Gradient Descent

Hoàng Nam Dũng

Khoa Toán - Cơ - Tin học, Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội

Gradient descent

Consider unconstrained, smooth convex optimization

$$\min_x f(x)$$

with convex and differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. Denote the optimal value by $f^* = \min_x f(x)$ and a solution by x^*.

Gradient descent

Consider unconstrained, smooth convex optimization

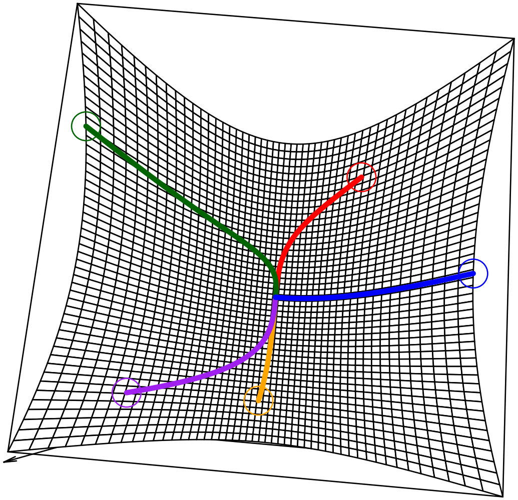
$$\min_x f(x)$$

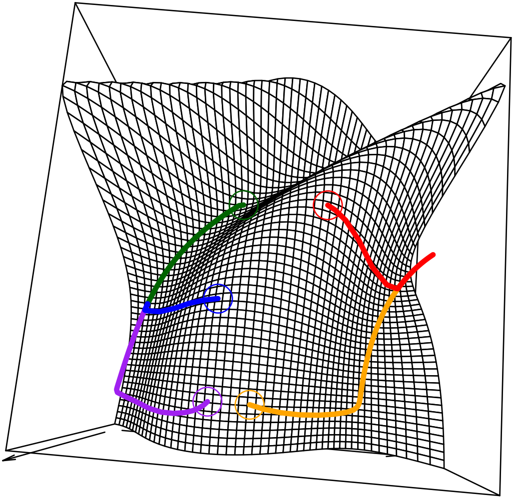
with convex and differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. Denote the optimal value by $f^* = \min_x f(x)$ and a solution by x^*.

Gradient descent: choose initial point $x^{(0)} \in \mathbb{R}^n$, repeat:

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f(x^{(k-1)}), \quad k = 1, 2, 3, \dots$$

Stop at some point.





Gradient descent interpretation

At each iteration, consider the expansion

$$f(y) \approx f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t} \|y - x\|_2^2 .$$

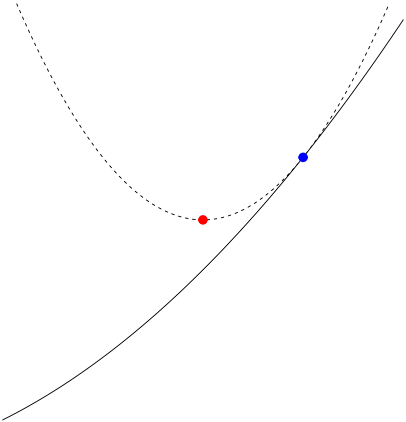
Quadratic approximation, replacing usual Hessian $\nabla^2 f(x)$ by $\frac{1}{t} I$.

$f(x) + \nabla f(x)^T (y - x)$	linear approximation to f
$\frac{1}{2t} \|y - x\|_2^2$	proximity term to x, with weight $1/2t$

Choose next point $y = x^+$ to minimize quadratic approximation

$$x^+ = x - t \nabla f(x).$$

Gradient descent interpretation



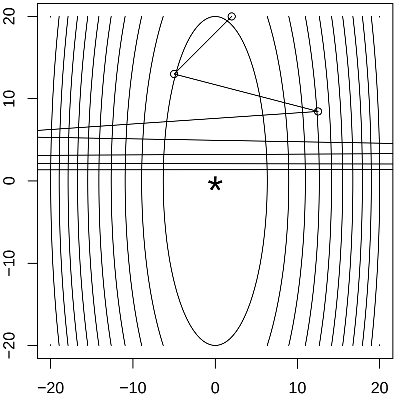
Blue point is x, red point is
$$x^* = \text{argmin}_y \, f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t} \|y - x\|_2^2$$

Outline

- How to choose step sizes
- Convergence analysis
- Nonconvex functions
- Gradient boosting

