Stochasitic Gradient Descent

Peng Lingwei

August 1, 2019

Contents

14 Stochastic Gradient Descent **1**
14.1 GRADIENT DESCENT 2
14.1.1 Analysis of GD for Convex-Lipschitz Functions 2
14.2 SUBGRADIENTS 3
14.2.1 Calculating Subgradients 3
14.2.2 Subgradients of Lipschitz Functions 3
14.2.3 Subgradient Descent 3
14.3 STOCHASTIC GRADIENT DESCENT (SGD) 3
14.3.1 Anslysis of SGD for Convex-Lipschitz-Bounded Functions 4
14.4 VARIANTS 4
14.4.1 Adding a Projection Step 4
14.4.2 Variable Step Size 4
14.4.3 Other Averaging Techniques 5
14.4.4 Strongly Convex Function 5
14.5 LEARNING WITH SGD 6
14.5.1 SGD for Risk Minimization 6
14.5.2 Analyzing SGD for Convex-Smooth Learning Problems . 6
14.5.3 SGD for Regularized Loss Minimization 7

14 Stochastic Gradient Descent

The simplicity of SGD also allows us to use it in situations when it is not possible to apply method that based on the empirical risk.

14.1 GRADIENT DESCENT

14.1.1 Analysis of GD for Convex-Lipschitz Functions

We are interested in: $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}^{(t)}$. By using Jensen's inequality,

$$\begin{aligned} f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) =& f\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{w}^{(t)}\right) - f(\mathbf{w}^*) \\ \leq& \frac{1}{T}\sum_{t=1}^{T} f\left(\mathbf{w}^{(t)}\right) - f(\mathbf{w}^*) \\ =& \frac{1}{T}\sum_{t=1}^{T} \left(f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*)\right) \\ \leq& \frac{1}{T}\sum_{t=1}^{T} \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla f(\mathbf{w}^{(t)})\rangle \end{aligned}$$

Lemma 14.1. *$\forall \mathbf{v}_1, \ldots, \mathbf{v}_T$, s.t. $\mathbf{w}^{(1)} = 0$, $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t$, we have:*

$$\sum_{t=1}^{T} \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t\rangle \leq \frac{\|\mathbf{w}^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \|\mathbf{v}_t\|^2 \tag{14.1}$$

Proof. First, key step:

$$\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t\rangle = \frac{1}{2\eta}(-\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2 + \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2) + \frac{\eta}{2}\|\mathbf{v}_t\|^2$$

Second, key step:

$$\begin{aligned} \sum_{t=1}^{T} \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t\rangle =& \frac{1}{2\eta}(-\|\mathbf{w}^{(T+1)} - \mathbf{w}^*\|^2 + \|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2) + \frac{\eta}{2}\sum_{t=1}^{T} \|\mathbf{v}_t\|^2 \\ \leq& \frac{1}{2\eta}\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2 + \frac{\eta}{2}\sum_{t=1}^{T} \|\mathbf{v}_t\|^2 \\ =& \frac{1}{2\eta}\|\mathbf{w}^*\|^2 + \frac{\eta}{2}\sum_{t=1}^{T} \|\mathbf{v}_t\|^2 \end{aligned}$$

□

$$
\begin{aligned}
f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \leq & \frac{1}{T}\sum_{t=1}^{T}\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla f(\mathbf{w}^{(t)})\rangle \\
= & \frac{1}{T}\left(\frac{1}{2\eta}\|\mathbf{w}^*\|^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{v}_t\|^2\right) \\
\leq & \frac{1}{2T}\sqrt{\|\mathbf{w}^*\|^2 \cdot \sum_{t=1}^{T}\|\mathbf{v}_t\|^2} \\
\leq & \frac{B\rho}{\sqrt{T}}
\end{aligned}
$$

(f is a convex, $\rho - Lipschitz$ function. $\|\vec{v}_t\| = \|\nabla f(\mathbf{w}_t)\| \leq \rho$)

14.2 SUBGRADIENTS

Definition 14.1. *(Subgradient) ∂f:*

$$
\forall \mathbf{u}, \quad f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \partial f(\mathbf{w})\rangle \tag{14.2}
$$

14.2.1 Calculating Subgradients

14.2.2 Subgradients of Lipschitz Functions

Lemma 14.2. *Let A be a convex open set and let $f : A \to \mathbb{R}$ be a convex function. Then, f is $\rho - Lipschitz$ over A iff $\forall \mathbf{w} \in A$ and $\mathbf{v} \in \partial f(\mathbf{w})$ we have that $\|\mathbf{v}\| \leq \rho$*

Sufficiency: $f(\mathbf{w}) - f(\mathbf{u}) \leq \langle \mathbf{v}, \mathbf{w} - \mathbf{u}\rangle \leq \|\mathbf{v}\|\|\mathbf{w} - \mathbf{u}\| \leq \rho\|\mathbf{w} - \mathbf{u}\|$

