ML and Opt Lecture 12

- Strong convexity, convex condition numbers
- Polyak-Lojasiewicz (PL) inequality
- Gradient Descent converges linearly for well-conditioned convex problems
- Gradient Descent convergence rates for different classes of convex functions
- Subgradient Descent for nonsmooth cvx optim

α-Strongly convex

We say a cvx function f is α-strongly convex if

$$f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle + \frac{\alpha}{2}\|x-y\|_2^2$$

$$D = \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix}$$

$$\|Dx\|_2^2 = x^T D^T D x$$
$$= \lambda_1 x_1^2 + \lambda_2 x_2^2$$

assume $\lambda_1 \ll \lambda_2$, then
$\|Dx\|_2^2 \geq \lambda_1 (x_1^2 + x_2^2)$

Note

$$\|Dy\|_2^2 = \|Dx\|_2^2 + \langle 2D^T Dx, y-x \rangle + \frac{1}{2}(y-x)^T (2D^T D)(y-x)$$
$$= \|Dx\|_2^2 + 2\langle D^2 x, y-x \rangle + \|D(y-x)\|_2^2$$
$$\geq \|Dx\|_2^2 + 2\langle D^2 x, y-x \rangle + \lambda_1 \|y-x\|_2^2$$

$\|Dx\|_2^2$ is $2\lambda_1$-strongly convex

Ex function that is β-smooth and α-strongly cvx

$$f(x) = \frac{1}{2}\|Ax-b\|_2^2 \quad \text{where} \quad \alpha \cdot I \preceq A^TA \preceq \beta \cdot I$$

lower bound on curvature of f

upper bound on curvature of f

Pf

$$f(y) = f(x) + \langle \nabla f(x), y-x \rangle + \frac{1}{2}(y-x)^TA^TA(y-x)$$

Because $A^TA \preceq \beta \cdot I \Rightarrow (y-x)^TA^TA(y-x) \leq \beta\|y-x\|_2^2$

and $A^TA \succeq \alpha \cdot I \Rightarrow (y-x)^TA^TA(y-x) \geq \alpha\|y-x\|_2^2$

we have

$$f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle + \frac{1}{2}\alpha\|y-x\|_2^2$$ α-strong convexity

$$f(y) - f(x) - \langle \nabla f(x), y-x \rangle \leq \frac{\beta}{2}\|y-x\|_2^2$$ β-smoothness

Facts Identify α-strongly cvx & β-smooth cvx fcns

If f is twice continuously differentiable in $\text{dom}(f)$, then

1) f is α-strongly cvx $\iff \forall x \in \text{dom}(f)$:

$$\nabla^2 f(x) \succcurlyeq \alpha \cdot I$$

2) f is β-smooth $\iff \forall x \in \text{dom}(f)$:

$$\nabla^2 f(x) \preccurlyeq \beta \cdot I$$

Outline of proof

$\Leftarrow$ (1)

Assume $\forall x \in \text{dom}(f)$: $\nabla^2 f(x) \succcurlyeq \alpha \cdot I$. Consider a pt $z \in \text{dom}(f)$

$$f(y) = f(z) + \langle \nabla f(z), y - x \rangle + \frac{1}{2}(y-z)^T \nabla^2 f(a)(y-z)$$

where $a \in [z, y]$ by T.S. w/ Lagrange remainder

Then because $a \in \text{dom}(f)$, $\nabla^2 f(a) \succcurlyeq \alpha \cdot I \Rightarrow$

$$f(y) \geq f(z) + \langle \nabla f(z), y - z \rangle + \frac{\alpha}{2} \|y - z\|_z^2$$

This is true for all y (although a depends on y)

so f is α-strongly cvx on $\text{dom}(f)$

Gradient Descent converges very fast for β-smooth, α-strongly cvx functions

$$f(x_T) - f(x_*) \leq \left(1 - \frac{1}{\kappa_f}\right)^T \left(f(x_0) - f(x^*)\right)$$

where $\kappa_f = \frac{\beta}{\alpha} \geq 1$, if we use GD w/ stepsize $\frac{1}{\beta}$

Fact Strongly cvx functions satisfy the PL-inequality

$$\frac{1}{2\alpha}\|\nabla f(x)\|_2^2 \geq f(x) - f(x^*)$$

Pf

By strong cvxty we have that for all y,

$$f(y) \geq f(x) + \langle \nabla f(x), y-x\rangle + \frac{\alpha}{2}\|y-x\|_2^2$$

