Machine Learning

Lecture 8: Optimization 2

Hao Tang

October 10, 2022

- For log loss

$$L = \sum_{i=1}^{N} \log\left(1 + \exp(-y_i w^\top \phi(x_i))\right) \tag{1}$$

we cannot even solve $\frac{\partial L}{\partial w} = 0$.

- How do we find the optimal solution?

- Could we find an approximate solution?

Approximate solutions in optimization

- We say that $\hat{x}$ is an approximate solution if, for a given $\epsilon > 0$,

$$f(\hat{x}) - f(x^*) < \epsilon. \tag{2}$$

- Note that it is close in function value, not close in the input.

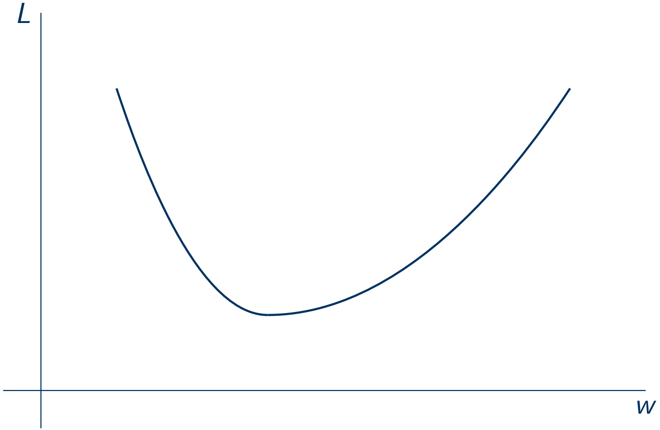
Gradient descent

- Gradient descent is an iterative algorithm, consisting of the steps

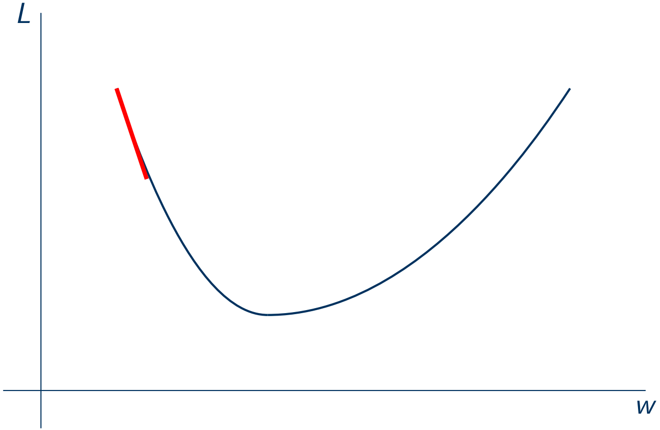
$$w_{t+1} = w_t - \eta_t \nabla L(w_t). \tag{3}$$

- The variable $\eta_t > 0$ is called the step size, and can depend on t.

