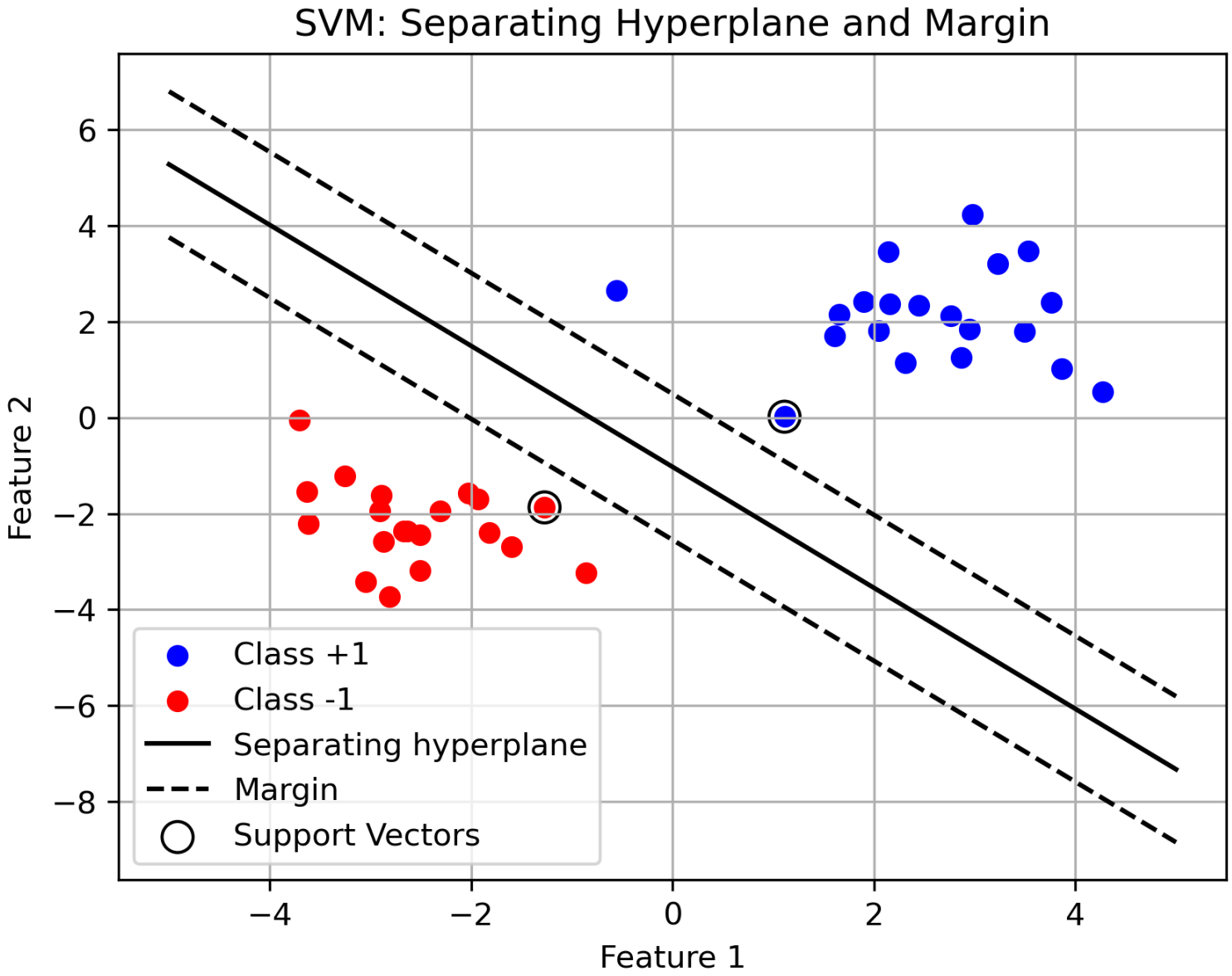
Stochastic Gradient Methods

DSAI5104 (2025-26)

Support Vector Machine (SVM)



Part I: Perceptron Machine

- Supervised learning problem: Suppose we have data $\{(\mathbf{x}_i, y_i)\}$, $i \in [n]$ with $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \{+1, -1\}$ (sample data).
- Linear separator: We would like to find a linear function $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$ such that the positive labeled data stays on one side of the hyperplane:

$$\mathcal{H} := \{\mathbf{x} \in \mathbb{R}^p \mid f(\mathbf{x}) = 0\}$$

  while the negative labeled data stay the other side.
- Separation by unit margin:

$$\begin{aligned} \langle \mathbf{w}, \mathbf{x} \rangle &\geq 1 \qquad \text{for } y_i = +1 \\ \langle \mathbf{w}, \mathbf{x} \rangle &\leq -1 \qquad \text{for } y_i = -1. \end{aligned}$$