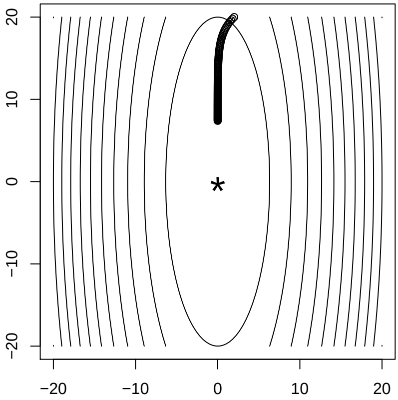
Fixed step size

Simply take $t_k = t$ for all $k = 1, 2, 3, \ldots$, can diverge if t is too big. Consider $f(x) = (10x_1^2 + x_2^2)/2$, gradient descent after 8 steps:



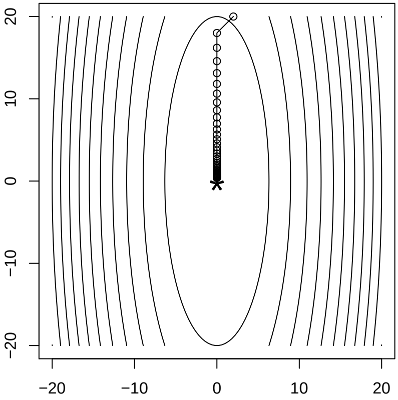
Fixed step size

Can be slow if t is too small. Same example, gradient descent after 100 steps:



Fixed step size

Converges nicely when t is "just right". Same example, 40 steps:



Convergence analysis later will give us a precise idea of "just right".

Backtracking line search

One way to adaptively choose the step size is to use backtracking line search:

- First fix parameters $0 < \beta < 1$ and $0 < \alpha \leq 1/2$.
- At each iteration, start with $t = t_{\text{init}}$, and while

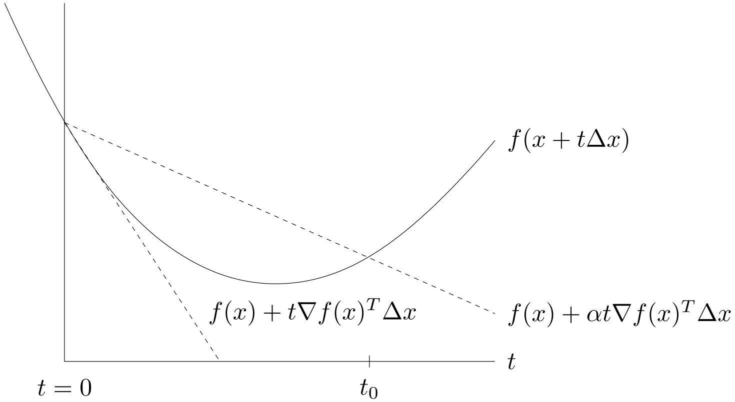
  $$f(x - t\nabla f(x)) > f(x) - \alpha t \|\nabla f(x)\|_2^2$$

  shrink $t = \beta t$. Else perform gradient descent update

  $$x^+ = x - t\nabla f(x).$$

Simple and tends to work well in practice (further simplification: just take $\alpha = 1/2$).

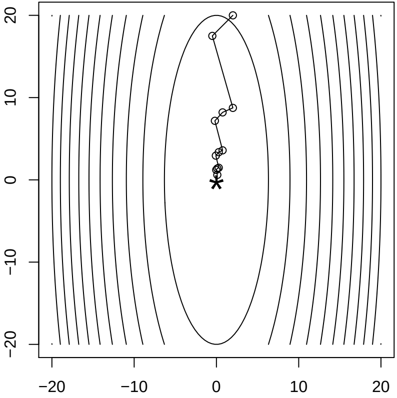
Backtracking interpretation



For us $\Delta x = -\nabla f(x)$

Backtracking line search

Setting $\alpha = \beta = 0.5$, backtracking picks up roughly the right step size (12 outer steps, 40 steps total).



Exact line search

We could also choose step to do the best we can along direction of negative gradient, called exact line search:

$$t = \mathrm{argmin}_{s \geq 0} \, f(x - s\nabla f(x)).$$

Usually not possible to do this minimization exactly.

Approximations to exact line search are typically not as efficient as backtracking and it's typically not worth it.

Convergence analysis

Assume that $f : \mathbb{R}^n \to \mathbb{R}$ convex and differentiable and additionally

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2 \quad \text{for any } x, y$$

i.e., ∇f is Lipschitz continuous with constant $L > 0$.

Theorem

Gradient descent with fixed step size $t \leq 1/L$ satisfies

$$f(x^{(k)}) - f^* \leq \frac{1}{2tk} \|x^{(0)} - x^*\|_2^2$$

and same result holds for backtracking with t replaced by β/L.

Convergence analysis

Assume that $f : \mathbb{R}^n \to \mathbb{R}$ convex and differentiable and additionally

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2 \quad \text{for any } x, y$$

i.e., ∇f is Lipschitz continuous with constant $L > 0$.

