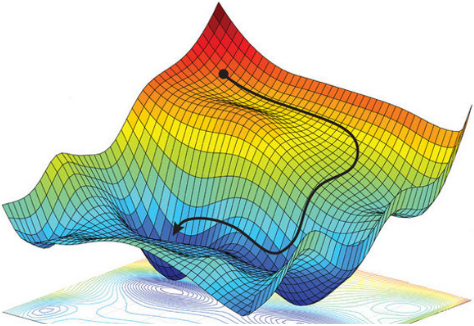
Optimization

Chi Jin

Princeton University.

Optimization

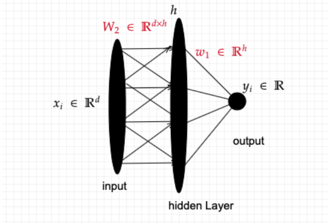
$$\min_{x \in \mathcal{X}} \; f(x)$$



ML Examples

Regression: given training dataset $\{(x_i, y_i)\}_{i=1}^n$ s.t.

$$y_i = (w_1^\star)^T \sigma((W_2^\star)^T x_i) + \epsilon_i$$



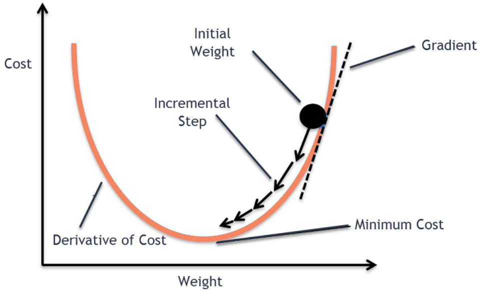
Learning unknown weights by

$$\min_{w=(w_1, W_2)} f(w) := \sum_{i=1}^n (y_i - w_1^T \sigma(W_2^T x_i))^2.$$

Gradient Descent (GD)

$$x_0 = \text{initialization}$$
$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

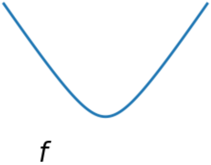
η: learning rate



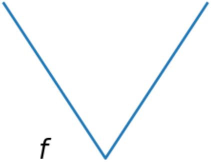
Smooth Functions

ℓ-smoothness

$$\|\nabla^2 f(x)\| \leq \ell$$



smooth



nonsmooth

$$f(x) \leq \underbrace{f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{\ell}{2}\|x - x_t\|^2}_{\text{2nd order Taylor expansion as an upper bound}}$$

GD Interpretation

GD as optimizing the smooth upper bound:

$$x_{t+1} = \operatorname*{argmin}_x \underbrace{f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{\ell}{2} \|x - x_t\|^2}_{\text{2nd order Taylor expansion as an upper bound}}$$

equivalent to

$$x_{t+1} = x_t - \frac{1}{\ell} \nabla f(x_t)$$

Natural choice of learning rate $\eta = 1/\ell$.

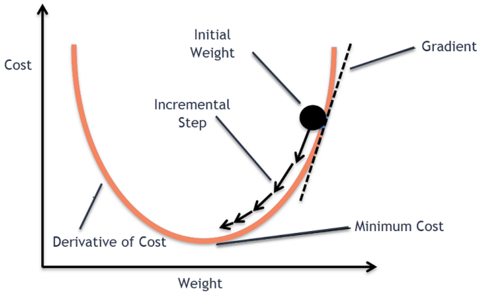
Descent Lemma

Lemma

If f is ℓ-smooth, then GD with learning rate $\eta \leq 1/\ell$ satisfies

$$f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|^2.$$

GD monotonically decreases function value!