  Equivalently,

$$y_i \langle \mathbf{w}, \mathbf{x} \rangle \geq 1, \quad i \in [n].$$

Margin Mistake

- We say a data point $(\mathbf{x}_i, y_i)$ makes margin mistake with respect to the hyperplane $\mathcal{H} := \{\mathbf{x} \mid \langle \mathbf{w}, \mathbf{x} \rangle = 0\}$ if

$$y_i \langle \mathbf{w}, \mathbf{x}_i \rangle < 1$$

- When data $(\mathbf{x}_i, y_i)$ makes a margin mistake, either it is correctly classified

$$y_i = \text{sgn}\Big(\langle \mathbf{w}, \mathbf{x}_i \rangle\Big),$$

or it is wrongly classified. But in both cases, the data is close to the hyperplane in the sense $y_i \langle \mathbf{w}, \mathbf{x}_i \rangle < 1$.

Perceptron machine: How it works

The purpose is to find the "best" $\mathbf{w}$ hoping to satisfy $y_i\langle \mathbf{w},\ \mathbf{x}_i\rangle \geq 1$ for all data.

Start with $\mathbf{w}^0 = 0$, randomly pick data $(\mathbf{x}_k, y_k)$, do

$$\mathbf{w}^{k+1} = \begin{cases} \mathbf{w}^k & \text{if } y_k\langle \mathbf{w}^k,\ \mathbf{x}_k\rangle \geq 1 \\ \mathbf{w}^k + y_k\mathbf{x}_k & \text{otherwise} \quad \text{(correction step)} \end{cases}$$

The intuition is that when the current separating plane works well for data $\mathbf{x}_k$, we keep the plane unchanged; otherwise we update it so that the condition is likely to be satisfied at $\mathbf{x}_k$. This cane be seen below

$$y_k\langle \mathbf{w}^{k+1},\ \mathbf{x}_k\rangle = y_k\langle \mathbf{w}^k,\ \mathbf{x}_k\rangle + \|\mathbf{x}_k\|^2 > y_k\langle \mathbf{w}^k,\ \mathbf{x}_k\rangle.$$

Perceptron machine: Theoretical guarantee

Assumption A (Boundedness Assumption): There exists $D > 0$ such that

$$\|\mathbf{x}_i\| \leq D, \quad i = 1, \ldots, .$$

Assumption B (Separability Assumption): There exist a weight vector $\mathbf{w}_0$ and $\rho_0 > 0$ such that

$$\min_{(\mathbf{x}_i, y_i)} \frac{y_i \langle \mathbf{w}_0, \mathbf{x}_i \rangle}{\|\mathbf{w}_0\|} \geq \rho_0.$$

Theorem (Novikoff)

Suppose Assumptions A and B hold. The Perceptron machine makes at most

$$M = \left\lfloor \frac{2 + D^2}{\rho_0^2} \right\rfloor.$$

corrections.

Proof

Let the weight vectors after the first N corrections by

$$\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(N)}.$$

We assume at the Nth correction happens at data $(\mathbf{x}_t.y_t)$. We must have

$$y_t \langle \mathbf{w}^{(N-1)},\ \mathbf{x}_t \rangle < 1 \quad \text{and} \quad \mathbf{w}^{(N)} = \mathbf{w}^{(N-1)} + y_t \mathbf{x}_t.$$

Therefore, we have an upper bound for $\|\mathbf{w}^{(N)}\|$:

$$\begin{aligned} \|\mathbf{w}^{(N)}\|^2 &= \|\mathbf{w}^{(N-1)}\|^2 + 2y_t \langle \mathbf{w}^{(N-1)},\ \mathbf{x}_t \rangle + \|\mathbf{x}_t\|^2 \\ &\leq \|\mathbf{w}^{(N-1)}\|^2 + 2 + D \\ &\vdots \\ &\leq N(2 + D) \end{aligned}$$

using the fact that the initial weight vector is $\mathbf{w}^0 = 0$.

