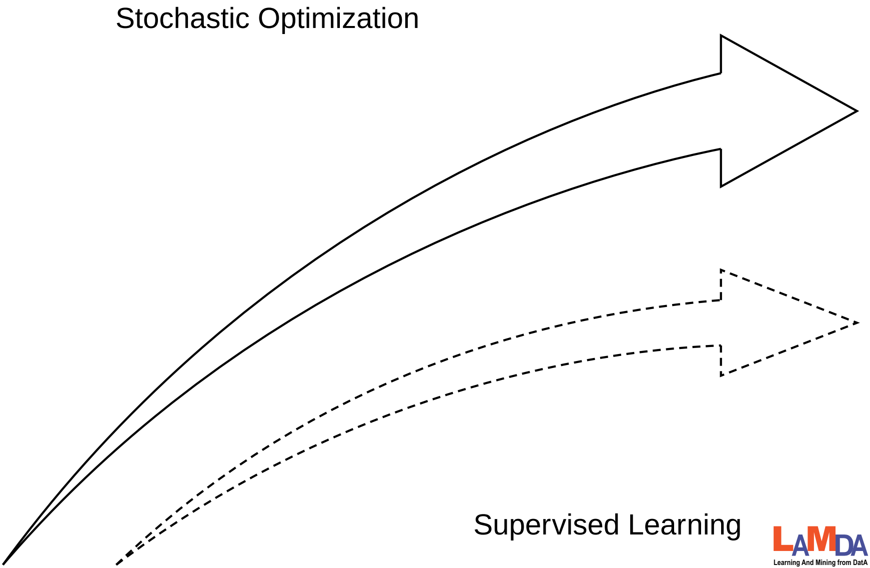
  - n is the number of samples in empirical risk minimization
  - T is the number of stochastic gradients in first-order stochastic approximation

Outline

Related Work

	Stochastic Optimization	Supervised Learning
Empirical Risk Minimization	[Shalev-Shwartz et al., 2009] [Koren and Levy, 2015] [Feldman, 2016] [Zhang et al., 2017]	[Vapnik, 1998] [Lee et al., 1996] [Panchenko, 2002] [Bartlett and Mendelson, 2002] [Tsybakov, 2004] [Bartlett et al., 2005] [Sridharan et al., 2009] [Srebro et al., 2010] . . .
Stochastic Approximation	[Zinkevich, 2003] [Hazan et al., 2007] [Hazan and Kale, 2011] [Rakhlin et al., 2012] [Lan, 2012] [Agarwal et al., 2012] [Shamir and Zhang, 2013] [Mahdavi et al., 2015] . . .	[Bach and Moulines, 2013]

Risk Bounds of Empirical Risk Minimization



Risk Bounds of Empirical Risk Minimization

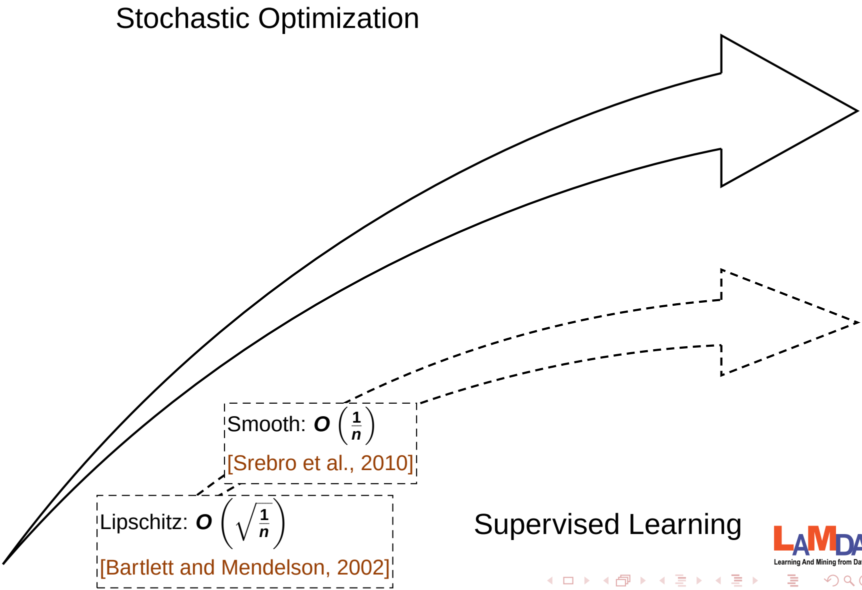
Stochastic Optimization

Lipschitz: $O\left(\sqrt{\frac{1}{n}}\right)$

[Bartlett and Mendelson, 2002]

Supervised Learning

Risk Bounds of Empirical Risk Minimization



Risk Bounds of Empirical Risk Minimization

Stochastic Optimization

Strongly Convex: $O\left(\frac{1}{\lambda n}\right)$
[Sridharan et al., 2009]

Smooth: $O\left(\frac{1}{n}\right)$
[Srebro et al., 2010]

Lipschitz: $O\left(\sqrt{\frac{1}{n}}\right)$
[Bartlett and Mendelson, 2002]

Supervised Learning

Risk Bounds of Empirical Risk Minimization

Stochastic Optimization

Strongly Convex & Smooth: $O\left(\frac{1}{\lambda n^2}\right)$

[Zhang et al., 2017]

Strongly Convex: $O\left(\frac{1}{\lambda n}\right)$

[Sridharan et al., 2009]

Smooth: $O\left(\frac{1}{n}\right)$

[Srebro et al., 2010]

Lipschitz: $O\left(\sqrt{\frac{1}{n}}\right)$

[Bartlett and Mendelson, 2002]

Supervised Learning

Learning And Mining from DatA

Risk Bounds of Empirical Risk Minimization

Stochastic Optimization

Strongly Convex & Smooth: $O\left(\frac{1}{\lambda n^2}\right)$

[Zhang et al., 2017]

Strongly Convex: $O\left(\frac{1}{\lambda n}\right)$

[Sridharan et al., 2009]

Lipschitz: $O\left(\sqrt{\frac{d}{n}}\right)$

[Shalev-Shwartz et al., 2009]