Descent Lemma II

Lemma

If f is ℓ-smooth, then GD with learning rate $\eta \le 1/\ell$ satisfies

$$f(x_{t+1}) \le f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|^2.$$

Proof. By smoothness

$$\begin{aligned}
f(x_{t+1}) &\le f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{\ell}{2}\|x_{t+1} - x_t\|^2 \\
&\overset{(a)}{\le} f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{1}{2\eta}\|x_{t+1} - x_t\|^2 \\
&\overset{(b)}{=} f(x_t) - \eta\|\nabla f(x_t)\|^2 + \frac{\eta}{2}\|\nabla f(x_t)\|^2 = f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|^2.
\end{aligned}$$

(a) by $\eta \le 1/\ell$; (b) by GD update.

Optimization Questions

- When can GD find the minimizer?
- How fast can GD find the minimizer?
- Faster algorithms?
- Gradient has noise?
- ...

Overview

- Introduction to Optimization
  - Gradient descent
  - Smooth function and descent lemma
- Convex Optimization
  - Convexity and GD guarantees
  - Acceleration
  - Stochastic gradient descent
- Nonconvex Optimization
  - Finding stationary points
  - Escaping saddle points

Convex Optimization

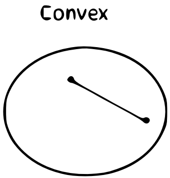
Convex sets and functions

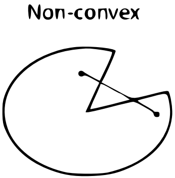
- A set $\mathcal{X} \subseteq \mathbb{R}^n$ is convex if

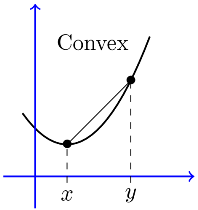
$$\forall (x, y, \gamma) \in \mathcal{X} \times \mathcal{X} \times [0, 1] : \; (1 - \gamma)x + \gamma y \in \mathcal{X}.$$

- A function $f : \mathcal{X} \to \mathbb{R}$ is convex if

$$\forall (x, y, \gamma) \in \mathcal{X} \times \mathcal{X} \times [0, 1] : \; f\big((1 - \gamma)x + \gamma y\big) \leq (1 - \gamma) f(x) + \gamma f(y).$$



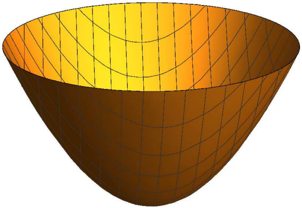




Convex Optimization

$$\min_{x \in \mathcal{X}} f(x)$$

$\mathcal{X}$ is convex set, f is convex function



Properties of Convex Optimization

Proposition

A local min of a convex function is also a global min!

This enables local search (GD) as global optimizers.

Proof. Suppose x is a local min of f, for any $y \in \mathcal{X}$, there exists $\gamma \in (0, 1]$ s.t.

$$f(x) \leq f((1-\gamma)x + \gamma y) \leq (1-\gamma)f(x) + \gamma f(y).$$

This implies $f(x) \leq f(y)$, i.e., x is also a global min of f.

GD Convergence Guarantees

Theorem

If f is ℓ-smooth and convex, then GD with learning rate $\eta = 1/\ell$ satisfies

$$f(x_t) - f(x^\star) \le \frac{2\ell \|x_0 - x^\star\|^2}{t}.$$

To achieve ϵ-optimality, i.e., $f(x_t) - f(x^\star) \le \epsilon$, GD needs no more than $\mathcal{O}(\ell \|x_0 - x^\star\|^2/\epsilon)$ iterations, independent of dimension!

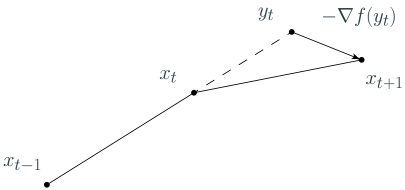
Acceleration

Can we find ϵ-optimal points faster than GD?