On the other hand,

$$\begin{aligned}\frac{\langle \mathbf{w}^{(N)},\ \mathbf{w}_0\rangle}{\|\mathbf{w}_0\|} &= \frac{\langle \mathbf{w}^{(N-1)},\ \mathbf{w}_0\rangle}{\|\mathbf{w}_0\|} + \frac{y_t\langle \mathbf{x}^t,\ \mathbf{w}_0\rangle}{\|\mathbf{w}_0\|} \\ &\geq \frac{\langle \mathbf{w}^{(N-1)},\ \mathbf{w}_0\rangle}{\|\mathbf{w}_0\|} + \rho_0 \\ &\vdots \\ &\geq N\rho_0.\end{aligned}$$

By Cauchy-Schwartz inequality, we have a lower bound:

$$\|\mathbf{w}^{(N)}\| \geq N\rho_0.$$

Combining the upper and lower bounds, we see N cannot be bigger than

$$\frac{2+D}{\rho_0^2}.$$

Link to Optimization

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \underbrace{\max\{1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle, 0\}}_{=f_i(\mathbf{w})\ \text{(hinge loss)}} + \underbrace{\frac{\lambda}{2}\|\mathbf{w}\|^2}_{\text{regularization}} \tag{1}$$

- randomly pick $i \in [n]$, calculates the partial gradient at $\mathbf{w}^k$:

$$\nabla f_i(\mathbf{w}^k) + \lambda \mathbf{w}^k = \begin{cases} -y_i \mathbf{x}_i + \lambda \mathbf{w}^k & \text{if } y_i\langle \mathbf{w}, \mathbf{x}_i\rangle < 1 \\ \lambda \mathbf{w}^k & \text{otherwise.} \end{cases}$$

- Perform the gradient decent step:

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \alpha\Big(\nabla f_i(\mathbf{w}^k) + \lambda \mathbf{w}^k\Big) = (1 - \alpha\lambda)\mathbf{w}^k - \alpha \nabla f_i(\mathbf{w}^k),$$

  where $\alpha > 0$ is steplength or learning rate.

- When $\lambda = 0$, $\alpha = 1$, we recover the Perceptron algorithm.

Perceptron machine can be put in a general framework that is known as SGD (Stochastic Gradient Descent) methods. Consider

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; f(\mathbf{x}).$$

- At the kth iterate $\mathbf{x}^k$, we compute stochastic gradient $\mathbf{g}_k$:

$$\mathbf{g}_k \approx \nabla f(\mathbf{x}^k).$$

- Perform the stochastic gradient step:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k \mathbf{g}_k,$$

  where $\alpha_k > 0$ is the steplength.

Under what conditions, the framework would work?

Example: Computing a mean

Suppose we have m real numbers y_i, $i = 1, \ldots, m$. Its mean $\bar{x} = (y_1 + \cdots + y_m)/m$. It is the solution of the following quadratic problem:

$$\min_x \ f(x) = \frac{1}{2} \sum_{i=1}^{m} \underbrace{\Big(x - y_i\Big)^2}_{f_i(x)}.$$

We have

$$\nabla f_i = x - y_i.$$

Suppose we start from $x^{(1)}$ and $\alpha_k = \frac{1}{k}$. We draw the data from 1 to m .

Example is from the book by M. Hardt and B. Recht, Patterns, Predictions, and Actions – A Story about Machine Learning, 2023.

We then have

$$x^{(2)} = x^{(1)} - \alpha_1 \nabla f_1(x^{(1)}) = x^{(1)} - (x^{(1)} - y_1) = y_1$$
$$x^{(3)} = x^{(2)} - \alpha_2 \nabla f_1(x^{(2)}) = x^{(2)} - \frac{1}{2}(x^{(2)} - y_2) = \frac{1}{2}(y_1 + y_2)$$
$$x^{(4)} = x^{(3)} - \alpha_3 \nabla f_1(x^{(3)}) = x^{(3)} - \frac{1}{3}(x^{(3)} - y_3) = \frac{1}{3}(y_1 + y_2 + y_3)$$

Generalizing the calculation, we have by induction that

$$x^{(k+1)} = \left(\frac{k-1}{k}\right) x^{(k)} + \frac{1}{k} y_k = \frac{1}{k} \sum_{i=1}^{k} y_i.$$

Key messages:

- SGD works!
- $\alpha_k \to 0$, but $\sum \alpha_k \to +\infty$ as $k \to +\infty$.

How Good is the Sample Mean

Let us apply the above computation to the problem:

$$f(x) = \frac{1}{2}E[(x - Y)^2]$$

where Y is a random variable with $E[Y] = \mu$ and $\text{Var}[Y] = \sigma^2$.
We have

$$f(x) = \frac{1}{2}x^2 - \mu x + \frac{1}{2}E[Y^2].$$

The optimal solution satisfies

$$f'(x) = 0 \quad \Longrightarrow \quad x_* = \mu,$$

The optimal functional value is given by:

$$f_* = f(x_*) = -\frac{1}{2}\mu^2 + \frac{1}{2}E[Y^2] = \frac{1}{2}\sigma^2.$$

We sample i.i.d. Y_k and update x by

$$x^{(k)} = \frac{1}{k}(Y_1 + Y_2 + \cdots + Y_k).$$

We calculate

$$
\begin{aligned}
f(x^{(k)}) &= \frac{1}{2}E[(x^{(k)} - Y)^2] \\
&= \frac{1}{2}\left\{\frac{1}{k}E[Y_i^2] + \frac{k(k-1)}{k^2}\Big(E[Y]\Big)^2 - 2\Big(E[Y]\Big)^2 + E[Y^2]\right\} \\
&= \frac{k+1}{2k}\Big(E[Y^2] - (E[Y])^2\Big) \\
&= \frac{k+1}{2k}\sigma^2 \\
&= \frac{1}{2k}\sigma^2 + \frac{1}{2}\sigma^2.
\end{aligned}
$$