Theorem

Gradient descent with fixed step size $t \leq 1/L$ satisfies

$$f(x^{(k)}) - f^* \leq \frac{1}{2tk} \|x^{(0)} - x^*\|_2^2$$

and same result holds for backtracking with t replaced by β/L.

We say gradient descent has convergence rate $O(1/k)$, i.e., it finds ε-suboptimal point in $O(1/\varepsilon)$ iterations.

Convergence analysis

Assume that $f : \mathbb{R}^n \to \mathbb{R}$ convex and differentiable and additionally

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2 \quad \text{for any } x, y$$

i.e., ∇f is Lipschitz continuous with constant $L > 0$.

Theorem

Gradient descent with fixed step size $t \leq 1/L$ satisfies

$$f(x^{(k)}) - f^* \leq \frac{1}{2tk} \|x^{(0)} - x^*\|_2^2$$

and same result holds for backtracking with t replaced by β/L.

We say gradient descent has convergence rate $O(1/k)$, i.e., it finds ε-suboptimal point in $O(1/\varepsilon)$ iterations.

Chứng minh.

Slide 20-25 in `http://www.seas.ucla.edu/~vandenbe/236C/lectures/gradient.pdf` □

Convergence under strong convexity

Reminder: strong convexity of f means $f(x) - \frac{m}{2}\|x\|_2^2$ is convex for some $m > 0$.

Assuming Lipschitz gradient as before and also strong convexity:

Theorem

Gradient descent with fixed step size $t \le 2/(m+L)$ or with backtracking line search search satisfies

$$f(x^{(k)}) - f^* \le c^k \frac{L}{2}\|x^{(0)} - x^*\|_2^2$$

where $0 < c < 1$.

Convergence under strong convexity

Reminder: strong convexity of f means $f(x) - \frac{m}{2}\|x\|_2^2$ is convex for some $m > 0$.

Assuming Lipschitz gradient as before and also strong convexity:

Theorem

Gradient descent with fixed step size $t \le 2/(m+L)$ or with backtracking line search search satisfies

$$f(x^{(k)}) - f^* \le c^k \frac{L}{2} \|x^{(0)} - x^*\|_2^2$$

where $0 < c < 1$.

Rate under strong convexity is $O(c^k)$, exponentially fast, i.e., we find ε-suboptimal point in $O(\log(1/\varepsilon))$ iterations.

Convergence under strong convexity

Reminder: strong convexity of f means $f(x) - \frac{m}{2}\|x\|_2^2$ is convex for some $m > 0$.

Assuming Lipschitz gradient as before and also strong convexity:

Theorem

Gradient descent with fixed step size $t \le 2/(m+L)$ or with backtracking line search search satisfies

$$f(x^{(k)}) - f^* \le c^k \frac{L}{2}\|x^{(0)} - x^*\|_2^2$$

where $0 < c < 1$.

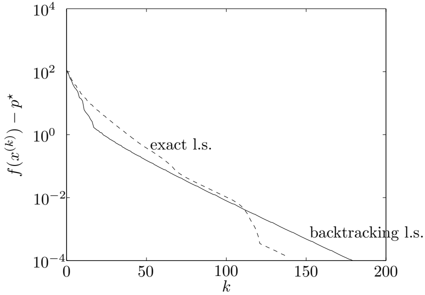
Rate under strong convexity is $O(c^k)$, exponentially fast, i.e., we find ε-suboptimal point in $O(\log(1/\varepsilon))$ iterations.

Chứng minh.

Slide 26-27 in http://www.seas.ucla.edu/~vandenbe/236C/lectures/gradient.pdf □

Convergence rate

Called linear convergence, because looks linear on a semi-log plot.



(From B & V page 487)

Important note: contraction factor c in rate depends adversely on condition number L/m: higher condition number $\Rightarrow$ slower rate.

Affects not only our upper bound... very apparent in practice too.

A look at the conditions

A look at the conditions for a simple problem, $f(\beta) = \frac{1}{2} \|y - X\beta\|_2^2$.

Lipschitz continuity of ∇f:

- This mean $\nabla^2 f(x) \preceq LI$.
- As $\nabla^2 f(\beta) = X^T X$, we have $L = \sigma_{\max}(X^T X)$.

Strong convexity of f:

- This mean $\nabla^2 f(x) \succeq mI$.
- As $\nabla^2 f(\beta) = X^T X$, we have $m = \sigma_{\min}(X^T X)$.
- If X is wide (i.e., X is $n \times p$ with $p > n$), then $\sigma_{\min}(X^T X) = 0$, and f can't be strongly convex.
- Even if $\sigma_{\min}(X^T X) > 0$, can have a very large condition number $L/m = \sigma_{\max}(X^T X)/\sigma_{\min}(X^T X)$.

