Topics in Advanced Optimization

Gradient Descent

Lecturer: Prof. CAI Hongmin

School of Computer Science & Engineering
South China University of Technology

Oct. 7, 2020

Last time : canonical convex programs

- Linear program (LP) : takes the form

$$\begin{aligned} \min_x \quad & c^T x \\ \textit{subject to} \quad & Gx \le h \\ & Ax = b \end{aligned}$$

- Quadratic program (QP) : like an LP, but with a quadratic criterion ;
- Semidefinite program (SDP) : like an LP, but with matrices ;
- Conic program : the most general form of all.

Gradient descent

Consider unconstrained, smooth convex optimization

$$\min_{x} \quad f(x)$$

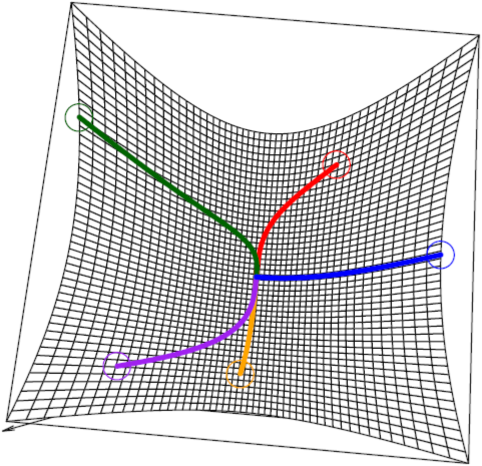
i.e., f is convex and differentiable with $dom(f) = \mathbb{R}^n$. Denote the optimal criterion value by $f^* = \min_x \ f(x)$, and a solution by x^*

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

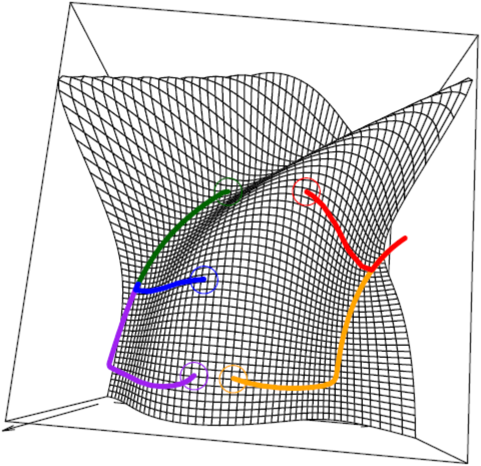
$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f\left(x^{(k-1)}\right), \ k = 1, 2, 3, \cdots$$

Stop at some point.

Gradient descent



Gradient descent



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 $\nabla^2 f(x)$ by $\frac{1}{t} I$

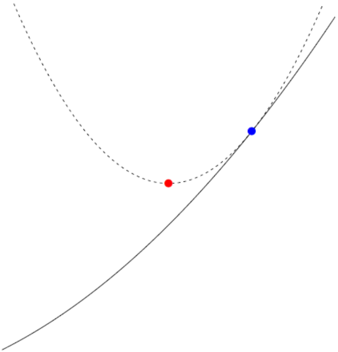
$$f(x) + \nabla f(x)^T (y - x) \qquad \leftarrow \text{linear approximation to } f$$

$$\frac{1}{2t} \|y - x\|_2^2 \qquad \leftarrow \text{proximity term to } x, \text{with weight } \frac{1}{2t}$$

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 :

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

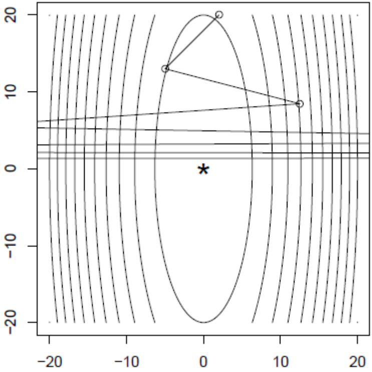
Outline

Today :

- How to choose step sizes
- Convergence analysis
- Forward stagewise regression
- Gradient boosting