Therefore, after k steps, the optimality gap is

$$f(x^{(k)}) - f_* = \frac{1}{2k}\sigma^2.$$

SGD: General Framework

- Problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; f(\mathbf{x}), \quad f \text{ is differential and convex.}$$

- stochastic gradient oracle: We have a random function $\mathbf{g}(\mathbf{x}, \boldsymbol{\xi}) : \mathbb{R}^n \times \mathbb{R}^m \mapsto \mathbb{R}$ such that

$$E_\xi[\mathbf{g}(\mathbf{x}, \boldsymbol{\xi})] = \nabla f(\mathbf{x}), \qquad \forall\, \mathbf{x} \in \mathbb{R}^n$$

  where ξ is a random variable following certain distribution.

Example: In Perceptron machine, the random variable $\boldsymbol{\xi}$ takes the value of the randomly selected data $\mathbf{x}_i$, the gradient is taken to be

$$\mathbf{g}(\mathbf{w}, \mathbf{x}_i) = \text{sgn}\Big(1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle\Big) \times (-y_i\mathbf{x}_i).$$

Note that $\mathbf{x}_i$ here denotes data point, while $\mathbf{x}$ is is also used as the variable in our optimization problem. In Perceptron machine, $\mathbf{w}$ is $\mathbf{x}$ in the optimization problem. Do not get confused.

At iterate $\mathbf{x}^k$, we randomly draw $\boldsymbol{\xi}^k$, which is independent of $\boldsymbol{\xi}^j$ for $j < k$ (i.e., i.i.d. sampling). We update

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k g(\mathbf{x}^k, \boldsymbol{\xi}^k), \quad k = 1, 2, \ldots.$$

where $\alpha_k > 0$ is a steplength or step size.

Theorem (General Bound)

Let f be differentiable and convex. Let f_* be its minimal value at $\mathbf{x}_*$. Let $\{\mathbf{x}^k\}_{k=1}^T$ be the sequence generated by SGD upto T iterates and

$$\overline{\mathbf{x}}_T := \frac{1}{\lambda_T}\sum_{k=1}^{T}\alpha_k\mathbf{x}^k, \quad \lambda_T = \sum_{k=0}^{T}\alpha_k, \quad d_1 := \|\mathbf{x}^1 - \mathbf{x}_*\|.$$

Suppose

$$E[\|\mathbf{g}(\mathbf{x}^k, \boldsymbol{\xi}^k)\|^2] \leq B^2 \ \ \forall\ k \ \ \text{and for some } B > 0.$$

We then have

$$E[f(\overline{\mathbf{x}}_T) - f_*] \leq \frac{d_1^2 + B^2\sum_{k=1}^{T}\alpha_k^2}{2\sum_{k=1}^{T}\alpha_k}$$

Proof

We expand:

$$\begin{aligned}\|\mathbf{x}^{k+1}-\mathbf{x}_*\|^2 &= \|\mathbf{x}^k-\alpha_k\mathbf{g}(\mathbf{x}^k,\boldsymbol{\xi}^k)-\mathbf{x}_*\|^2\\ &= \|\mathbf{x}^k-\mathbf{x}_*\|^2-2\alpha_k\langle\mathbf{g}(\mathbf{x}^k,\boldsymbol{\xi}^k),\mathbf{x}^k-\mathbf{x}_*\rangle+\alpha_k^2\|\mathbf{g}(\mathbf{x}^k,\boldsymbol{\xi}^k)\|^2.\end{aligned}$$

Use the law of iterated expectation, we obtain:

$$\begin{aligned}&E[\langle\mathbf{g}(\mathbf{x}^k,\boldsymbol{\xi}^k),\mathbf{x}^k-\mathbf{x}_*\rangle] = E[E_{\boldsymbol{\xi}^k}[\langle\mathbf{g}(\mathbf{x}^k,\boldsymbol{\xi}^k),\mathbf{x}^k-\mathbf{x}_*\rangle \mid (\boldsymbol{\xi}^1,\cdots,\boldsymbol{\xi}^{k-1})]]\\ &= E[\langle E_{\boldsymbol{\xi}^k}[\mathbf{g}(\mathbf{x}^k,\boldsymbol{\xi}^k)\mid(\boldsymbol{\xi}^1,\cdots,\boldsymbol{\xi}^{k-1})],\mathbf{x}^k-\mathbf{x}_*\rangle]\\ &= E[\langle\nabla f(\mathbf{x}^k),\mathbf{x}^k-\mathbf{x}_*\rangle].\end{aligned}$$

