THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

SDS 385: Stat Models for Big Data

Lecture 4: GD with momentum.

Purnamrita Sarkar
Department of Statistics and Data Science
The University of Texas at Austin

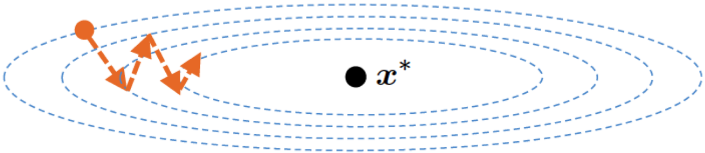
https://psarkar.github.io/teaching

Polyak's heavy ball method

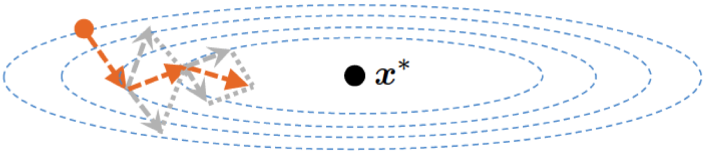
Figure 1: B. Polyak

- $$\beta_{t+1} = \beta_t - \alpha \nabla f(\beta_t) + \underbrace{\theta(\beta_t - \beta_{t-1})}_{\text{momentum term}}$$

Momentum



gradient descent



heavy-ball method

Recall GD?

- For a L smooth and μ convex optimization problem, i.e. $\mu I \preceq H \preceq LI$,

$$\|\beta_t - \beta^*\| \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^t \|\beta_0 - \beta^*\|$$

  where $\kappa = L/\mu$ i.e. the condition number of the Hessian.

- For the same problem, using Polyak's method we can show that,

$$\left\| \begin{bmatrix} \beta_{t+1} - \beta^* \\ \beta_t - \beta^* \end{bmatrix} \right\| \leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^t \left\| \begin{bmatrix} \beta_1 - \beta^* \\ \beta_0 - \beta^* \end{bmatrix} \right\|$$

Momentum method

- Recall we have:

$$\begin{aligned}\beta_{t+1} - \beta^* &= (1+\theta)(\beta_t - \beta^*) - \alpha\nabla f(\beta_t) - \theta(\beta_{t-1} - \beta^*) \\ &= ((1+\theta)I - \alpha\nabla^2 f(z_t))(\beta_t - \beta^*) - \theta(\beta_{t-1} - \beta^*)\end{aligned}$$

- This gives the dynamic system:

$$\begin{bmatrix}\beta_{t+1} - \beta^* \\ \beta_t - \beta^*\end{bmatrix} \leq \begin{bmatrix}(1+\theta)I - \alpha\nabla^2 f(z_t) & -\theta I \\ I & 0\end{bmatrix}\begin{bmatrix}\beta_t - \beta^* \\ \beta_{t-1} - \beta^*\end{bmatrix}$$

Momentum method

- We need to upper bound the norm of $M := \begin{bmatrix} (1+\theta)I - \alpha\nabla^2 f(z_t) & -\theta I \\ I & 0 \end{bmatrix}$
- It can be shown that:

$$\begin{aligned} \|M\| &= \left\| \begin{bmatrix} (1+\theta) - \alpha\Lambda & -\theta I \\ I & 0 \end{bmatrix} \right\| \\ &= \max_i \left\| \begin{bmatrix} (1+\theta) - \alpha\lambda_i & -\theta \\ 1 & 0 \end{bmatrix} \right\| \end{aligned}$$

- Eigenvalues of the 2×2 matrix can be written as a solution of the following quadratic:

$$\sigma^2 - \sigma((1+\theta) - \alpha\lambda_i) + \theta = 0$$

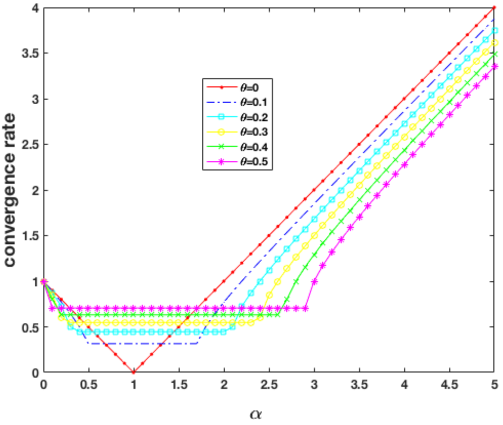
Momentum method - simple example

- Take $f(x) = \frac{h}{2}x^2$.
- Now $M := \begin{bmatrix} 1+\theta-\alpha h & -\theta \\ 1 & 0 \end{bmatrix}$
- The two eigenvalues of this matrix are:

$$\sigma_1 = \frac{1}{2}\left(1 - \alpha h + \theta + \sqrt{(1+\theta-\alpha h)^2 - 4\theta}\right)$$
$$\sigma_2 = \frac{1}{2}\left(1 - \alpha h + \theta - \sqrt{(1+\theta-\alpha h)^2 - 4\theta}\right)$$