Smooth: $O\left(\frac{1}{n}\right)$

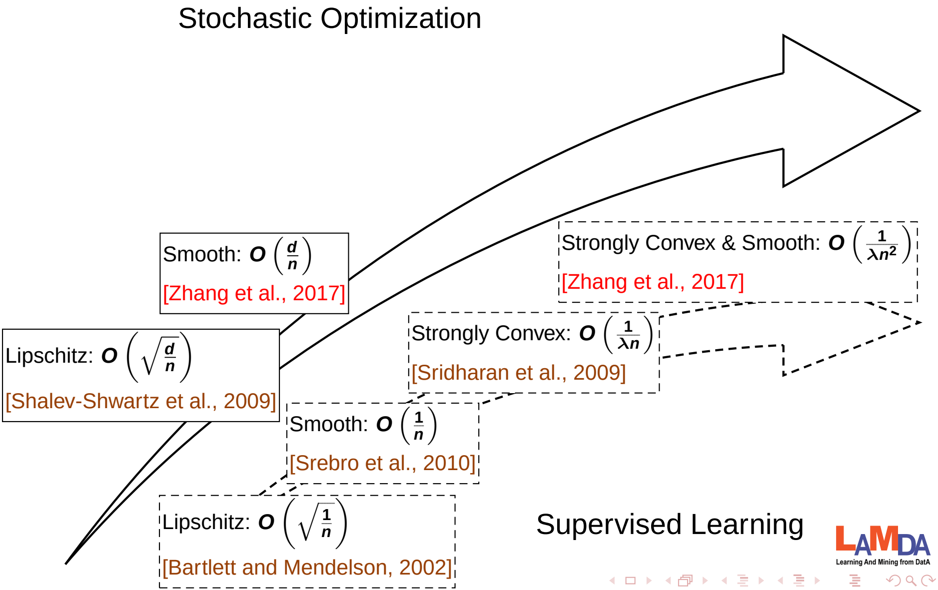
[Srebro et al., 2010]

Lipschitz: $O\left(\sqrt{\frac{1}{n}}\right)$

[Bartlett and Mendelson, 2002]

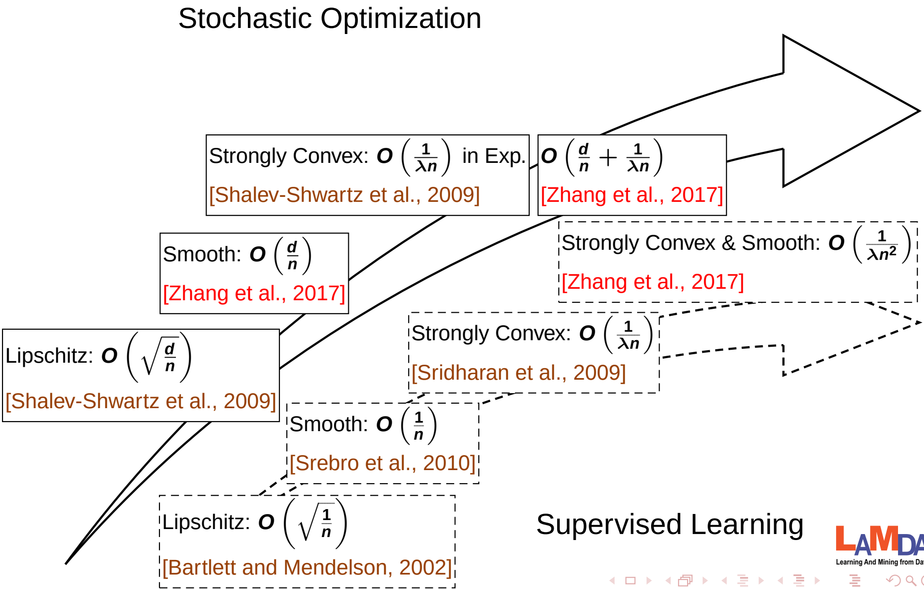
Supervised Learning

Risk Bounds of Empirical Risk Minimization



LAMDA Learning And Mining from DatA

Risk Bounds of Empirical Risk Minimization



Risk Bounds of Empirical Risk Minimization

Stochastic Optimization

Strongly Convex & Smooth: $O\left(\frac{1}{\lambda n^2}\right)$
[Zhang et al., 2017]

Strongly Convex: $O\left(\frac{1}{\lambda n}\right)$ in Exp.
[Shalev-Shwartz et al., 2009]

$O\left(\frac{d}{n} + \frac{1}{\lambda n}\right)$
[Zhang et al., 2017]

Smooth: $O\left(\frac{d}{n}\right)$
[Zhang et al., 2017]

Lipschitz: $O\left(\sqrt{\frac{d}{n}}\right)$
[Shalev-Shwartz et al., 2009]

Strongly Convex & Smooth: $O\left(\frac{1}{\lambda n^2}\right)$
[Zhang et al., 2017]

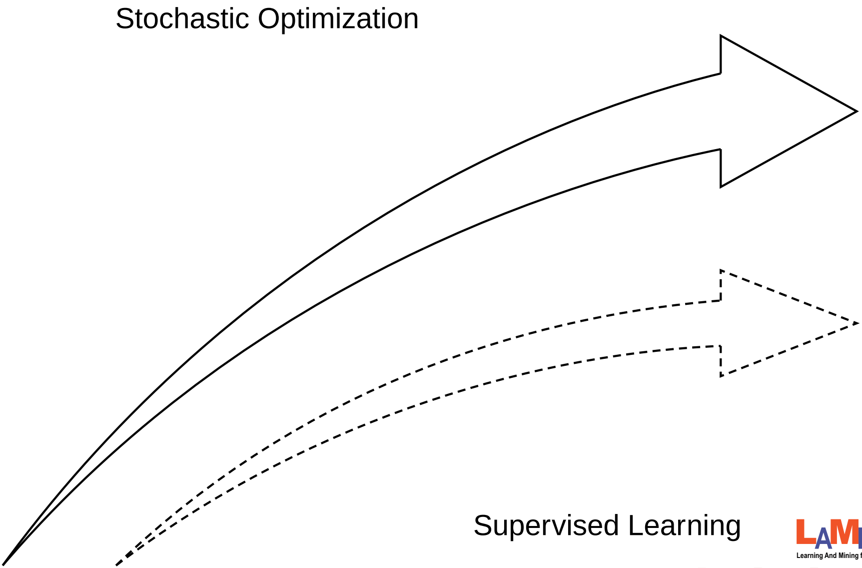
Strongly Convex: $O\left(\frac{1}{\lambda n}\right)$
[Sridharan et al., 2009]

Smooth: $O\left(\frac{1}{n}\right)$
[Srebro et al., 2010]