The key argument here is that $\mathbf{x}^k$ is a random variable of $\boldsymbol{\xi}^1,\ldots,\mathbf{x}^{k-1}$, and is independent of $\boldsymbol{\xi}^k$.

Let $d_k^2 := E[\|\mathbf{x}^k - \mathbf{x}_*\|^2]$. Then we have

$$d_{k+1}^2 \leq d_k^2 - 2\alpha_k E[\langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}_* \rangle] + \alpha_k^2 B^2. \tag{2}$$

By the convexity of f, we have

$$\begin{aligned}
E[f(\overline{\mathbf{x}}_T) - f_*] &\leq E\left[\frac{1}{\lambda_T}\sum_{k=0}^{T} \alpha_k (f(\mathbf{x}^k) - f(\mathbf{x}_*))\right] \\
&\leq \frac{1}{\lambda_T}\sum_{k=1}^{T} \alpha_k E[\langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}_* \rangle] \\
&\leq \frac{1}{\lambda_T}\sum_{k=1}^{T} \left(\frac{1}{2}(d_k^2 - d_{k+1}^2) + \frac{1}{2}\alpha_k^2 B^2\right) \\
&= \frac{d_1^2 - d_{T+1}^2 + B^2 \sum_{k=1}^{T} \alpha_k^2}{2\lambda_T} \leq \frac{d_1^2 + B^2 \sum_{k=1}^{T} \alpha_k^2}{2\lambda_T}.
\end{aligned}$$

□

Optimal Bound under Constant Step size

Define

$$\Xi(\alpha_1, \ldots, \alpha_T) := \frac{d_1^2 + B^2 \sum_{k=1}^T \alpha_k^2}{2\lambda_T}.$$

Consider the constant step size rule: $\alpha_k \equiv \alpha$ and denote:

$$\Xi(\alpha) := \Xi(\alpha, \ldots, \alpha) = \frac{d_1^2 + TB^2\alpha^2}{2T\alpha} = \frac{d_1^2}{2T} \times \frac{1}{\alpha} + \frac{B^2}{2}\alpha.$$

We note that $\Xi(\alpha)$ is convex. It reaches its optimal solution at

$$\alpha_* = \frac{d_1}{B\sqrt{T}} \quad \text{with} \quad \Xi_* = \Xi(\alpha_*) = \frac{Bd_1}{\sqrt{T}}.$$

That is, the optimal constant step size is α_* giving the lowest error bound Ξ_* among all constant step sizes.

Part-II: SGD (Result 1)

Theorem ($1/\sqrt{T}$-complexity under constant step size)

Let f be differentiable and convex. Let f_* be its minimal value at $\mathbf{x}_*$. Consider the SGD with a constant step size α. Let

$$\alpha_* = \frac{d_1}{B\sqrt{T}}, \qquad \theta := \frac{\alpha}{\alpha_*}.$$

Then we have the bound

$$E[f(\overline{\mathbf{x}}_T) - f_*] \leq \left(\frac{1}{2}\theta + \frac{1}{2}\theta^{-1}\right)\frac{Bd_1}{\sqrt{T}}.$$

Proof. We have from General Bound Theorem

$$E[f(\overline{\mathbf{x}}_T) - f_*] \leq \Xi(\alpha) = \frac{d_1^2}{2T} \times \frac{1}{\alpha} + \frac{B^2}{2}\alpha$$
$$= \frac{d_1^2}{2T}\frac{1}{\theta}\frac{1}{\alpha_*} + \frac{\theta B^2}{2}\alpha_* = \left(\frac{1}{2}\theta + \frac{1}{2}\theta^{-1}\right)\frac{Bd_1}{\sqrt{T}}. \qquad \square$$

Theorem ($\log T/T$-complexity under diminishing step size)

Let f be differentiable and μ-strongly convex. Let f_* be its minimal value at $\mathbf{x}_*$. Consider the SGD with diminishing step size

$$\alpha_k = \frac{1}{k\mu}. \quad \textit{Denote} \quad \overline{\mathbf{x}}_T = \frac{1}{T}\sum_{k=1}^{T} \mathbf{x}^k.$$