Necessity:Let $\mathbf{w} \in A, \mathbf{v} \in \partial f(\mathbf{w}), \mathbf{u} = \mathbf{w} + \epsilon\mathbf{v}/\|\mathbf{v}\|$*, Then, we have:*

$$
\rho\epsilon = \rho\|\mathbf{u} - \mathbf{w}\| \geq f(\mathbf{u}) - f(\mathbf{w}) \geq \langle \mathbf{v}, \mathbf{u} - \mathbf{w}\rangle = \epsilon\|\mathbf{v}\|
$$

14.2.3 Subgradient Descent

The analysis of the convergence rate remains unchanged.

14.3 STOCHASTIC GRADIENT DESCENT (SGD)

Algorithm 1 Stochastic Gradient Descent (SGD) for minimizing $f(\mathbf{w})$.

Require: Scalar $\eta > 0$, integer $T > 0$
Ensure: $\mathbf{w}^{(1)} = \mathbf{0}$
  for $t = 1, 2, \ldots, T$ **do**
    random choose $\mathbf{v}_t$ (make sure that $\mathbb{E}_{\mathcal{D}}[\mathbf{v}_t|\mathbf{w}^{(t)}] \in \partial f(\mathbf{w}^{(t)})$)
    $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta\mathbf{v}_t$
  end for.
  return $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}$

14.3.1 Anslysis of SGD for Convex-Lipschitz-Bounded Functions

Theorem 14.1. *Let B,$\rho > 0$, and $\mathbb{P}\{\|\mathbf{v}_t\| \leq \rho\} = 1$. Then,*

$$\mathbf{E}_{\mathcal{D}}[f(\bar{\mathbf{w}})] - f(\mathbf{w}^*) \leq \frac{B\rho}{\sqrt{T}} \tag{14.3}$$

Proof. Key step: proof

$$\mathop{\mathbb{E}}_{\mathbf{v}_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}(f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*))\right] \leq \mathop{\mathbb{E}}_{\mathbf{v}_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t\rangle\right] \tag{14.4}$$

Subproof:

$$\mathop{\mathbb{E}}_{\mathbf{v}_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t\rangle\right] = \frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\mathbf{v}_{1:T}}\left[\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t\rangle\right] = \frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\mathbf{v}_{1:t}}\left[\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t\rangle\right]$$

$$\mathop{\mathbb{E}}_{\mathbf{v}_{1:t}}\left[\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t\rangle\right] = \mathop{\mathbb{E}}_{\mathbf{v}_{1:t-1}}\mathop{\mathbb{E}}_{\mathbf{v}_t}\left[\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t\rangle | \mathbf{v}_{1:t-1}\right] = \mathop{\mathbb{E}}_{\mathbf{v}_{1:t-1}}\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathop{\mathbb{E}}_{\mathbf{v}_t}[\mathbf{v}_t|\mathbf{v}_{1:t-1}]\rangle$$

We have $\mathop{\mathbb{E}}_{\mathbf{v}_t}[\mathbf{v}_t|\mathbf{w}^{(t)}] \in \partial f(\mathbf{w}^{(t)})$, which equals to $\mathop{\mathbb{E}}_{\mathbf{v}_t}[\mathbf{v}_t|\mathbf{v}_{1:t-1}] \in \partial f(\mathbf{w}^{(t)})$, so

$$\mathop{\mathbb{E}}_{\mathbf{v}_{1:T}}\left[\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t\rangle\right] \geq \mathop{\mathbb{E}}_{\mathbf{v}_{1:t-1}}\left[f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*)\right] = \mathop{\mathbb{E}}_{\mathbf{v}_{1:T}}\left[f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*)\right]$$

□

14.4 VARIANTS

14.4.1 Adding a Projection Step

In previous analyses of the GD and SGD algorithms, we require $\|\mathbf{w}^*\| \leq B$, but there is no guarantee that $\bar{\mathbf{w}}$ satisfies it. So here comes the projection step.

Definition 14.2. *(Porjection step).*

1. $\mathbf{w}^{(t+\frac{1}{2})} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t$
2. $\mathbf{w}^{(t+1)} = \arg\min_{\mathbf{w}\in\mathcal{H}} \|\mathbf{w} - \mathbf{w}^{(t+\frac{1}{2})}\|$

Lemma 14.3. *(Projection Lemma)*

$$\mathbf{v} = \mathop{\arg\min}_{\mathbf{x}\in\mathcal{H}}\|\mathbf{x} - \mathbf{w}\|^2 \Rightarrow \|\mathbf{w} - \mathbf{u}\|^2 - \|\mathbf{v} - \mathbf{u}\|^2 \geq 0$$

So we have:

$$\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \leq \|\mathbf{w}^{(t+\frac{1}{2})} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2$$

14.4.2 Variable Step Size

We can set $\eta_t = \frac{B}{\rho\sqrt{t}}$ and achieve a similar bound.