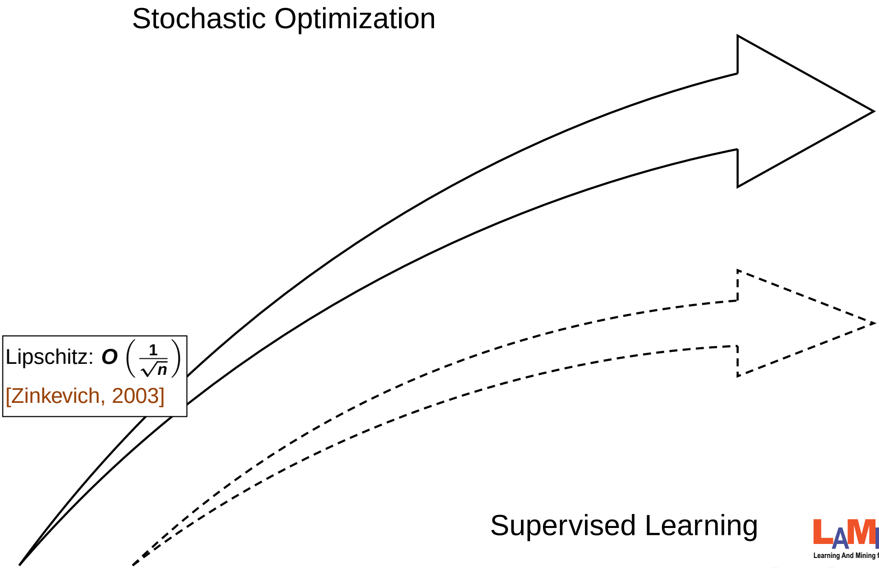
Lipschitz: $O\left(\sqrt{\frac{1}{n}}\right)$
[Bartlett and Mendelson, 2002]