Then we have

$$E[f(\overline{\mathbf{x}}_T) - f_*] \leq \frac{B^2}{2\mu T}\Big(1 + \log T\Big).$$

Proof

- In general, we have (see (2)):

$$d_{k+1}^2 \leq d_k^2 - 2\alpha_k E[\langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}_* \rangle] + \alpha_k^2 B^2,$$

where $d_k^2 := E[\|\mathbf{x}^k - \mathbf{x}_*\|^2]$. Therefore,

$$E[\langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}_* \rangle] \leq \frac{d_k^2 - d_{k+1}^2}{2\alpha_k} + \frac{\alpha_k}{2} B^2.$$

- From strong convexity, we have

$$f(\mathbf{x}_*) \geq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x}_* - \mathbf{x}^k \rangle + \frac{\mu}{2}\|\mathbf{x}^k - \mathbf{x}^*\|^2$$

and hence

$$E[\langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}_* \rangle] \geq E[f(\mathbf{x}^k) - f_* + \frac{\mu}{2}\|\mathbf{x}^k - \mathbf{x}^*\|^2]$$

Proof

- Putting the two bounds together yields

$$E[f(\mathbf{x}^k) - f_*] \leq \frac{(1-\mu\alpha_k)d_k^2 - d_{k+1}^2}{2\alpha_k} + \frac{\alpha_k}{2}B^2$$

$$= \frac{\mu}{2}\Big((k-1)d_k^2 - kd_{k+1}^2\Big) + \frac{B^2}{2\mu} \times \frac{1}{k} \quad \text{(used } \alpha_k = 1/(k\mu)\text{)}$$

By convexity, we have

$$E[f(\overline{\mathbf{x}}_T) - f_*] \leq \frac{1}{T}E[\sum_{k=1}^{T}(f(\mathbf{x}^k) - f_*)] = \frac{1}{T}\sum_{k=1}^{T}E[f(\mathbf{x}^k) - f_*]$$

$$\leq \frac{\mu}{2T}\underbrace{\sum_{k=1}^{T}\Big((k-1)d_k^2 - kd_{k+1}^2\Big)}_{=-Td_{T+1}^2} + \frac{B^2}{2\mu T}\underbrace{\sum_{k=1}^{T}\frac{1}{k}}_{\leq(1+\log T)}$$

$$\leq \frac{B^2}{2\mu T}\Big(1+\log T\Big) \qquad \square$$

Part-III: Randomized Coordinate Descent (RCD)

- Problem:

$$\min_{x_i} f(x_1, \cdots, x_n)$$

  where $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex and differentiable.

- L-smoothness in Coordinates:

$$\left|\nabla_i f(\mathbf{x} + h\mathbf{e}_i) - \nabla_i f(\mathbf{x})\right| \leq L_i|h| \quad \forall\ h \in \mathbb{R} \text{ and } \forall\ \mathbf{x} \in \mathbb{R}^n,$$

  where $\nabla_i f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial x_i}$ and $\mathbf{e}_i$ is the ith unit vector with the ith entry equal 1 and everywhere else equal 0.

Note: L-smoothness (encountered in AGD) implies L-smoothness in Coordinates.

Randomized Coordinate Descent Method

Given an initial point $\mathbf{x}^0$; Do for $k = 0, 1, 2, \ldots$

(1) Choose $i_k \in [n]$ randomly with a uniform distribution.

(2) Update

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L_i} \nabla_{i_k} f(\mathbf{x}^k) \mathbf{e}_{i_k}.$$

Note-1: Let $\boldsymbol{\xi}_k$ denotes the random sequence drawn upto the point $\mathbf{x}^k$:

$$\boldsymbol{\xi}_k = \{i_0, i_1, \ldots, i_k\}.$$

Therefore, $\mathbf{x}^{k+1}$ only depends on $\boldsymbol{\xi}_k$ and $\mathbf{x}^k$ only depends on $\boldsymbol{\xi}_{k-1}$.

Note-2: RCD is a stochastic algorithm, but blue cannot be as SGD framework.

Theorem

Convergence of RCD: Lipschitz Convex Case Let f be convex and L-smooth in coordinates with $L_i > 0$, $i \in [n]$. Let $\mathbf{x}_*$ be a minimizer of $f(\mathbf{x})$ and let $\{\mathbf{x}^k\}$ be a sequence generated by RCD algorithm. We then have for $k \geq 1$

$$E_{\boldsymbol{\xi}_{k-1}}[f(\mathbf{x}^k) - f(\mathbf{x}_*)] \leq \frac{n}{n+k}\left(\frac{1}{2}\|\mathbf{x}^0 - \mathbf{x}_*\|_L^2 + f(\mathbf{x}^0) - f(\mathbf{x}_*)\right),$$

where $\|\mathbf{x}\|_L^2 := \sum_{i=1}^n L_i x_i^2$ (i.e., weighted norm of $\mathbf{x}$).