14.4.3 Other Averaging Techniques

- $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}$
- $\bar{\mathbf{w}} = \mathbf{w}^{(t)}$, for some random $t \in [t]$
- $\bar{\mathbf{w}} = \frac{1}{\alpha T}\sum_{t=T-\alpha T}^{T}\mathbf{w}^{(t)}$ for $\alpha \in (0,1)$

14.4.4 Strongly Convex Function

Algorithm 2 SGD for minimizing a $\lambda -$ *strongly* convex function

---

Ensure: $\mathbf{w}^{(1)} = \mathbf{0}$
  for $t = 1, \ldots, T$ **do**
    Choose a random vector $\mathbf{v}_t$ (s.t. $\mathbb{E}[\mathbf{v}_t|\mathbf{w}^{(t)}] \in \partial f(\mathbf{w}^{(t)})$)
    $\eta_t = 1/(\lambda t)$
    $\mathbf{w}^{(t+\frac{1}{2})} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t$
    $\mathbf{w}^{(t+1)} = \arg\min_{\mathbf{w}\in\mathcal{H}} \|\mathbf{w} - \mathbf{w}^{(t+\frac{1}{2})}\|^2$
  end for.
  return $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}$

Theorem 14.2. *Assume that f is $\lambda -$ strongly convex and that $\mathbb{E}[\|\mathbf{v}_t\|^2] \leq \rho^2$. Let $\mathbf{w}^* \in \arg\min_{\mathbf{w}\in\mathcal{H}} f(\mathbf{w})$ be an optimal solution. Then,*

$$\mathbb{E}[f(\bar{\mathbf{w}})] - f(\mathbf{w}^*) \leq \frac{\rho^2}{2\lambda T}(1 + \log(T)) \tag{14.5}$$

Proof. We already have:

$$\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla^{(t)} \rangle \leq \frac{\mathbb{E}[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2]}{2\eta_t} + \frac{\eta_t}{2}\rho^2 \tag{14.6}$$

So:

$$\begin{aligned}
&\sum_{t=1}^{T}(\mathbb{E}[f(\mathbf{w}^{(t)})] - f(\mathbf{w}^*)) \\
\leq &\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2}{2\eta_t} - \frac{\lambda}{2}\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2\right)\right] + \frac{\rho^2}{2}\sum_{t=1}^{T}\eta_t
\end{aligned}$$

When we use the definition $\eta_t = 1/(\lambda t)$, then we can telescope the right side:

$$\sum_{t=1}^{T}(\mathbb{E}[f(\mathbf{w}^{(t)})] - f(\mathbf{w}^*)) \leq -\frac{\lambda T}{2}\|\mathbf{w}^{(T+1)} - \mathbf{w}^*\|^2 + \frac{\rho^2}{2\lambda}\sum_{t=1}^{T}\frac{1}{t} \leq \frac{\rho^2}{2\lambda}(1 + \ln(T)).$$

(Because $\int_1^T 1/x dx > \sum_{t=2}^{T} 1/t$) □

14.5 LEARNING WITH SGD

14.5.1 SGD for Risk Minimization

SGD allows us to take a different approach and minimize $L_{\mathcal{D}}(\mathbf{w})$ directly.

Definition 14.3. *(Risk function)* $L_{\mathcal{D}}(\mathbf{w}) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}}[l(\mathbf{w}, z)]$.

We set

$$\mathbf{v}_t = \nabla l(\mathbf{w}_{(t)}, z), \quad where\ z \sim \mathcal{D}.$$

Then,

$$\mathbb{E}_{\mathbf{v}_t}[\mathbf{v}_t|\mathbf{w}^{(t)}] = \mathop{\mathbb{E}}_{z\sim\mathcal{D}}[\nabla l(\mathbf{w}^{(t)}, z)] = \nabla \mathop{\mathbb{E}}_{z\sim\mathcal{D}}[l(\mathbf{w}^{(t)}, z)] = \nabla L_{\mathcal{D}}(\mathbf{w}^{(t)})$$

Algorithm 3 Stochastic Gradient (SGD) for minimizing $L_{\mathcal{D}}(\mathbf{w})$

Ensure: $\mathbf{w}^{(1)} = \mathbf{0}$
  for $t = 1, 2, \ldots, T$ **do**
    sample $z \sim \mathcal{D}$
    pick $\mathbf{v}_t \in \partial l(\mathbf{w}^{(t)}, z)$
    $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta\mathbf{v}_t$
  end for.
  return $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}$

We can get $\mathbb{E}[L_{\mathcal{D}}(\bar{\mathbf{w}})] \leq \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}(\mathbf{w}) + \frac{B\rho}{\sqrt{T}}$ on a convex-Lipschitz-bounded learning problem.When setting $T \geq \frac{B^2\rho^2}{\epsilon^2}$, we can get the accuracy to ϵ.