Supervised Learning

Risk Bounds of Stochastic Approximation

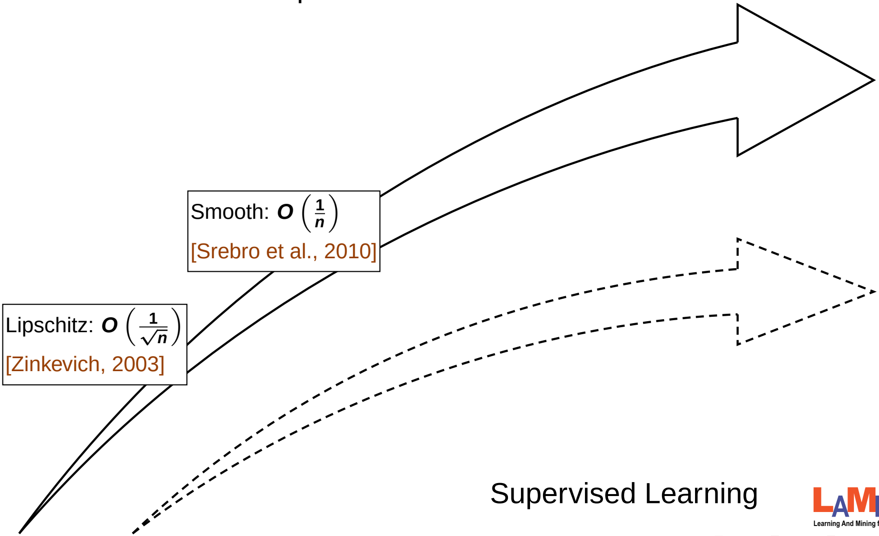


Risk Bounds of Stochastic Approximation



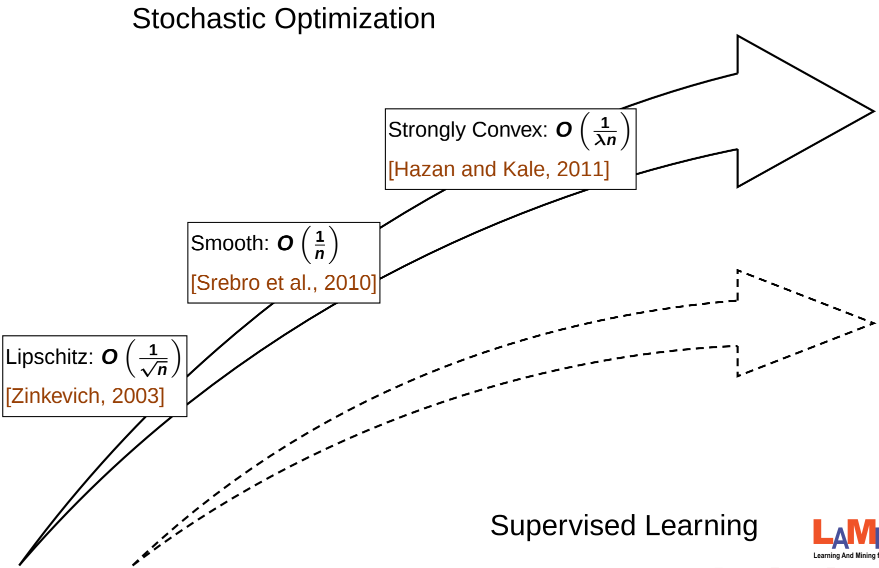
Risk Bounds of Stochastic Approximation

Stochastic Optimization

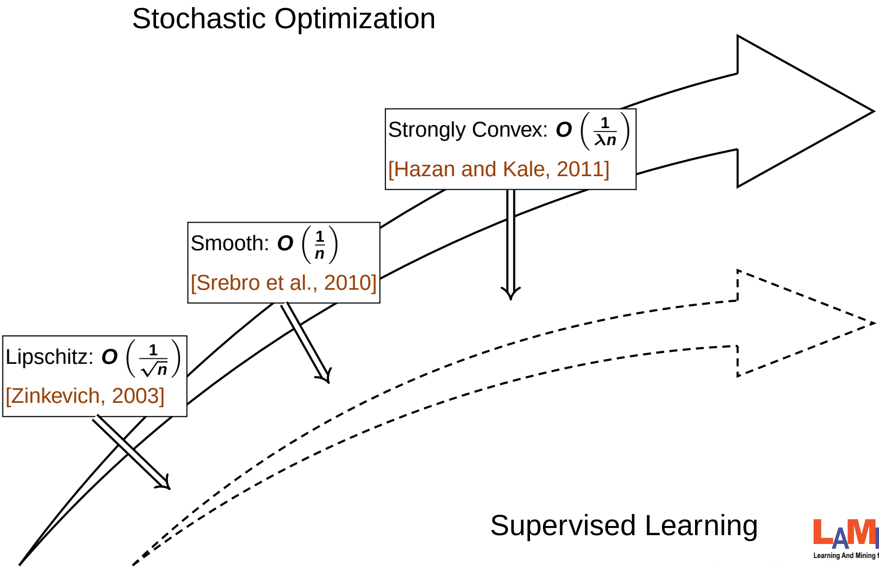


Supervised Learning

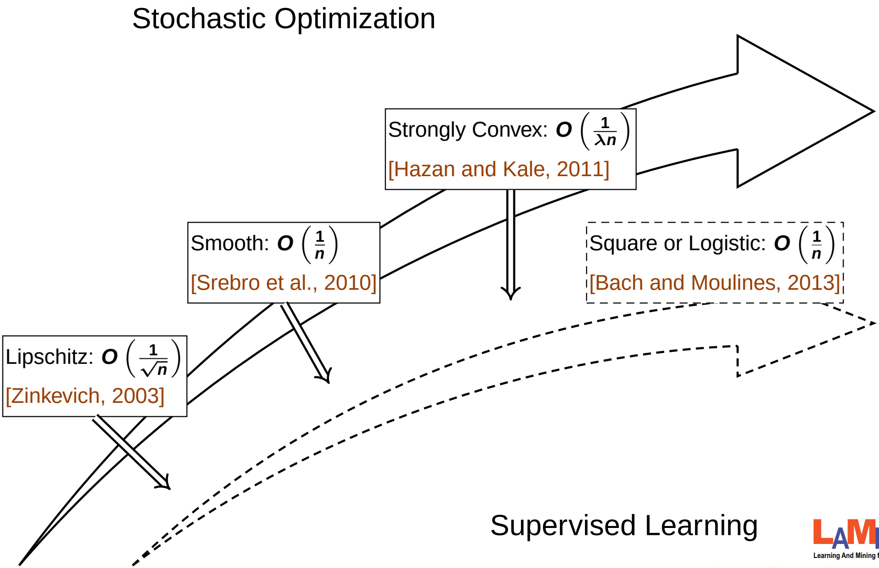
Risk Bounds of Stochastic Approximation



Risk Bounds of Stochastic Approximation



Risk Bounds of Stochastic Approximation



Outline

Stochastic Gradient Descent

- The Algorithm

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

2:
$$\mathbf{w}'_{t+1} = \mathbf{w}_t - \eta_t \mathbf{g}_t$$

3:
$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}(\mathbf{w}'_{t+1})$$

4: **end for**

5: **return** $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$

- Requirement

$$\mathrm{E}_t[\mathbf{g}_t] = \nabla F(\mathbf{w}_t)$$

where $\mathrm{E}_t[\cdot]$ is the conditional expectation

Stochastic Gradient Descent

- The Algorithm

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

2:
$$\mathbf{w}'_{t+1} = \mathbf{w}_t - \eta_t \mathbf{g}_t$$

3:
$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}(\mathbf{w}'_{t+1})$$

4: **end for**

5: **return** $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$

- Stochastic Optimization

$$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{\xi}\left[f(\mathbf{w}, \xi)\right]$$

  - Sample ξ_t from the underlying distribution
  - $\mathbf{g}_t = \nabla f(\mathbf{w}_t, \xi_t)$

Stochastic Gradient Descent

- The Algorithm

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

2:

$$\mathbf{w}'_{t+1} = \mathbf{w}_t - \eta_t \mathbf{g}_t$$

3:

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}(\mathbf{w}'_{t+1})$$

4: **end for**

5: **return** $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$

- Supervised Learning

$$\min_{h \in \mathcal{H}} F(h) = \mathrm{E}_{(\mathbf{x},y)}\left[\ell(h(\mathbf{x}), y)\right]$$

  - Sample $(\mathbf{x}_t, y_t)$ from the underlying distribution
  - $\mathbf{g}_t = \nabla \ell(h_t(\mathbf{x}_t), y_t)$

Outline

1. Introduction
2. Related Work
3. Stochastic Gradient Descent
   - Expectation Bound
   - High-probability Bound
4. Epoch Gradient Descent
   - Expectation Bound
   - High-probability Bound

Analysis I

For any $\mathbf{w} \in \mathcal{W}$, we have

$$
\begin{aligned}
& F(\mathbf{w}_t) - F(\mathbf{w}) \\
\leq & \langle \nabla F(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle \\
= & \langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{w} \rangle + \underbrace{\langle \nabla F(\mathbf{w}_t) - \mathbf{g}_t, \mathbf{w}_t - \mathbf{w} \rangle}_{\delta_t} \\
= & \frac{1}{\eta_t} \langle \mathbf{w}_t - \mathbf{w}'_{t+1}, \mathbf{w}_t - \mathbf{w} \rangle + \delta_t \\
= & \frac{1}{2\eta_t} \left(\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}'_{t+1} - \mathbf{w}\|_2^2 + \|\mathbf{w}_t - \mathbf{w}'_{t+1}\|_2^2 \right) + \delta_t \\
= & \frac{1}{2\eta_t} \left(\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}'_{t+1} - \mathbf{w}\|_2^2 \right) + \frac{\eta_t}{2} \|\mathbf{g}_t\|_2^2 + \delta_t \\
\leq & \frac{1}{2\eta_t} \left(\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 \right) + \frac{\eta_t}{2} \|\mathbf{g}_t\|_2^2 + \delta_t
\end{aligned}
$$

Analysis II

To simplify the above inequality, we assume

$$\eta_t = \eta, \text{ and } \|\mathbf{g}_t\|_2 \leq G$$

Then, we have

$$F(\mathbf{w}_t) - F(\mathbf{w}) \leq \frac{1}{2\eta}\left(\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2\right) + \frac{\eta}{2}G^2 + \delta_t$$

By adding the inequalities of all iterations, we have

$$\sum_{t=1}^{T} F(\mathbf{w}_t) - TF(\mathbf{w}) \leq \frac{1}{2\eta}\|\mathbf{w}_1 - \mathbf{w}\|_2^2 + \frac{\eta T}{2}G^2 + \sum_{t=1}^{T}\delta_t$$

Analysis III

Then, we have

$$
\begin{aligned}
& F(\bar{\mathbf{w}}_T) - F(\mathbf{w}) = F\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t\right) - F(\mathbf{w}) \\
\leq & \frac{1}{T}\sum_{t=1}^{T} F(\mathbf{w}_t) - F(\mathbf{w}) \leq \frac{1}{T}\left(\frac{1}{2\eta}\|\mathbf{w}_1 - \mathbf{w}\|_2^2 + \frac{\eta T}{2}G^2 + \sum_{t=1}^{T}\delta_t\right) \\
= & \frac{1}{2\eta T}\|\mathbf{w}_1 - \mathbf{w}\|_2^2 + \frac{\eta}{2}G^2 + \frac{1}{T}\sum_{t=1}^{T}\delta_t
\end{aligned}
$$

In summary, we have

$$
F(\bar{\mathbf{w}}_T) - F(\mathbf{w}) \leq \frac{1}{2\eta T}\|\mathbf{w}_1 - \mathbf{w}\|_2^2 + \frac{\eta}{2}G^2 + \frac{1}{T}\sum_{t=1}^{T}\delta_t
$$

Expectation Bound

Under the condition $\mathrm{E}_t[\mathbf{g}_t] = \nabla F(\mathbf{w}_t)$, it is easy to verify

$$\begin{aligned}&\mathrm{E}[\delta_t] = \mathrm{E}[\langle \nabla F(\mathbf{w}_t) - \mathbf{g}_t, \mathbf{w}_t - \mathbf{w}\rangle] \\ =&\mathrm{E}_{1,\ldots,t-1}[\mathrm{E}_t[\langle \nabla F(\mathbf{w}_t) - \mathbf{g}_t, \mathbf{w}_t - \mathbf{w}\rangle]] = 0\end{aligned}$$