Practicalities

Stopping rule: stop when $\|\nabla f(x)\|_2$ is small

- Recall $\nabla f(x^*) = 0$ at solution x^*
- If f is strongly convex with parameter m, then

$$\|\nabla f(x)\|_2 \le \sqrt{2m\varepsilon} \Longrightarrow f(x) - f^* \le \varepsilon.$$

Practicalities

Stopping rule: stop when $\|\nabla f(x)\|_2$ is small

- Recall $\nabla f(x^*) = 0$ at solution x^*
- If f is strongly convex with parameter m, then

$$\|\nabla f(x)\|_2 \leq \sqrt{2m\varepsilon} \Longrightarrow f(x) - f^* \leq \varepsilon.$$

Pros and cons of gradient descent:

- Pro: simple idea, and each iteration is cheap (usually)
- Pro: fast for well-conditioned, strongly convex problems
- Con: can often be slow, because many interesting problems aren't strongly convex or well-conditioned
- Con: can't handle nondifferentiable functions.

Can we do better?

Gradient descent has $O(1/\varepsilon)$ convergence rate over problem class of convex, differentiable functions with Lipschitz gradients.

First-order method: iterative method, which updates $x^{(k)}$ in

$$x^{(0)} + \mathrm{span}\{\nabla f(x^{(0)}), \nabla f(x^{(1)}), \ldots, \nabla f(x^{(k-1)})\}.$$

Can we do better?

Gradient descent has $O(1/\varepsilon)$ convergence rate over problem class of convex, differentiable functions with Lipschitz gradients.

First-order method: iterative method, which updates $x^{(k)}$ in

$$x^{(0)} + \text{span}\{\nabla f(x^{(0)}), \nabla f(x^{(1)}), \ldots, \nabla f(x^{(k-1)})\}.$$

Theorem (Nesterov)

For any $k \leq (n-1)/2$ and any starting point $x^{(0)}$, there is a function f in the problem class such that any first-order method satisfies

$$f(x^{(k)}) - f^* \geq \frac{3L \left\| x^{(0)} - x^* \right\|_2^2}{32(k+1)^2}.$$

Can attain rate $O(1/k^2)$, or $O(1/\sqrt{\varepsilon})$? Answer: yes (we'll see)!

What about nonconvex functions?

Assume f is differentiable with Lipschitz gradient as before, but now nonconvex. Asking for optimality is too much. So we'll settle for x such that $\|\nabla f(x)\|_2 \leq \varepsilon$, called ε-stationarity.

[1] Carmon et al. (2017), "Lower bounds for finding stationary points I"

What about nonconvex functions?

Assume f is differentiable with Lipschitz gradient as before, but now nonconvex. Asking for optimality is too much. So we'll settle for x such that $\|\nabla f(x)\|_2 \leq \varepsilon$, called ε-stationarity.

Theorem

Gradient descent with fixed step size $t \leq 1/L$ satisfies

$$\min_{i=0,\ldots,k} \|\nabla f(x^{(i)})\|_2 \leq \sqrt{\frac{2(f(x^0) - f^*)}{t(k+1)}}.$$

Thus gradient descent has rate $O(1/\sqrt{k})$, or $O(1/\varepsilon^2)$, even in the nonconvex case for finding stationary points.

This rate cannot be improved (over class of differentiable functions with Lipschitz gradients) by any deterministic algorithm[1].

[1] Carmon et al. (2017), "Lower bounds for finding stationary points I"

Proof

Key steps:

- ∇f Lipschitz with constant L means

$$f(y) \le f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|_2^2, \quad \forall x, y.$$

- Plugging in $y = x^+ = x - t\nabla f(x)$,

$$f(x^+) \le f(x) - \left(1 - \frac{Lt}{2}\right) t \|\nabla f(x)\|_2^2.$$

- Taking $0 < t \le 1/L$, and rearranging,

$$\|\nabla f(x)\|_2^2 \le \frac{2}{t}(f(x) - f(x^+)).$$

- Summing over iterations

$$\sum_{i=0}^{k} \left\|\nabla f(x^{(i)})\right\|_2^2 \le \frac{2}{t}(f(x^{(0)}) - f(x^{(k+1)})) \le \frac{2}{t}(f(x^{(0)}) - f^*).$$

- Lower bound sum by $(k+1)\min_{i=0,1,\ldots} \left\|\nabla f(x^{(i)})\right\|_2^2$, conclude.

References and further reading

- S. Boyd and L. Vandenberghe (2004), *Convex optimization*, Chapter 9
- T. Hastie, R. Tibshirani and J. Friedman (2009), *The elements of statistical learning*, Chapters 10 and 16
- Y. Nesterov (1998), *Introductory lectures on convex optimization: a basic course*, Chapter 1
- L. Vandenberghe, *Lecture notes for EE 236C*, UCLA, Spring 2011-2012