Accelerated Gradient Descent (AGD):

$$\begin{aligned} y_t &\leftarrow x_t + \gamma(x_t - x_{t-1}), \\ x_{t+1} &\leftarrow y_t - \eta \nabla f(y_t). \end{aligned}$$

γ: momentum parameter



AGD Guarantees

Theorem

If f is ℓ-smooth and convex, then AGD with proper parameters satisfies

$$f(x_t) - f(x^\star) \leq \mathcal{O}\left(\frac{\ell\|x_0 - x^\star\|^2}{t^2}\right).$$

- faster than GD rate — $\mathcal{O}(\ell\|x_0 - x^\star\|^2/t)$.
- information-theoretically optimal!

Stochastic Optimization

What if we only have access to noisy version of gradient $g(\cdot)$:

- $\mathbb{E}g(x) = \nabla f(x)$
- $\mathrm{Var}(g(x)) := \mathbb{E}\|g(x) - \mathbb{E}g(x)\|^2 \le \sigma^2$

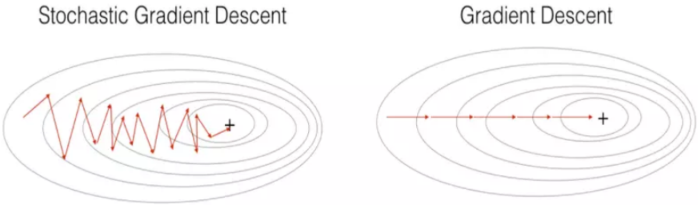
In ML, $\min_x F(x) := \sum_{i=1}^n f_i(x)$, then

$\nabla f_i(x)$ with uniformly random $i \in [n]$ is a stochastic gradient for F.

Stochastic Gradient Descent (SGD)

$$x_{t+1} = x_t - \eta g(x_t).$$

Despite noise, SGD makes gradient update on average with small η.



SGD Guarantees

Theorem

If f is ℓ-smooth and convex, let $R = \|x^* - x_1\|^2$, then SGD with $\eta = \min\{\frac{1}{\ell}, \frac{R}{\sigma\sqrt{2t}}\}$ gives

$$\mathbb{E} f(\underbrace{\bar{x}_t}_{\text{average iterate}}) - f(x^*) \leq \underbrace{\frac{\ell R^2}{2t}}_{\text{rate for GD}} + \underbrace{R\sigma\sqrt{\frac{2}{t}}}_{\text{extra error due to SG}}$$

SGD is still capable of finding global min efficiently.

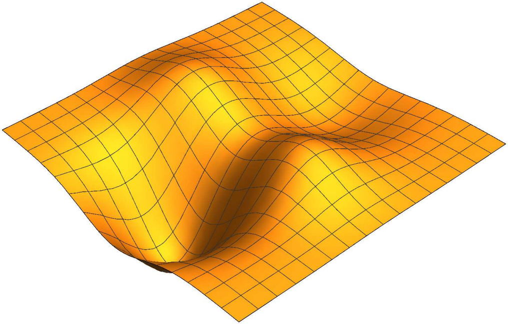
Larger noise σ leads to smaller learning rate and slower convergence.

Nonconvex Optimization

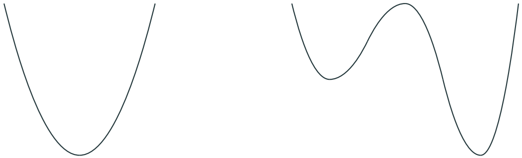
Nonconvex Optimization

$$\min_{x \in \mathcal{X}} \; f(x)$$

f is not a convex function



Hardness of nonconvex optimization



Convex vs nonconvex functions.

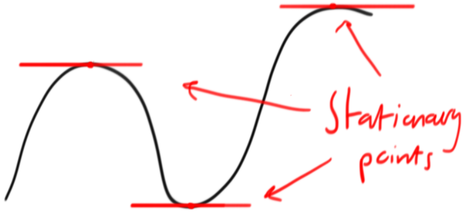
Proposition for nonconvex functions

(1) local min is not necessarily a global min; (2) Finding global min requires an exponential number of gradient queries in the worst case!