Thus, by taking expectation over both sides, we have

$$\begin{aligned}&\mathrm{E}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}) \\ \leq&\frac{1}{2\eta T}\|\mathbf{w}_1 - \mathbf{w}\|_2^2 + \frac{\eta}{2}G^2 \\ \overset{\eta=\frac{c}{\sqrt{T}}}{=}&\frac{1}{2\sqrt{T}}\left(\frac{1}{c}\|\mathbf{w}_1 - \mathbf{w}\|_2^2 + cG^2\right)\end{aligned}$$

LAMDA Learning And Mining from DatA

Outline

Martingale

A sequence of random variables $Y_1, Y_2, Y_3, \ldots$ that satisfies for any time t

$$\mathrm{E}[|Y_t|] \leq \infty$$
$$\mathrm{E}[Y_{t+1}|Y_1, \ldots, Y_t] = Y_t$$

An Example

Y_t is a gambler's fortune after t tosses of a fair coin.

Let δ_t a random variable such that $\delta_t = 1$ if the coin comes up heads and $\delta_t = -1$ if the coin comes up tails. Then $Y_t = \sum_{i=1}^{t} \delta_t$.

$$\begin{aligned} \mathrm{E}[Y_{t+1}|Y_1, \ldots, Y_t] &= \mathrm{E}[\delta_{t+1} + Y_t|Y_1, \ldots, Y_t] \\ =& Y_t + \mathrm{E}[\delta_{t+1}|Y_1, \ldots, Y_t] = Y_t \end{aligned}$$

Martingale

A sequence of random variables $Y_1, Y_2, Y_3, \ldots$ that satisfies for any time t

$$\mathrm{E}[|Y_t|] \leq \infty$$
$$\mathrm{E}[Y_{t+1}|Y_1, \ldots, Y_t] = Y_t$$

An Example

Y_t is a gambler's fortune after t tosses of a fair coin.

Martingale Difference

Suppose $Y_1, Y_2, Y_3, \ldots$ is a martingale, then

$$X_t = Y_t - Y_{t-1}$$

is a martingale difference sequence.

$$\mathrm{E}[X_{t+1}|X_1, \ldots, X_t] = \mathrm{E}[Y_{t+1} - Y_t|X_1, \ldots, X_t] = 0$$

Azuma's inequality for Martingales

Suppose $X_1, X_2, \ldots$ is a martingale difference sequence and

$$|X_i| \leq c_i$$

almost surely. Then, we have

$$\Pr\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^{n} c_i^2}\right)$$

$$\Pr\left(\sum_{i=1}^{n} X_i \leq -t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^{n} c_i^2}\right)$$

Corollary 1

Assume $c_i \leq c$. With a probability at least $1 - 2\delta$, we have

$$\left|\sum_{i=1}^{n} X_i\right| \leq \sqrt{2nc \log \frac{1}{\delta}}$$

High-probability Bound I

Assume

$$\|\mathbf{g}_t\|_2 \leq G, \text{ and } \|\mathbf{x}-\mathbf{y}\|_2 \leq D, \ \forall \mathbf{x},\mathbf{y} \in \mathcal{W}$$

Under the condition $\mathrm{E}_t[\mathbf{g}_t] = \nabla F(\mathbf{w}_t)$, it is easy to verify

$$\delta_t = \langle \nabla F(\mathbf{w}_t) - \mathbf{g}_t, \mathbf{w}_t - \mathbf{w} \rangle$$

is a martingale difference sequence with

$$\begin{aligned}|\delta_t| \leq & \|\nabla F(\mathbf{w}_t) - \mathbf{g}_t\|_2 \|\mathbf{w}_t - \mathbf{w}\|_2 \\ \leq & D(\|\nabla F(\mathbf{w}_t)\|_2 + \|\mathbf{g}_t\|_2) \leq 2GD\end{aligned}$$

where we use Jensen's inequality

$$\|\nabla F(\mathbf{w}_t)\|_2 = \|\mathrm{E}[\mathbf{g}_t]\|_2 \leq \mathrm{E}\|\mathbf{g}_t\|_2 \leq G$$

High-probability Bound II

From Azuma's inequality, with a probability at least $1-\delta$

$$\sum_{t=1}^{T} \delta_t \leq 2\sqrt{TGD \log \frac{1}{\delta}}$$

Thus, with a probability at least $1-\delta$

$$\begin{aligned}
F(\bar{\mathbf{w}}_T) - F(\mathbf{w}) \leq & \frac{1}{2\eta T} \|\mathbf{w}_1 - \mathbf{w}\|_2^2 + \frac{\eta}{2} G^2 + \frac{1}{T} \sum_{t=1}^{T} \delta_t \\
\leq & \frac{1}{2\eta T} \|\mathbf{w}_1 - \mathbf{w}\|_2^2 + \frac{\eta}{2} G^2 + 2\sqrt{\frac{1}{T} GD \log \frac{1}{\delta}} \\
\overset{\eta = \frac{c}{\sqrt{T}}}{=} & \frac{1}{\sqrt{T}} \left(\frac{1}{2c} \|\mathbf{w}_1 - \mathbf{w}\|_2^2 + \frac{c}{2} G^2 + 2\sqrt{GD \log \frac{1}{\delta}} \right) \\
= & O\left(\frac{1}{\sqrt{T}} \right)
\end{aligned}$$

Outline

1. Introduction
2. Related Work
3. Stochastic Gradient Descent
   - Expectation Bound
   - High-probability Bound
4. Epoch Gradient Descent
   - Expectation Bound
   - High-probability Bound

Strong Convexity

A General Definition

F is λ-strongly convex if for all $\alpha \in [0,1]$, $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$,

$$F(\alpha \mathbf{w} + (1-\alpha)\mathbf{w}') \leq \alpha F(\mathbf{w}) + (1-\alpha)F(\mathbf{w}') - \frac{\lambda}{2}\alpha(1-\alpha)\|\mathbf{w} - \mathbf{w}'\|_2^2$$

A More Popular Definition

F is λ-strongly convex if for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$,

$$F(\mathbf{w}) + \langle \nabla F(\mathbf{w}), \mathbf{w}' - \mathbf{w} \rangle + \frac{\lambda}{2}\|\mathbf{w}' - \mathbf{w}\|_2^2 \leq F(\mathbf{w}')$$

Learning And Mining from DatA

Strong Convexity

A General Definition

F is λ-strongly convex if for all $\alpha \in [0,1]$, $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$,

$$F(\alpha\mathbf{w}+(1-\alpha)\mathbf{w}') \leq \alpha F(\mathbf{w})+(1-\alpha)F(\mathbf{w}')-\frac{\lambda}{2}\alpha(1-\alpha)\|\mathbf{w}-\mathbf{w}'\|_2^2$$

A More Popular Definition

F is λ-strongly convex if for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$,

$$F(\mathbf{w}) + \langle \nabla F(\mathbf{w}), \mathbf{w}' - \mathbf{w} \rangle + \frac{\lambda}{2}\|\mathbf{w}' - \mathbf{w}\|_2^2 \leq F(\mathbf{w}')$$

An Important Property

If F is λ-strongly convex, then

$$\frac{\lambda}{2}\|\mathbf{w} - \mathbf{w}_*\|^2 \leq F(\mathbf{w}) - F(\mathbf{w}_*)$$

where $\mathbf{w}_* = \operatorname{argmin}_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$.