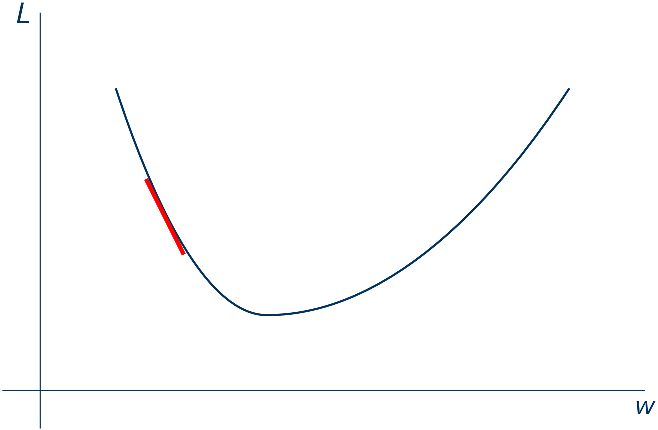
Gradient descent



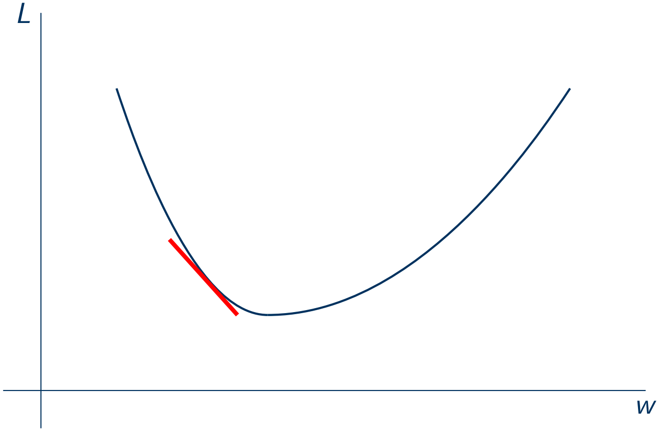
Gradient descent



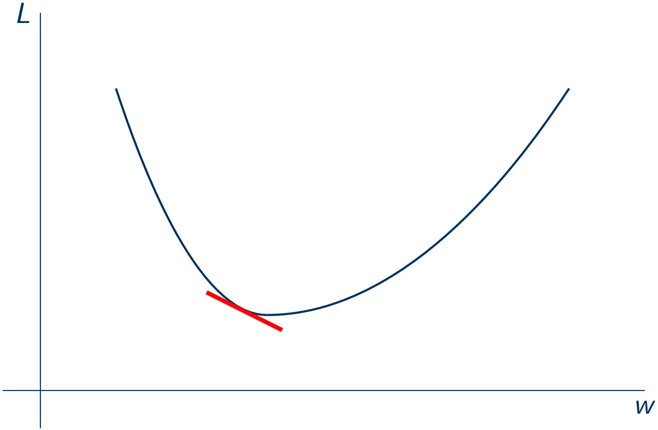
Gradient descent



Gradient descent



Gradient descent



Approximate solutions for iterative algorithms

- An iterative algorithm creates a sequence $x_1, \ldots, x_t$.
- How many updates do we need to achieve an approximate solution?
- Given $\epsilon > 0$, how large does t needs to be to achieve

$$f(x_t) - f(x^*) < \epsilon? \tag{4}$$

- We want to express ϵ as a function of t.

Potential results

- Sublinear
  - $f(x_t) - f(x^*) \leq \frac{c}{t^2}$
- Linear
  - $f(x_t) - f(x^*) \leq cr^t$ for $0 < r < 1$
- Quadratic
  - $f(x_t) - f(x^*) \leq cr^{2^t}$ for $0 < r < 1$

Potential results

- Sublinear
  - $f(x_t) - f(x^*) \leq \frac{c}{t^2}$
  - $\epsilon = O\left(\frac{1}{t^2}\right)$ or $t = O(\frac{1}{\sqrt{\epsilon}})$
- Linear
  - $f(x_t) - f(x^*) \leq cr^t$ for $0 < r < 1$
  - $\epsilon = O\left(2^{-t}\right)$ or $t = O(\log \frac{1}{\epsilon})$
- Quadratic
  - $f(x_t) - f(x^*) \leq cr^{2^t}$ for $0 < r < 1$
  - $\epsilon = O\left(2^{2^{-t}}\right)$ or $t = O(\log \log \frac{1}{\epsilon})$

Implications of smoothness

- Based on the definition of smoothness and gradient update,

$$
\begin{aligned}
f(x_t) &\le f(x_{t-1}) + (x_t - x_{t-1})^\top \nabla f(x_{t-1}) + L\|x_t - x_{t-1}\|_2^2 && (5)\\
&= f(x_{t-1}) - \eta_t \|\nabla f(x_{t-1})\|_2^2 + L\eta_t^2 \|\nabla f(x_{t-1})\|_2^2 && (6)\\
&= f(x_{t-1}) - \eta_t(1 - L\eta_t)\|\nabla f(x_{t-1})\|_2^2 && (7)
\end{aligned}
$$

- In other words, $f(x_{t-1}) - f(x_t) \ge \eta_t(1 - L\eta_t)\|\nabla f(x_{t-1})\|_2^2$.

- The expression $\eta_t(1 - L\eta_t)$ has a maximum $\frac{1}{4L}$ when $\eta_t = \frac{1}{2L}$, and reaches 0 when $\eta_t = \frac{1}{L}$.

- Choosing any $\eta_t \in \left[\frac{1}{2L}, \frac{1}{L}\right)$ is able to stricly decrease the objective.

- For simplicity, we choose $\eta_t = \frac{1}{2L}$ for the rest of the analysis.

Implications of strong convexity

- Based on the definition of strong convexity,

$$f(x) \geq f(y) + (x-y)^\top \nabla f(y) + \frac{\mu}{2}\|x-y\|_2^2. \tag{8}$$

- The best x on the right-hand side is $x = y - \frac{1}{\mu}\nabla f(y)$.