Notice the RHS is a linear term plus a quadratic that curves upward. This implies that the RHS has a finite minimum. Let's minimize the RHS as a function of y by setting grad to 0:

$$\nabla f(x) + \alpha(y^* - x) = 0$$

$$\Leftrightarrow \quad y^* = x - \frac{1}{\alpha}\nabla f(x)$$

the corresponding minimum value is

$$f(x) + \langle \nabla f(x), \left(-\frac{1}{\alpha}\nabla f(x)\right)\rangle + \frac{\alpha}{2}\left\|-\frac{1}{\alpha}\nabla f(x)\right\|_2^2$$

$$= f(x) - \frac{1}{\alpha}\|\nabla f(x)\|_2^2 + \frac{1}{2\alpha}\|\nabla f(x)\|_2^2$$

$$= f(x) - \frac{1}{2\alpha}\|\nabla f(x)\|_2^2 .$$

So for <u>any</u> $y \in \mathrm{dom}(f)$:

$$f(y) \geq f(x) - \frac{1}{2\alpha}\|\nabla f(x)\|_2^2$$

In particular, take $y = x^*$:

$$\frac{1}{2\alpha}\|\nabla f(x)\|_2^2 \geq f(x) - f(x^*).$$

Prf Linear convergence of GD for well-conditioned cvx optim probs.

Consider the descent lemma's guarantee that when the stepsize is $\frac{1}{\beta}$:

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2$$

so

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2\beta}\|\nabla f(x_t)\|_2^2$$

and by the PL-inequality

$$\|\nabla f(x_t)\|_2^2 \geq 2\alpha\left(f(x_t) - f(x_*)\right)$$

so

$$f(x_{t+1}) - f(x_t) \leq -\frac{2\alpha}{2\beta}\left(f(x_t) - f(x_*)\right)$$

Now add $f(x_t) - f(x_*)$ to both sides to get

$$f(x_{t+1}) - f(x_*) \leq \left(1 - \frac{\alpha}{\beta}\right)\left[f(x_t) - f(x_*)\right]$$

By induction on t,

$$f(x_T) - f(x_*) \leq \left(1 - \frac{1}{\kappa_f}\right)^T \left[f(x_0) - f(x_*)\right]$$

Convergence rates for different classes of cvx functions using GD (Revisiting Polyak's step size. Hazan & Kakade. 2019)

properties	cvx	β-smooth	α-strongly cvx	$\kappa = \frac{\beta}{\alpha}$ well-conditioned
convergence rate	$\frac{1}{\sqrt{T}}$	$\frac{\beta}{T}$	$\frac{1}{\alpha T}$	$e^{-T/\kappa}$
stepsize	$\frac{1}{\sqrt{T}}$	$\frac{1}{\beta}$	$\frac{1}{\alpha T}$	$\frac{1}{\beta}$

What about nonsmooth convex optimization?

Subgradient Descent Alg

f – cvx function

$x^* \in \operatorname{argmin}_x f(x)$

Inputs

∂f – oracle for subgradient

T – # iterates

x_0 – starting pt

$\alpha_1, \dots, \alpha_T$ – stepsizes

Alg

for $t = 0, \dots, T-1$:

$g_t \leftarrow$ a vector in $\partial f(x_t)$

$x_{t+1} \leftarrow x_t - \alpha_{t+1} g_t$

Options for Return

1) Monitor the function values

$x_{best} = x_0$

for $t = 0, \dots, T-1$:

if $f(x_{t+1}) < f(x_{best})$:

$x_{best} \leftarrow x_{t+1}$

return x_{best}

2) Polyak averaging

return

$$\hat{x} = \frac{1}{T} \sum_{t=1}^{T} x_t$$

(Polyak average)

Guarantee

$$f(\hat{x}) - f(x^*) \leq O\left(\frac{1}{\sqrt{T}}\right)$$

How to handle constraints?

(In both GD and subgradient descent)

$$x^* \in \underset{x \in C}{\operatorname{argmin}} \; f(x)$$

Update step becomes

$$x_{t+1} = P_C\left(x_t - \alpha_{t+1} g_t\right)$$