Epoch Gradient Descent

- The Algorithm [Hazan and Kale, 2011]

$T_1 = 4$, $\eta_1 = 1/\lambda$
while $\sum_{i=1}^{k} T_k \leq T$ **do**
  for $t = 1, \ldots, T_k$ **do**

$$\mathbf{w}_{t+1}^k = \Pi_{\mathcal{W}}(\mathbf{w}_t^k - \eta_k \mathbf{g}_t^k)$$

  end for
  $\mathbf{w}_1^{k+1} = \frac{1}{T_k} \sum_{t=1}^{T_k} \mathbf{w}_t^k$
  $\eta_{k+1} = \eta_k/2$, $T_{k+1} = 2T_k$
  $k = k + 1$
end while

Epoch Gradient Descent

- The Algorithm [Hazan and Kale, 2011]

1: $T_1 = 4$, $\eta_1 = 1/\lambda$
2: **while** $\sum_{i=1}^{k} T_k \leq T$ **do**
3: **for** $t = 1, \ldots, T_k$ **do**
4:
$$\mathbf{w}_{t+1}^k = \Pi_{\mathcal{W}}(\mathbf{w}_t^k - \eta_k \mathbf{g}_t^k)$$
5: **end for**
6: $\mathbf{w}_1^{k+1} = \frac{1}{T_k} \sum_{t=1}^{T_k} \mathbf{w}_t^k$
7: $\eta_{k+1} = \eta_k/2$, $T_{k+1} = 2T_k$
8: $k = k + 1$
9: **end while**

- Our Goal
  - Establish an $O(1/T)$ excess risk bound

Outline

Analysis I

If we have

$$\mathrm{E}\left[F(\mathbf{w}_1^k)\right] - F(\mathbf{w}_*) = O\left(\frac{1}{\lambda T_k}\right), \ \forall k$$

Then, the last solution $\mathbf{w}_1^{\bar{k}}$ satisfies

$$\mathrm{E}\left[F(\mathbf{w}_1^{\bar{k}})\right] - F(\mathbf{w}_*) = O\left(\frac{1}{\lambda T_{\bar{k}}}\right) = O\left(\frac{1}{\lambda T}\right)$$

Suppose

$$\mathrm{E}\left[F(\mathbf{w}_1^k)\right] - F(\mathbf{w}_*) = c\frac{G^2}{\lambda T_k}$$

From the analysis of SGD, we have

$$\mathrm{E}\left[F(\mathbf{w}_1^{k+1})\right] - F(\mathbf{w}_*) \leq \frac{1}{2\eta_k T_k}\mathrm{E}\left[\|\mathbf{w}_1^k - \mathbf{w}_*\|_2^2\right] + \frac{\eta_k}{2}G^2$$

Analysis II

Recall that

$$\|\mathbf{w}_1^k - \mathbf{w}_*\|_2^2 \leq \frac{2}{\lambda}\left(F(\mathbf{w}_1^k) - F(\mathbf{w}_*)\right)$$

Then, we have

$$\begin{aligned} \mathrm{E}\left[F(\mathbf{w}_1^{k+1})\right] - F(\mathbf{w}_*) \leq & \frac{1}{\lambda \eta_k T_k} \mathrm{E}\left[F(\mathbf{w}_1^k) - F(\mathbf{w}_*)\right] + \frac{\eta_k}{2} G^2 \\ \overset{\lambda \eta_k T_k = 4}{=} & \frac{1}{4} \mathrm{E}\left[c \frac{G^2}{\lambda T_k}\right] + \frac{4}{2\lambda T_k} G^2 \\ \overset{T_{k+1} = 2T_k}{=} & c \frac{G^2}{\lambda T_{k+1}} \left(\frac{1}{2} + \frac{4}{c}\right) \overset{c=8}{=} c \frac{G^2}{\lambda T_{k+1}} \end{aligned}$$

Outline

What do We Need?

- Key Inequality in the Expectation Bound

$$\mathrm{E}\left[F(\mathbf{w}_1^{k+1})\right] - F(\mathbf{w}_*) \leq \frac{1}{2\eta_k T_k}\mathrm{E}\left[\|\mathbf{w}_1^k - \mathbf{w}_*\|_2^2\right] + \frac{\eta_k}{2}G^2$$

What do We Need?

- Key Inequality in the Expectation Bound

$$\mathrm{E}\left[F(\mathbf{w}_1^{k+1})\right] - F(\mathbf{w}_*) \leq \frac{1}{2\eta_k T_k}\mathrm{E}\left[\|\mathbf{w}_1^k - \mathbf{w}_*\|_2^2\right] + \frac{\eta_k}{2}G^2$$

- A High-probability Inequality based on Azuma's Inequality

$$F(\mathbf{w}_1^{k+1}) - F(\mathbf{w}_*) \leq \frac{1}{2\eta_k T_k}\|\mathbf{w}_1^k - \mathbf{w}_*\|_2^2 + \frac{\eta_k}{2}G^2 + 2\sqrt{\frac{GD}{T_k}\log\frac{1}{\delta}}$$

What do We Need?