Solution: find local surrogates.

Stationary Points

- x is a stationary point if $\|\nabla f(x)\| = 0$.
- x is an ϵ-stationary point if $\|\nabla f(x)\| \leq \epsilon$.



Guarantees for Stationary Points

Theorem (Nesterov 1998)

If f is ℓ-smooth, then after running GD with $\eta = 1/\ell$ for

$$\frac{2\ell(f(x_0) - f(x^\star))}{\epsilon^2}$$

iterations, at least one of the iterates will be an ϵ-stationary point.

Not a last iterate guarantee.

Guarantees for Stationary Points

Theorem (Nesterov 1998)

If f is ℓ-smooth, then after running GD with $\eta = 1/\ell$ for

$$\frac{2\ell(f(x_0) - f(x^\star))}{\epsilon^2}$$

iterations, at least one of the iterates will be an ϵ-stationary point.

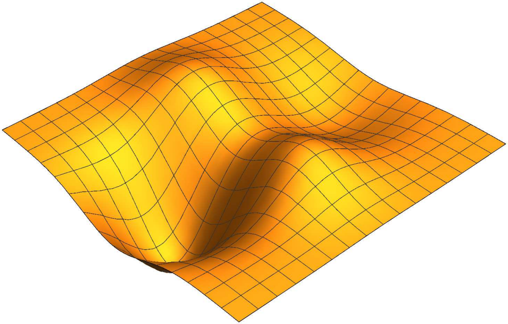
Proof. Assume the contrary, all iterates are non-ϵ-stationary, by descent lemma

$$f(x_{t+1}) - f(x_t) \leq -\frac{\eta}{2}\|\nabla f(x_t)\|^2 \leq -\frac{\epsilon^2}{2\ell}$$

Note function value can not decrease by more than $f(x_0) - f(x^\star)$.

Drawbacks of Stationary Points

Stationary points can be local min, local max or even saddle points.



Second-order Stationary Points

Want to only find "approximate local min".

- x is a second-order stationary point if

$$\|\nabla f(x)\| = 0, \text{ and } \nabla^2 f(x) \succeq 0.$$

- x is an ϵ-second-order stationary point if

$$\|\nabla f(x)\| \leq \epsilon, \text{ and } \nabla^2 f(x) \succeq -\sqrt{\rho\epsilon}.$$

where ρ is the second-order smooth parameter s.t. $\|\nabla^3 f(x)\| \leq \rho$.

Escaping Saddle Points

GD will stuck if initialized at local max or saddle points.

Solution: add perturbations!

Perturbed Gradient Descent (PGD)

$$x_{t+1} = x_t - \eta(\nabla f(x_t) + \zeta_t),$$

where $\zeta_t \sim \mathcal{N}(0, (r^2/d) \cdot I)$ and $r = \tilde{\Theta}(\epsilon)$.

Guarantees for Second-order Stationary Points

Theorem (Jin et al. 2015)

If f is ℓ-smooth and ρ-second-order smooth, then after running PGD with $\eta = 1/\ell$ and $r = \tilde{\Theta}(\epsilon)$ for

$$\tilde{\mathcal{O}}\left(\frac{\ell(f(x_0) - f(x^\star))}{\epsilon^2}\right)$$

iterations, one of the iterates will be an ϵ-second-order stationary point.

Strengthen the original stationary point guarantee to approximate local min by paying only logarithmic factors in iteration complexity!

Summary

- Introduction to Optimization
  - Gradient descent
  - Smooth function and descent lemma
- Convex Optimization
  - Convexity and GD guarantees
  - Acceleration
  - Stochastic gradient descent
- Nonconvex Optimization
  - Finding stationary points
  - Escaping saddle points

Advanced Topics in Optimization

- High-order algorithms.
- Nonsmooth optimization.
- Adaptive / parameter-free algorithms.
- Distributed optimization.
- Minimax optimization.
- ...