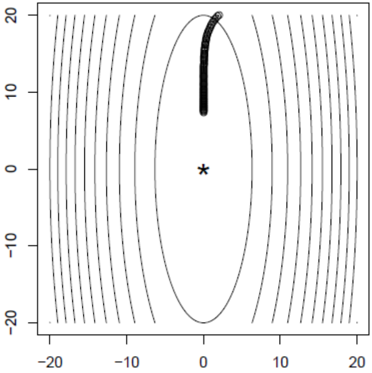
Fixed step size

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



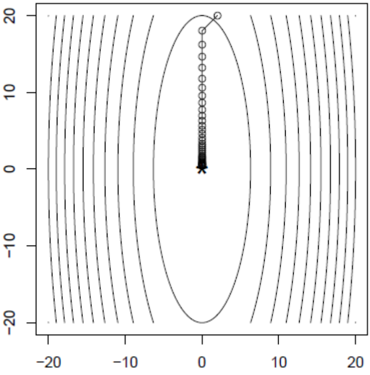
Fixed step size

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



Fixed step size

Same example, gradient descent after 40 appropriately sized steps :



Clearly there's a tradeoff.
Convergence analysis later will give us a better idea.

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 $t = 1$, and while

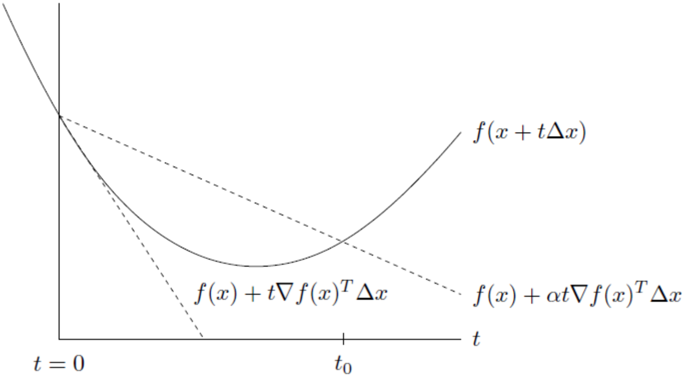
$$f(x - t\nabla f(x)) > f(x) - \alpha t \left\| \nabla f(x) \right\|_2^2$$

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 $\alpha = 1/2$)

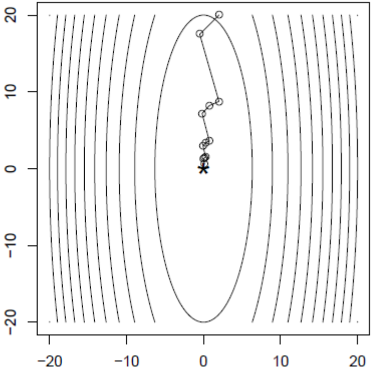
Backtracking interpretation



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

Backtracking interpretation

Backtracking picks up roughly the right step size (12 outer steps, 40 steps total) :



Here $\alpha = \beta = 0,5$

Exact line search

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

$$t = \arg\min_{s \geq 0} f\left(x - s\nabla f(x)\right)$$

Usually not possible to do this minimization exactly.
Approximations to exact line search are often not much more efficient than backtracking, and it's usually not worth it.

Convergence analysis

Assume that f convex and differentiable, with $dom(x) = \mathbb{R}^n$, and additionally

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

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

Theorem

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

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

We say gradient descent has convergence rate $0(\frac{1}{k})$

i.e., to get $f(x^{(k)}) - f^* \leq \epsilon$, we need $0(\frac{1}{\epsilon})$

Démonstration.

The function ∇f satisfies Lipschitz with constant L implies that :
$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|_2^2$ for all x, y
By plugging in $y = x^+ = x - t\nabla f(x)$,

$$f(x^+) \leq f(x) - (1 - \frac{Lt}{2})t\|\nabla f(x)\|_2^2$$

If $t \leq \frac{1}{L}$, we have

$$f(x^+) \leq (f(x) - \frac{t}{2})\|\nabla f(x)\|_2^2 \tag{1}$$

By the convexity of f, we have

$$f(x^*) \geq f(x) + \nabla f(x)^T(x^* - x) \tag{2}$$

$$\Rightarrow \quad f(x) \leq f(x^*) + \nabla f(x)^T(x^* - x) \tag{3}$$

□

Proof

By combining Eq. (1) and (2) together, we have

$$f(x^+) \le f(x^*) + \nabla f(x)^T(x^* - x) - \frac{t}{2}\|\nabla f(x)\|_2^2 \quad (4)$$

$$\Rightarrow \quad f(x^+) - f(x^*) \le \nabla f(x)^T(x^* - x) - \frac{t}{2}\|\nabla f(x)\|_2^2 \quad (5)$$