- Key Inequality in the Expectation Bound

$$\mathrm{E}\left[F(\mathbf{w}_1^{k+1})\right] - F(\mathbf{w}_*) \leq \frac{1}{2\eta_k T_k}\mathrm{E}\left[\|\mathbf{w}_1^k - \mathbf{w}_*\|_2^2\right] + \frac{\eta_k}{2}G^2$$

- A High-probability Inequality based on Azuma's Inequality

$$F(\mathbf{w}_1^{k+1}) - F(\mathbf{w}_*) \leq \frac{1}{2\eta_k T_k}\|\mathbf{w}_1^k - \mathbf{w}_*\|_2^2 + \frac{\eta_k}{2}G^2 + 2\sqrt{\frac{GD}{T_k}\log\frac{1}{\delta}}$$

- The Inequality that We Need

$$F(\mathbf{w}_1^{k+1}) - F(\mathbf{w}_*) \leq \frac{1}{2\eta_k T_k}\|\mathbf{w}_1^k - \mathbf{w}_*\|_2^2 + \frac{\eta_k}{2}G^2 + O\left(\frac{1}{\lambda T_k}\right)$$

Concentration Inequality I

Bernstein's Inequality for Martingales

Let $X_1, \ldots, X_n$ be a bounded martingale difference sequence with respect to the filtration $\mathcal{F} = (\mathcal{F}_i)_{1\leq i\leq n}$ and with $|X_i| \leq K$. Let

$$S_i = \sum_{j=1}^{i} X_j$$

be the associated martingale. Denote the sum of the conditional variances by

$$\Sigma_n^2 = \sum_{t=1}^{n} \mathrm{E}\left[X_t^2 | \mathcal{F}_{t-1}\right].$$

Then for all constants $t, \nu > 0$,

$$\Pr\left(\max_{i=1,\ldots,n} S_i > t \text{ and } \Sigma_n^2 \leq \nu\right) \leq \exp\left(-\frac{t^2}{2(\nu + Kt/3)}\right)$$

Concentration Inequality II

Corollary 2

Suppose $\Sigma_n^2 \leq \nu$*. Then, with a probability at least* $1-\delta$

$$S_n \leq \sqrt{2\nu \log \frac{1}{\delta}} + \frac{2K}{3} \log \frac{1}{\delta}$$

Note

$$\sqrt{2\nu \log \frac{1}{\delta}} \leq \frac{1}{2\alpha} \log \frac{1}{\delta} + \alpha\nu, \ \forall \alpha > 0$$

The Basic Idea

For any $\mathbf{w} \in \mathcal{W}$, we have

$$
\begin{aligned}
& F(\mathbf{w}_t) - F(\mathbf{w}) \\
\leq & \langle \nabla F(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle - \frac{\lambda}{2} \|\mathbf{w}_t - \mathbf{w}\|_2^2 \\
= & \langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{w} \rangle + \underbrace{\langle \nabla F(\mathbf{w}_t) - \mathbf{g}_t, \mathbf{w}_t - \mathbf{w} \rangle}_{\delta_t} - \frac{\lambda}{2} \|\mathbf{w}_t - \mathbf{w}\|_2^2
\end{aligned}
$$

Then, following the same analysis, we arrive at

$$
F(\bar{\mathbf{w}}_T) - F(\mathbf{w}) \leq \frac{1}{2\eta T} \|\mathbf{w}_1 - \mathbf{w}\|_2^2 + \frac{\eta}{2} G^2 + \frac{1}{T} \left(\sum_{t=1}^{T} \delta_t - \frac{\lambda}{2} \sum_{t=1}^{T} \|\mathbf{w}_t - \mathbf{w}\|_2^2 \right)
$$

We are going to prove

$$
\sum_{t=1}^{T} \delta_t = \frac{\lambda}{2} \sum_{t=1}^{T} \|\mathbf{w}_t - \mathbf{w}\|_2^2 + O\left(\frac{\log \log T}{\lambda} \right)
$$

LAMDA Learning And Mining from DatA

Analysis I

$$\delta_t = \langle \nabla F(\mathbf{w}_t) - \mathbf{g}_t, \mathbf{w}_t - \mathbf{w} \rangle$$

is a martingale difference sequence with

$$|\delta_t| \leq \|\nabla F(\mathbf{w}_t) - \mathbf{g}_t\|_2 \|\mathbf{w}_t - \mathbf{w}\|_2 \leq D(\|\nabla F(\mathbf{w}_t)\|_2 + \|\mathbf{g}_t\|_2) \leq 2GD$$

where D could be a prior parameter or $2G/\lambda$ when $\mathbf{w} = \mathbf{w}_*$

The sum of the conditional variances is upper bounded by

$$\begin{aligned}\Sigma_T^2 = & \sum_{t=1}^{T} \mathrm{E}_t \left[\delta_t^2\right] \\ \leq & \sum_{t=1}^{T} \mathrm{E}_t \left[\|\nabla F(\mathbf{w}_t) - \mathbf{g}_t\|_2^2 \|\mathbf{w}_t - \mathbf{w}\|_2^2\right] \\ \leq & 4G^2 \sum_{t=1}^{T} \|\mathbf{w}_t - \mathbf{w}\|_2^2\end{aligned}$$

Analysis II

- A straightforward way

Since

$$\Sigma_T^2 \leq 4G^2 \sum_{t=1}^{T} \|\mathbf{w}_t - \mathbf{w}\|_2^2$$

Following Bernstein's inequality for martingales, with a probability at least $1 - \delta$, we have

$$\begin{aligned} \left| \sum_{t=1}^{T} \delta_t \right| \leq & \sqrt{2 \left(4G^2 \sum_{t=1}^{T} \|\mathbf{w}_t - \mathbf{w}\|_2^2 \right) \log \frac{1}{\delta}} + \frac{2}{3} 2GD \log \frac{1}{\delta} \\ \leq & \frac{\lambda}{2} \sum_{t=1}^{T} \|\mathbf{w}_t - \mathbf{w}\|_2^2 + \frac{4G^2}{\lambda} \log \frac{1}{\delta} + \frac{4}{3} GD \log \frac{1}{\delta} \end{aligned}$$

Analysis III

- We need "Peeling"
  - Divide $\sum_{t=1}^{T}\|\mathbf{w}_t-\mathbf{w}\|_2^2 \in [0, TD^2]$ in to a sequence of intervals
  - The initial: $\sum_{t=1}^{T}\|\mathbf{w}_t-\mathbf{w}\|_2^2 \in [0, D^2/T]$
    Since
    $$|\delta_t| \leq \|\nabla F(\mathbf{w}_t)-\mathbf{g}_t\|_2\|\mathbf{w}_t-\mathbf{w}\|_2 \leq 2G\|\mathbf{w}_t-\mathbf{w}\|_2$$
    we have
    $$\left|\sum_{t=1}^{T}\delta_t\right| \leq 2G\sum_{t=1}^{T}\|\mathbf{w}_t-\mathbf{w}\|_2 \leq 2G\sqrt{T}\sqrt{\sum_{t=1}^{T}\|\mathbf{w}_t-\mathbf{w}\|_2^2} \leq 2GD$$
  - The rest: $\sum_{t=1}^{T}\|\mathbf{w}_t-\mathbf{w}\|_2^2 \in (2^{i-1}D^2/T, 2^{i}D^2/T]$ where $i=1,\ldots,\lceil 2\log_2 T\rceil$

Analysis IV

Define

$$A_T = \sum_{t=1}^{T} \|\mathbf{w}_t - \mathbf{w}\|_2^2 \leq TD^2$$

We are going to bound

$$\Pr\left[\sum_{t=1}^{T} \delta_t \geq 2\sqrt{4G^2 A_T \tau} + \frac{2}{3} 2GD\tau + 2GD\right]$$