- We have $f(x) \geq f(y) - \frac{1}{2\mu}\|\nabla f(y)\|_2^2$, for any x and y.

- In particular, $f(y) - f(x^*) \leq \frac{1}{2\mu}\|\nabla f(y)\|_2^2$.

- In words, the gradient norm at any given point tells us how far we are from the optimal value.

Guarentee of gradient descent

- If we do gradient descent on a L-smooth and μ-strongly convex function,

$$
\begin{aligned}
f(x_t) - f(x^*) &\leq f(x_{t-1}) - f(x^*) - \frac{1}{4L}\|\nabla f(x_{t-1})\|_2^2 && (9)\\
&\leq f(x_{t-1}) - f(x^*) + \frac{\mu}{2L}(f(x^*) - f(x_{t-1})) && (10)\\
&= \left(1 - \frac{\mu}{2L}\right)(f(x_{t-1}) - f(x^*)) && (11)\\
&= \left(1 - \frac{\mu}{2L}\right)^t (f(x_0) - f(x^*)) && (12)
\end{aligned}
$$

- The convergence rate is linear.

Guarentee of gradient descent

- If we do gradient descent on a L-smooth convex function,

$$f(x_t) - f(x^*) \leq \frac{\|x_0 - x^*\|_2^2}{2\eta t} \tag{13}$$

  for $\eta \leq 1/L$.

- The convergence rate is sublinear.

- The proof is beyond the scope of this course.

Back to log loss

- The log loss in the binary case

$$L = \sum_{i=1}^{N} \log \left(1 + \exp(-y_i w^\top \phi(x_i)) \right). \tag{14}$$

- We have shown that L is convex in w.

Gradient descent on log loss

$$\frac{\partial L}{\partial w} = \sum_{i=1}^{N} \frac{\exp(-y_i w^\top \phi(x_i))}{1 + \exp(-y_i w^\top \phi(x_i))} (-y_i \phi(x_i)) \tag{15}$$

$$= \sum_{i=1}^{N} \left(1 - \frac{1}{1 + \exp(-y_i w^\top \phi(x_i))}\right) (-y_i \phi(x_i)) \tag{16}$$

$$= \sum_{i=1}^{N} (1 - p(y_i|x_i)) (-y_i \phi(x_i)) \tag{17}$$

The size of the data set

- For mean-squared error, recall that the solution for $\frac{\partial L}{\partial w} = 0$ is $w^* = (\Phi\Phi^\top)^{-1}\Phi y$.
- Computing $(\Phi\Phi^\top)^{-1}\Phi y$ takes $O(N^3)$.
- For gradient descent on log loss, computing the gradient itself takes $O(N)$.

Stochastic gradient descent

1. Sample x_t, y_t from a data set S.
2. $w_{t+1} = w_t - \eta_t \nabla \ell(w_t; x_t, y_t)$
   - Per sample L_2 loss $\ell(w; x, y) = (w^\top \phi(x) - y)^2$
   - Per sample log loss $\ell(w; x, y) = \log(1 + \exp(-y w^\top \phi(x)))$
3. Go to 1 until the solution is satisfactory.

Stochastic gradient descent

- $\nabla \ell(w_t; x_t, y_t)$ is now random, because x_t and y_t is random.
- The expectation

$$\mathbb{E}_{x,y \sim U(S)}[\nabla \ell(w; x, y)] = \nabla L(w) \tag{18}$$

where $U(S)$ is the uniform distribution over the samples in S.

Guarantee for stochastic gradient descent

- If we do SGD on an γ-smooth convex function,

$$\mathbb{E}_{x,y\sim U(S)}[L(\bar{w}_t)] \leq L(w^*) + \frac{\|w_0 - w^*\|_2^2}{2\eta t} + \frac{t\sigma^2}{2} \quad (19)$$

  where $\eta \leq 1/\gamma$.

- $\sigma^2 \geq \mathbb{E}_{x,y\sim U(S)}\left[\|\nabla\ell(w_t; x, y)\|^2\right] - \left\|\mathbb{E}_{x,y\sim U(S)}[\nabla\ell(w_t; x, y)]\right\|_2^2$ for any t

- $\bar{w}_t = \frac{w_1 + \cdots + w_t}{t}$

- The proof is beyond the scope of this course.

- The runtime is $O(t)$, independent of the data set size N!