Proof

Let $r_k^2 := \|\mathbf{x}^k - \mathbf{x}_*\|_L^2$. We then have

$$\begin{aligned} r_{k+1}^2 &= \|\mathbf{x}^k - \frac{1}{L_{i_k}} \nabla_{i_k} f(\mathbf{x}^k) \mathbf{e}_{i_k} - \mathbf{x}_*\|_L^2 \\ &= r_k^2 - 2\nabla_{i_k} f(\mathbf{x}^k)(\mathbf{x}^k[i_k] - \mathbf{x}_*[i_k]) + \frac{1}{L_{i_k}} \left(\nabla_{i_k} f(\mathbf{x}^k)\right)^2 \end{aligned}$$

It follows from the L-smoothness in coordinates that

$$\begin{aligned} f(\mathbf{x}^{k+1}) &\leq f(\mathbf{x}^k) + \nabla_{i_k} f(\mathbf{x}^k)(\mathbf{x}^{k+1}[i_k] - \mathbf{x}^k[i_k]) \\ &\quad + \frac{L_{i_k}}{2}(\mathbf{x}^{k+1}[i_k] - \mathbf{x}^k[i_k])^2 \\ &= f(\mathbf{x}^k) - \frac{1}{L_{i_k}}(\nabla_{i_k} f(\mathbf{x}^k))^2. \end{aligned} \tag{3}$$

Combining the above two relations, we have

$$r_{k+1}^2 \leq r_k^2 - 2\nabla_{i_k} f(\mathbf{x}^k)(\mathbf{x}^k[i_k] - \mathbf{x}_*[i_k]) + 2(f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}))$$

Taking expectation with respect to i_k yields

$$E_{i_k}[r_{k+1}^2/2] \leq \frac{1}{2}r_k^2 - \frac{1}{n}\langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}_* \rangle + f(\mathbf{x}^k) - E_{i_k}[f(\mathbf{x}^{k+1})].$$

By convexity of f, we have

$$\nabla f(\mathbf{x}^k), \mathbf{x}_* - \mathbf{x}^k \rangle \leq f(\mathbf{x}_*) - f(\mathbf{x}^k)$$

and hence

$$E_{i_k}[r_{k+1}^2/2] \leq \frac{1}{2}r_k^2 + \frac{1}{n}f(\mathbf{x}_*) + \frac{n-1}{n}f(\mathbf{x}^k) - E_{i_k}[f(\mathbf{x}^{k+1})]$$

Rearranging to get

$$E_{i_k}\left[\frac{1}{2}r_{k+1}^2 + f(\mathbf{x}^{k+1}) - f_*\right] \leq \left(\frac{1}{2}r_k^2 + f(\mathbf{x}^k) - f_*\right) - \frac{1}{n}(f(\mathbf{x}^k) - f_*)$$

Taking expectation with respect to $\mathbf{x}_{k-1}$ on both sides yields

$$E_{\boldsymbol{\xi}_k}\left[\frac{1}{2}r_{k+1}^2+f(\mathbf{x}^{k+1})-f_*\right] \leq E_{\boldsymbol{\xi}_{k-1}}\left[\frac{1}{2}r_k^2+f(\mathbf{x}^k)-f_*\right]-\frac{E_{\boldsymbol{\xi}_{k-1}}[f(\mathbf{x}^k)-f_*]}{n}$$

It follows from (3) that

$$E_{\boldsymbol{\xi}_j}[f(\mathbf{x}^{j+1})] \leq E_{\boldsymbol{\xi}_{j-1}}[f(\mathbf{x}^j)]$$

We therefore have

$$\begin{aligned} E_{\boldsymbol{\xi}_k}\left[f(\mathbf{x}^{k+1})-f_*\right] &\leq E_{\boldsymbol{\xi}_k}\left[\frac{1}{2}r_{k+1}^2+f(\mathbf{x}^{k+1})-f_*\right] \\ &\leq \frac{1}{2}r_0^2+f(\mathbf{x}^0)-f_*-\frac{1}{n}\sum_{j=0}^{k}\left(E_{\boldsymbol{\xi}_{j-1}}[f(\mathbf{x}^j)]-f_*\right) \\ &\leq \frac{1}{2}r_0^2+f(\mathbf{x}^0)-f_*-\frac{k+1}{n}\left(E_{\boldsymbol{\xi}_k}[f(\mathbf{x}^{k+1})]-f_*\right) \end{aligned}$$