14.5.2 Analyzing SGD for Convex-Smooth Learning Problems

Theorem 14.3. *Assume that for all z, the loss function $l(\cdot, z)$ is convex, $\beta-$ smooth, and nonnegative. Then, if we run the SGD algorithm for minimzing $L_{\mathcal{D}}(\mathbf{w})$ we have that:*

$$\forall \mathbf{w}^*, \quad \mathbb{E}[L_{\mathcal{D}}(\bar{\mathbf{w}})] \leq \frac{1}{1-\eta\beta}\left(L_{\mathcal{D}}(\mathbf{w}^*) + \frac{\|\mathbf{w}^*\|^2}{2\eta T}\right) \tag{14.7}$$

Proof. Let $z_1, \ldots, z_T$ be the random samples of SGD algorithm, and $f_t(\cdot) = l(\cdot, z_t)$. So

$$\sum_{t=1}^{T}(f_t(\mathbf{w}^{(t)}) - f_t(\mathbf{w}^*)) \leq \sum_{t=1}^{T}\langle\mathbf{v}_t, \mathbf{w}^{(t)} - \mathbf{w}^*\rangle \leq \frac{\|\mathbf{w}^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{v}_t\|^2 \leq \frac{\|\mathbf{w}^*\|^2}{2\eta} + \eta\beta\sum_{t=1}^{T}f_t(\mathbf{w}^{(t)})$$

The last inequation comes from self-bounded property.

$$\frac{1}{T}\sum_{t=1}^{T}f_t(\mathbf{w}^{(t)}) \leq \frac{1}{1-\eta\beta}\left(\frac{1}{T}\sum_{t=1}^{T}f_t(\mathbf{w}^*) + \frac{\|\mathbf{w}^*\|^2}{2\eta T}\right)$$

Remaining steps are taking expectation of both side and using Jensen's inequation. □

$$
\begin{aligned}
\mathbb{E}L_{\mathcal{D}}(\bar{\mathbf{w}}) \leq & L_{\mathcal{D}}(\mathbf{w}^*) + \frac{\eta\beta}{1-\eta\beta} L_{\mathcal{D}}(\mathbf{w}^*) + \frac{\|\mathbf{w}^*\|^2}{(1-\eta\beta)2\eta T} \\
\leq & L_{\mathcal{D}}(\mathbf{w}^*) + \frac{\eta\beta}{1-\eta\beta} + \frac{\|\mathbf{w}^*\|^2}{(1-\eta\beta)2\eta T}, \quad (L_{\mathcal{D}} \leq 1)
\end{aligned}
$$

It's easy to construct η, β, T letting $\mathbb{E}L_{\mathcal{D}}(\bar{\mathbf{w}}) \leq \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}(\mathbf{w}) + \epsilon$.

14.5.3 SGD for Regularized Loss Minimization

$$
\min_{\mathbf{w}} \left(\frac{\lambda}{2}\|\mathbf{w}\|^2 + L_S(\mathbf{w}) \right)
$$

Regularization function $f(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + L_S(\mathbf{w})$ is $\lambda-$*strongly* convex function.

Lemma 14.4. *Random choose $z_t \sim \mathcal{D}$, and pick $\mathbf{v}_t \in \partial l(\mathbf{w}^{(t)}, z)$, then $\mathbb{E}[\lambda \mathbf{w}^{(t)} + \mathbf{v}_t] \in \partial f(\mathbf{w}^{(t)}, z)$*

In this task, we set $\eta = \frac{1}{\lambda t}$, then:

$$
t\mathbf{w}^{(t+1)} = t\mathbf{w}^{(t)} - \frac{1}{\lambda}(\lambda\mathbf{w}^{(t)} + \mathbf{v}_t) = (t-1)\mathbf{w}^{(t)} - \frac{\mathbf{v}_t}{\lambda} = -\frac{\sum_{i=1}^{t} \mathbf{v}_i}{\lambda}
$$

If the loss function is $\rho-$*Lipschitz*,then we have $\|\lambda\mathbf{w}^{(t)}\| \leq \rho$, which also means $\|\lambda\mathbf{w}^{(t)} + \mathbf{v}_t\| \leq 2\rho$. Theorem 14.2 tells us that:

$$
\mathbb{E}[f(\bar{\mathbf{w}})] - f(\mathbf{w}^*) \leq \frac{2\rho^2}{\lambda T}(1 + \log(T)) \tag{14.8}
$$