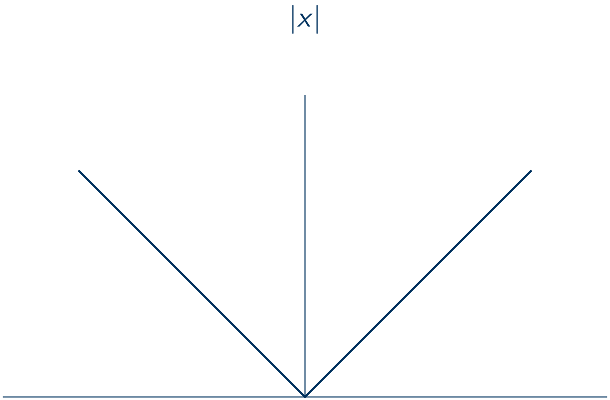
Subgradient

- A subgradient at x is a vector g that satisfies

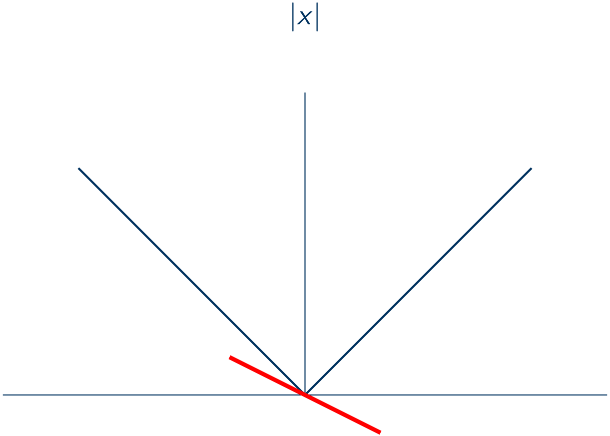
$$f(y) \geq f(x) + (y - x)^\top g \tag{20}$$

  for any y, and the set of subgradients at x is denoted as $\partial f(x)$.
- Obviously, $\nabla f(x) \in \partial f(x)$, if $\nabla f(x)$ exists.
- Convergence theorems can be ported to subgradient descent.

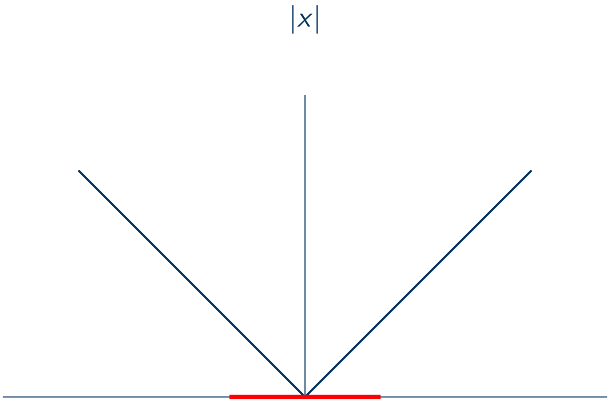
Subgradients for absolute values



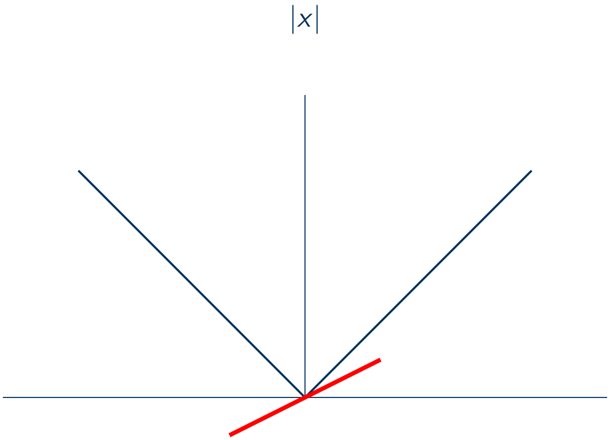
Subgradients for absolute values



Subgradients for absolute values



Subgradients for absolute values



Examples

- $f(x) = x^2$ is 2-strongly convex.
- $f(x) = |x|$ is convex and 1-Lipschitz.
- This also implies that mean-squared error is strongly convex function.
- $f(x) = \|x\|_2^2$ is 2-strongly convex.
- $g(x) = f(x) + \|x\|_2^2$ is strongly convex if f is convex.

Check your understanding

- What does it mean to get an approximate solution for an optimization problem?
- What is gradient descent?
- What is stochastic gradient descent?
- What does it mean to have an approximate solution for an iterative algorithm?
- What are sublinear, linear, quadratic convergence?