Today: Convergence rates

Assignment 5 out this weekend
Marked Assignment 3 due Mon Oct 20
No tutorial this week

Last week: Artificial Neural Networks (no longer convex)

instead of $\sum_{i=1}^{N} f_i(w^T x_i)$, we have $\sum_{i=1}^{N} f_i(w_1^T g(w_2^T x_i))$ (like a set of features)

linear

or $\sum_{i=1}^{N} f_i(w_1^T g(w_2^T h(w^T x_i)))$ (add more layers)

Subgradient Questions

Where do subgradients exist?
- everywhere for convex functions!

Why focus on gradient methods? $O(nd)$. Stochastic: $O(d)$
- Newton, linprog, quadprog, etc. cost: $O(d^2)$ or $O(n^2)$ or larger ✔
  (great for small problems, but modern problems can't apply them)

Convergence Rate of Gradient Descent

Gradient Method: $x^{t+1} = x^t - \alpha_t \nabla f(x^t)$

How fast does this converge? ($O(nd)$, but how many iterations?) ✔
- to answer this, you need assumptions
  (function or gradient can't change too quickly)

We'll assume $\mu I \preccurlyeq \nabla^2 f(x) \preccurlyeq LI$ (eigenvalues positive; Hessian; eigenvalues bounded by)

$\forall x \forall v \quad v^T \nabla^2 f(x) v \leq v^T (LI) v = L(v^T I v) = L v^T v = L\|v\|^2$

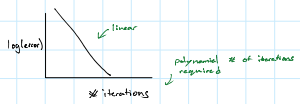
($\mu > 0$) "strongly convex" ① 0'th order $f(x) - \frac{\mu}{2}\|x\|^2$ is convex (regularizer)

($L < \infty$) "strongly smooth" ② 1'st order $\nabla f(x)$ is "L-Lipschitz" continuous: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$

Proposition: With $\alpha_t = 1/L$, the gradient method satisfies

$$f(x^t) - f(x_*) \leq \left(1 - \frac{\mu}{L}\right)^t \left[f(x^0) - f(x_*)\right]$$

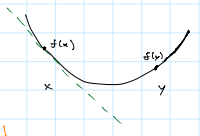
(optimal)

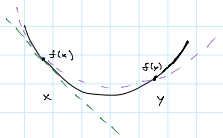


Proof: One version of the taylor expansion

$\forall x, y, \exists z$ st. $f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y-x)^T \nabla^2 f(z)(y-x)$

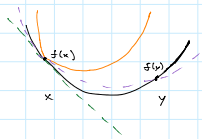
(quadratic function of y; linearization; Hessian in quadratic form)





upper bound

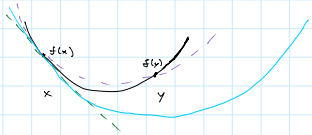
$f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|^2$



(always above the function)

lower bound

$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2}\|y-x\|^2$



(always below the function)

rearrange

Set $x = x^t$, $y = x^{t+1}$

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

use $x^{t+1} - x^t = -\alpha_t \nabla f(x^t)$, $\alpha_t = 1/L$

$f(x^{t+1}) \leq f(x^t) + \nabla f(x^t)^T \left(-\left(\frac{1}{L}\right) \nabla f(x^t)\right) + \frac{L}{2}\|\nabla f(x^t)\|^2$

$f(x^{t+1}) \leq f(x^t) - \frac{1}{2L}\|\nabla f(x^t)\|^2$ ("guaranteed progress")

(positive)

rearrange

Minimize in terms of y

$\left(y = x - \frac{1}{\mu}\nabla f(x)\right)$

$f(x_*) \geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2$

rearrange (upper bound)