Hence

$$E_{\boldsymbol{\xi}_k}\left[f(\mathbf{x}^{k+1})-f_*\right] \leq \frac{n}{n+k+1}\left(\frac{1}{2}r_0^2+f(\mathbf{x}^0)-f_*\right) \qquad \square$$

Numerical Example

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top(AA^\top + nI)\mathbf{x} + \mathbf{b}^\top\mathbf{x}.$$

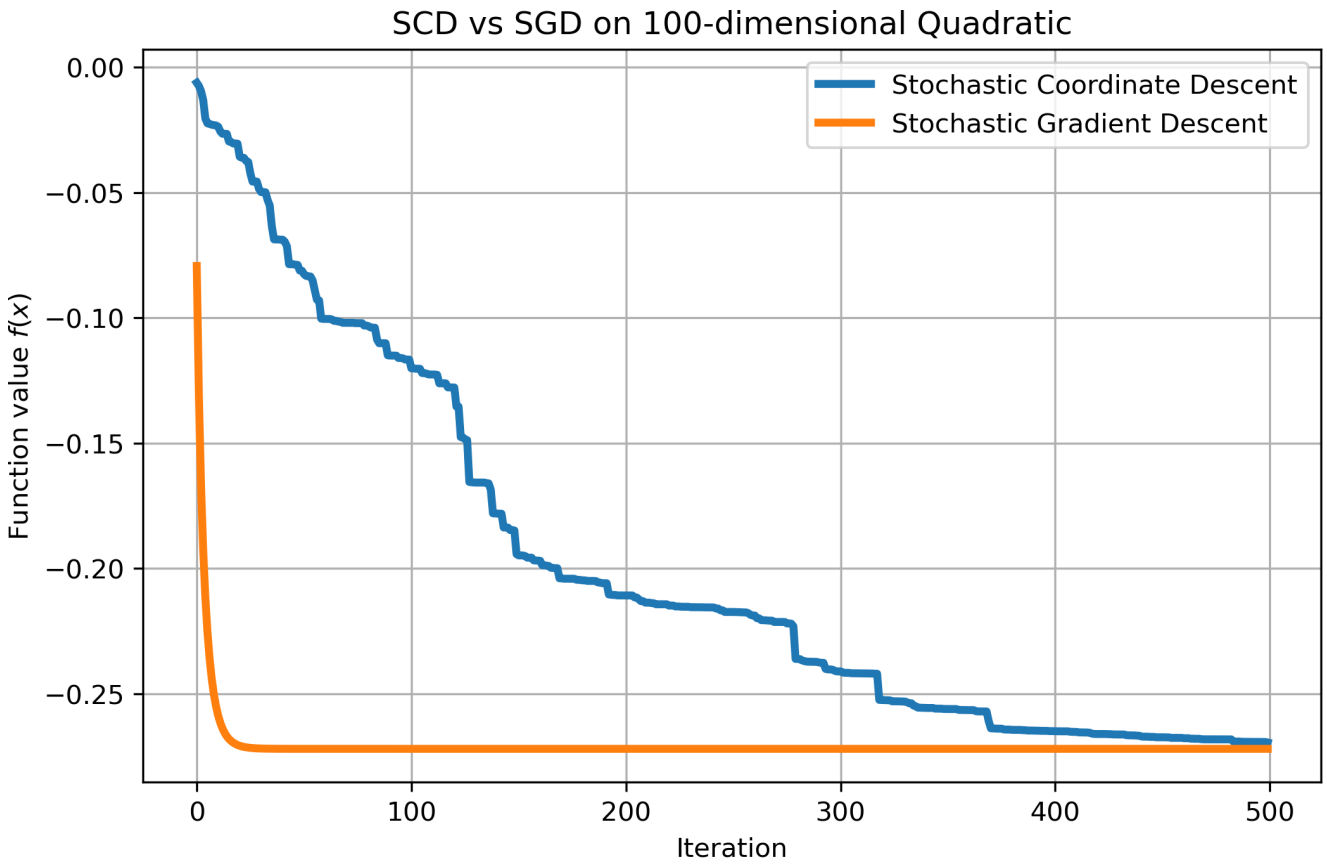


Figure: Comparison of SGD and SCD

Summary

- SGD naturally rises from ML (e.g., Perceptron Machine) with guarantee of correctness.
- For L-smooth and convex optimization, we can have $O(1/\sqrt{T})$-complexity.
- For L-smooth and strongly convex optimization, we can have $O(\log T/T)$-complexity. This can be improved to $O(1/T)$ complexity with appropriate step size.
- For L-smooth-in-coordinates and convex optimization, SCD has a complexity $O(n/(n+T))$-complexity. In practice, SCD is much slower than SGD.