Note that

$$\begin{aligned} & \frac{1}{2t}(\|x - x^*\|_2^2 - \|x - t\nabla f(x) - x^*\|_2^2) \quad (6) \\ = \quad & \frac{1}{2t}(\|x - x^*\|_2^2 + \|x - x^*\|_2^2 - t^2\|\nabla f(x)\|_2^2 - 2t\nabla f(x)^T(x - x^*)) \\ = \quad & -\frac{t}{2}\|\nabla f(x)\|_2^2 + \nabla f(x)^T(x - x^*) \end{aligned}$$

By substituting Eq. (6) into (4), one has

$$f(x^+) - f(x^*) \le \frac{1}{2t}(\|x - x^*\|_2^2 - \|x - t\nabla f(x) - x^*\|_2^2) \quad (7)$$

$$\le \quad \frac{1}{2t}(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2) \quad (8)$$

Proof

By summing over iterations, we have

$$\begin{aligned}\sum_{i=1}^{k}(f(x^{(i)})-f^*) &\leq \frac{1}{2t}(\|x^{(0)}-x^*\|_2^2-\|x^{(k)}-x^*\|_2^2) \\ &\leq \frac{1}{2t}\|x^{(0)}-x^*\|_2^2\end{aligned}$$

Since $f(x^{(k)})$ is nonincreasing, the inequality implies that

$$f(x^{(i)})-f^* \leq \frac{1}{k}\sum_{i=1}^{k}(f(x^{(i)})-f^*) \leq \frac{1}{2tk}\|x^{(0)}-x^*\|_2^2 \tag{9}$$

We are concluding the proof.

Convergence analysis for backtracking

Same assumptions, f is convex and differentiable, $dom(f) = \mathbb{R}^n$, and ∇f is Lipschitz continuous with constant $L > 0$

Same rate for a step size chosen by backtracking search

Theorem

Gradient descent with backtracking line search $t \leq 1/L$ *satisfies*

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

where $t_{min} = \min\{1, \frac{\beta}{L}\}$.

If β is not too small, then we don't lose much compared to fixed step size ($\frac{\beta}{L}$ vs $\frac{1}{L}$)

Convergence analysis under strong convexity

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

$$\nabla^2 f(x) \geq mI \quad \text{for any } x$$

Sharper lower bound than that from usual convexity :

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|_2^2 \quad \text{all } x, y$$

Under Lipschitz assumption as before, and also strong convexity :

Theorem

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

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

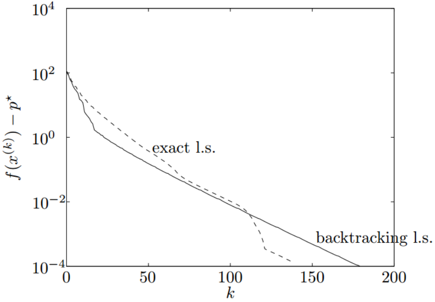
where $0 < c < 1$.

Convergence analysis under strong convexity

I.e., rate with strong convexity is $O(c^k)$, exponentially fast !

I.e., to get $f\left(x^{(k)}\right) - f^* \leq \epsilon$, need $O(log(\frac{1}{\epsilon}))$ iterations

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



Constant c depends adversely on condition number $\frac{L}{m}$ (higher condition number $\Rightarrow$ slower rate)

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 means $\nabla^2 f(x) \preceq LI$
- As $\nabla^2 f(\beta) = X^T X$, we have $L = \sigma^2_{\text{max}}(X)$

Strong convexity of f :

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

A look at the conditions

A function f having Lipschitz gradient and being strongly convex satisfies :

$$mI \preceq \nabla^2 f(x) \preceq LI \quad \text{for all } x \in \mathbb{R}^n,$$

for constants $L > m > 0$
Think of f begin sandwiched between two quadratics

May seem like a strong condition to hold globally (for all $x \in \mathbb{R}^n$). But a careful look a the proofs shows that we only need Lipschitz gradients/strong convexity over the sublevel set

$$S = \left\{ x : f(x) \le f(x^{(0)}) \right\}$$

This is less restrictive

Practicality

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

$$\|\nabla f(x)\|_2 \leq \sqrt{2m\epsilon} \Rightarrow f(x) - f^* \leq \epsilon$$

Pros and cons of gradient descent :

Pro :

- simple idea, and each iteration is cheap ;
- very fast for well-conditioned, strongly convex problems.