Analysis V

$$\Pr\left[\sum_{t=1}^{T}\delta_t \geq 2\sqrt{4G^2A_T\tau}+\frac{2}{3}2GD\tau+2GD\right]$$

$$=\Pr\left[\sum_{t=1}^{T}\delta_t \geq 2\sqrt{4G^2A_T\tau}+\frac{4}{3}GD\tau+2GD, A_T\leq D^2/T\right]$$

$$+\Pr\left[\sum_{t=1}^{T}\delta_t \geq 2\sqrt{4G^2A_T\tau}+\frac{4}{3}GD\tau+2GD, D^2/T< A_T\leq TD^2\right]$$

$$=\Pr\left[\sum_{t=1}^{T}\delta_t \geq 2\sqrt{4G^2A_T\tau}+\frac{4}{3}GD\tau+2GD, \Sigma_T^2\leq 4G^2A_T, D^2/T< A_T\leq TD^2\right]$$

$$\leq\sum_{i=1}^{m}\Pr\left[\sum_{t=1}^{T}\delta_t \geq 2\sqrt{4G^2A_T\tau}+\frac{4}{3}GD\tau+2GD, \Sigma_T^2\leq 4G^2A_T, \frac{D^2}{T}2^{i-1}< A_T\leq \frac{D^2}{T}2^{i}\right]$$

$$\leq\sum_{i=1}^{m}\Pr\left[\sum_{t=1}^{T}\delta_t \geq \sqrt{2\frac{4G^2D^22^i}{T}\tau}+\frac{4}{3}GD\tau, \Sigma_T^2\leq \frac{4G^2D^22^i}{T}\right]\leq me^{-\tau}$$

where $m=\lceil 2\log_2 T\rceil$

Analysis VI

By setting $\tau = \log \frac{m}{\delta}$, with a probability at least $1-\delta$,

$$
\begin{aligned}
\sum_{t=1}^{T} \delta_t \leq & 2\sqrt{4G^2 A_T \tau} + \frac{4}{3} GD\tau + 2GD \\
= & 2\sqrt{4G^2 \left(\sum_{t=1}^{T} \|\mathbf{w}_t - \mathbf{w}\|_2^2 \right) \tau} + \frac{4}{3} GD\tau + 2GD \\
\leq & \frac{\lambda}{2} \sum_{t=1}^{T} \|\mathbf{w}_t - \mathbf{w}\|_2^2 + \frac{8G^2}{\lambda}\tau + \frac{4}{3} GD\tau + 2GD \\
= & \frac{\lambda}{2} \sum_{t=1}^{T} \|\mathbf{w}_t - \mathbf{w}\|_2^2 + O\left(\frac{\log \log T}{\lambda} \right)
\end{aligned}
$$

where

$$
\tau = \log \frac{m}{\delta} = \log \frac{\lceil 2 \log_2 T \rceil}{\delta} = O(\log \log T)
$$

Reference I

Agarwal, A., Bartlett, P. L., Ravikumar, P., and Wainwright, M. J. (2012).
Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization.
IEEE Transactions on Information Theory, 58(5):3235–3249.

Bach, F. and Moulines, E. (2013).
Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$.
In *Advances in Neural Information Processing Systems 26*, pages 773–781.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005).
Local rademacher complexities.
The Annals of Statistics, 33(4):1497–1537.

Bartlett, P. L. and Mendelson, S. (2002).
Rademacher and gaussian complexities: risk bounds and structural results.
Journal of Machine Learning Research, 3:463–482.

Feldman, V. (2016).
Generalization of erm in stochastic convex optimization: The dimension strikes back.
ArXiv e-prints, arXiv:1608.04414.

Reference II

Hazan, E., Agarwal, A., and Kale, S. (2007).
Logarithmic regret algorithms for online convex optimization.
Machine Learning, 69(2-3):169–192.

Hazan, E. and Kale, S. (2011).
Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization.
In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 421–436.

Koren, T. and Levy, K. (2015).
Fast rates for exp-concave empirical risk minimization.
In *Advances in Neural Information Processing Systems 28*, pages 1477–1485.

Kushner, H. J. and Yin, G. G. (2003).
Stochastic Approximation and Recursive Algorithms and Applications.
Springer, second edition.

Lan, G. (2012).
An optimal method for stochastic composite optimization.
Mathematical Programming, 133:365–397.

Reference III

Lee, W. S., Bartlett, P. L., and Williamson, R. C. (1996).
The importance of convexity in learning with squared loss.
In *Proceedings of the 9th Annual Conference on Computational Learning Theory*, pages 140–146.

Mahdavi, M., Zhang, L., and Jin, R. (2015).
Lower and upper bounds on the generalization of stochastic exponentially concave optimization.
In *Proceedings of the 28th Conference on Learning Theory*.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009).
Robust stochastic approximation approach to stochastic programming.
SIAM Journal on Optimization, 19(4):1574–1609.

Nesterov, Y. (2011).
Random gradient-free minimization of convex functions.
Core discussion papers.

Panchenko, D. (2002).
Some extensions of an inequality of vapnik and chervonenkis.
Electronic Communications in Probability, 7:55–65.

Reference IV

- Rakhlin, A., Shamir, O., and Sridharan, K. (2012).
  Making gradient descent optimal for strongly convex stochastic optimization.
  In *Proceedings of the 29th International Conference on Machine Learning*, pages 449–456.
- Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2009).
  Stochastic convex optimization.
  In *Proceedings of the 22nd Annual Conference on Learning Theory*.
- Shamir, O. and Zhang, T. (2013).
  Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes.
  In *Proceedings of the 30th International Conference on Machine Learning*, pages 71–79.
- Srebro, N., Sridharan, K., and Tewari, A. (2010).
  Optimistic rates for learning with a smooth loss.
  ArXiv e-prints, arXiv:1009.3896.
- Sridharan, K., Shalev-shwartz, S., and Srebro, N. (2009).
  Fast rates for regularized objectives.
  In *Advances in Neural Information Processing Systems 21*, pages 1545–1552.

Reference V

Tsybakov, A. B. (2004).
Optimal aggregation of classifiers in statistical learning.
The Annals of Statistics, 32:135–166.

Vapnik, V. N. (1998).
Statistical Learning Theory.
Wiley-Interscience.

Zhang, L., Yang, T., and Jin, R. (2017).
Empirical risk minimization for stochastic convex optimization: $O(1/n)$- and $O(1/n^2)$-type of risk bounds.
In *Proceedings of the 30th Annual Conference on Learning Theory*.

Zinkevich, M. (2003).
Online convex programming and generalized infinitesimal gradient ascent.
In *Proceedings of the 20th International Conference on Machine Learning*, pages 928–936.

LAMDA Learning And Mining from DatA