$-\|\nabla f(x^t)\|^2 \leq -2\mu\left[f(x^t) - f(x_*)\right]$

Add $-f(x_*)$ to both sides of upper bound rearrange

$$f(x^{t+1}) - f(x_*) \le f(x^t) - f(x_*) - \frac{1}{2L}\|\nabla f(x^t)\|^2$$

↓ using lower bound rearrange

$$f(x^{t+1}) - f(x_*) \le f(x^t) - f(x_*) - \frac{2\mu}{2L}\left[f(x^t) - f(x_*)\right]$$

$$\le \left(1 - \frac{\mu}{L}\right)\left[f(x^t) - f(x_*)\right]$$

$$f(x^t) - f(x_*) \le \left(1 - \frac{\mu}{L}\right)\left[f(x^{t-1}) - f(x_*)\right]$$

$$\le \left(1 - \frac{\mu}{L}\right)\left(1 - \frac{\mu}{L}\right)\left[f(x^{t-1}) - f(x_*)\right]$$

$\vdots$ (keep doing this until we get x_0)

$$f(x^t) - f(x_*) \le \left(1 - \frac{\mu}{L}\right)^t \left[f(x_0) - f(x_*)\right]$$

$$\left(1 - \frac{\mu}{L}\right)^t \le \exp\left(-t\,\frac{\mu}{L}\right)$$

$f(x^t) - f(x_*) \le \epsilon \qquad t = O\left(\log\left(\frac{1}{\epsilon}\right)\right)$

total cost: $O\left(Nd \log\left(\frac{1}{\epsilon}\right)\right)$

Deterministic Gradient Complexity 200

Strongly Smooth

"linear" → Strongly Convex, error: $O\left(\left(1 - \frac{\mu}{L}\right)^t\right)$

"sublinear" → Convex $O\left(\frac{1}{t}\right)$ — longer you run, less progress you make

Strongly Convex → extrapolation: $O\left(\left(1 - \sqrt{\frac{\mu}{L}}\right)^t\right)$

Convex → extrapolation: heavy ball, momentum, Nesterov's accelerated gradient $O\left(\frac{1}{t^2}\right)$

Strongly Convex → Add: Lipschitz Hessian, Newtons Method → "super linear" $O\left(\prod_{i=1}^{t} \rho_i\right)$, $\rho_i \to 0$

Approximate Newton:
- Barzilai-Borwein
- Quasi-Newton (L-BfGS) (limited memory)
- Hessian-free Newton

$$x^{t+1} = x^t - \alpha_t \nabla f(x_t) + \beta_t (x_t - x_{t-1})$$

"approaching from one side" → extrapolation ⇒ some momentum, will "overshoot" the goal, but then come back

gets stuck → extrapolation ⇒ achieves the optimal

"saddle point"

(*) line search
- important in practice
- no effect on theory

Subgradient and/or Stochastic Gradient

x_* is not a fixed-point

$$x_* \neq x_* - \alpha g(x_*)$$

(g: subgradient)

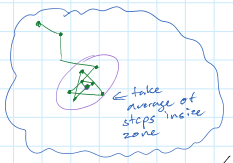
To get around this, need $\alpha_t \to 0$

Robbins-Monro $\sum_{t=1}^{\infty} \alpha_t = \infty$

$\sum_{t=1}^{\infty} (\alpha_t)^2 < \infty$

Eg., $\alpha_t = \frac{1}{\alpha t}$ gives $f(x^t) - f(x_*) = O\left(\frac{1}{t}\right)$

Subgradient / Stochastic Grad. Complexity 200



Lipschitz f (can't change too quickly)

Strongly Convex
$O\left(\frac{\log(t)}{t}\right)$

Convex
$O\left(\frac{1}{\sqrt{t}}\right)$

finite, strongly smooth → SAG
$O\left(\frac{1}{t}\right)$

Tail Averaging
$O\left(\frac{1}{t}\right)$

Strongly Smooth
$O\left(\frac{1}{t}\right)$

Smooth approximation
$O\left(\frac{1}{\sqrt{t}}\right)$

(Deterministic) extrapolation (exact gradient)
$O\left(\frac{1}{t}\right)$

Constant Step-Size
$O\left((1-2\mu\alpha)^t\right) + O(\alpha)$
↑ additional error based off step size

finite SAG
$O(\rho^t)$