Con :

- often slow, because interesting problems aren't strongly convex or well-conditioned
- can't handle nondifferentiable functions

Forward stagewide regression

Let's stick with $f(\beta) = \frac{1}{2}\|y - X\beta\|_2^2$, linear regression setting X is $n \times p$, its columns

$X_1, ...X_p$ are predictor variables

Forward stage-wise regression : start with $\beta^{(0)} = 0$, repeat :

- Find variable i s.t. $|X_i^T r|$ is largest, where $r = y - X\beta^{(k-1)}$ (largest absolute correlation with residual)
- Update $\beta_i^{(k)} = \beta_i^{(k-1)} + \gamma \cdot sign(X_i^T r)$

Here $\gamma > 0$ is small and fixed, called learning rate. This looks kind of like gradient

descent.

Steepest descent

Close cousin to gradient descent,just change the choice of norm. Let p, q be complementary (dual) : $1/p + 1/q = 1$

Steepest descent updates are $x^+ = x + t \cdot \Delta x$, where

$$\Delta x = \|\nabla f(x)\|_q \cdot u$$

$$u = \arg\min_{\|v\|_p \le 1} \nabla f(x)^T v$$

- If $p = 2$, then $\Delta x = -\nabla f(x)$, gradient descent
- If $p = 1$, then $\Delta x = -\partial f(x)/\partial x_i \cdot e_i$, where

$$\left|\frac{\partial f}{\partial x_i}(x)\right| = \max_{j=1,\dots n} \left|\frac{\partial f}{\partial x_j}(x)\right| = \|\nabla f(x)\|_\infty$$

Normalized steepest descent just takes $\Delta x = u$ (unit q-norm)

An interesting equivalence

Normalized steepest descent with respect to ℓ_1 norm :updates are

$$x_i^+ = x_i - t \cdot sign\left(\frac{\partial f}{\partial x_i}(x)\right)$$

where i is the largest component of $\nabla f(x)$ in absolute value

Compare forward stage-wise : updates are

$$\beta_i^+ = \beta_i + \gamma \cdot sign(X_i^T r), r = y - X\beta$$

But here $f(\beta) = \frac{1}{2}\|y - X\beta\|_2^2$, so $\nabla f(\beta) = -X^T(y - X\beta)$ and $\partial f(\beta)/\partial \beta_i = -X_i^T(y - X\beta)$

Hence forward stagewise regression is normalized steepest descent under ℓ_1 norm (with fixed step size $t = \gamma$)

Early stopping and sparse approximation

If we run forward stagewise to completion, then we will minimize $f(\beta) = \|y - X\beta\|_2^2$, i.e.,we will produce a least squares solution

What happens if we stop early ?

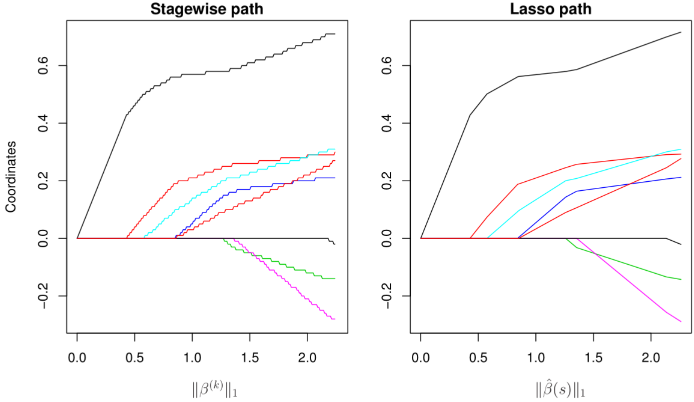
- May seem strange from an optimization perspective (we are "under-optimizing")...
- Interesting from a statistical perspective, because stopping early gives us a sparse approximation to the least squares solution

Well-known sparse regression estimator, the lasso :

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 \quad \text{subject to } \|\beta\|_1 \leq s$$

How do lasso solutions and forward stagewise estimates compare ?

Early stopping and sparse approximation



For some problems(some y, X), they are exactly as the learning rate $\gamma \to 0$!

Can we do better ?

Recall $O(1/\epsilon)$ rate for gradient descent over problem class of convex, differentiable functions with Lipschitz continuous gradients

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

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

Theorem

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

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

Can attain rate $O(1/k^2)$, or $O(1/\sqrt{\epsilon})$? Answer : **yes** (and more) !

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 2
- R. J. Tibshirani (2014), “A general framework for fast stagewise algorithms"
- L. Vandenberghe, Lecture notes for EE 236C, UCLA, Spring 2011-2012