- When $(1+\theta-\alpha h)^2 < 4\theta$, then the roots are complex conjugates, and each have the same absolute value $\sqrt{\theta}$

Momentum method - simple example



Momentum method

- If $((1+\theta) - \alpha\lambda_i)^2 \le 4\theta$, the roots are imaginary and the magnitude is $\sqrt{\theta}$
- This is satisfied if

$$\alpha \in \left[\frac{(1-\sqrt{\theta})^2}{\lambda_i}, \frac{(1+\sqrt{\theta})^2}{\lambda_i}\right]$$

- But recall that $\lambda_i \in [\mu, L]$.
- If we set $1 - \sqrt{\alpha L} = -(1 - \sqrt{\alpha\mu})$, then we have

$$\alpha = \left(\frac{2}{\sqrt{L} + \sqrt{\mu}}\right)^2 \qquad \theta = \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^2$$

- So the new contraction factor becomes $\dfrac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}$

Nesterov's Accelerated Gradient

- If we only assume that $\|\nabla^2 f(x)\| \leq L$ and not strong convexity, then in your homework you will prove that

$$f(\beta_t) - f(\beta^*) \leq c_L \frac{\|\beta_0 - \beta^*\|^2}{t}$$

- Note that this is much weaker than the linear convergence we saw before.
- Question is can we do better?

Nesterov's Accelerated Gradient

Figure 2: Y. Nesterov

- Keep track of two vectors x_t and y_t
- $x_{t+1} = y_t - \alpha_t \nabla f(y_t)$
- $y_{t+1} = x_{t+1} + \underbrace{\frac{t}{t+3}}_{\mu_{t+1}}(x_{t+1} - x_t)$

Nesterov's Accelerated Gradient

- Can be re-written as:

$$x_{t+1} = x_t + \mu(x_t - x_{t-1}) - \alpha_t \nabla f(x_t + \mu_t(x_t - x_{t-1}))$$

- Very much like the momentum method, but computes the derivative at a future step.

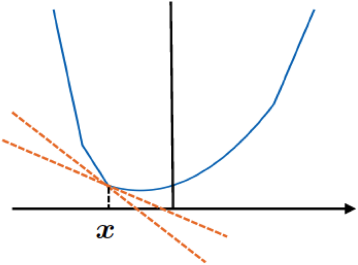
Nesterov's Accelerated Gradient

- Not a descent method.
- If f is convex and L smooth and the learning rate is $1/L$, this obtains the optimal $O(1/t^2)$ error after t steps.
- Proof is complicated, but can be simplified using intuitions from differential equations.

Subgradient methods

- So far we have assumed differentiable f.
- What if f is not differentiable?
- Instead of a gradient we will define a subgradient.

Subgradient methods

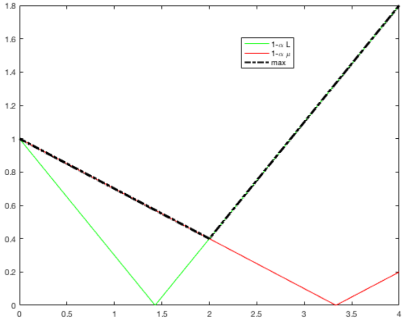


- We will say that g is a subgradient of f at point x if

$$f(z) \geq f(x) + g^T(z - x), \qquad \forall z$$

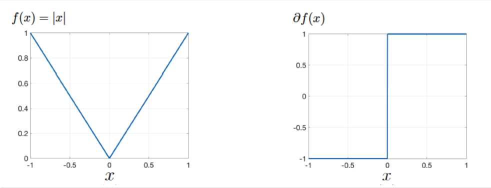
- Set of all gradients is called the sub-differential of f at point x and is denoted by $\partial f(x)$

Example



$$f(x) = \max(g(x), h(x)) \qquad \delta f(x) = \begin{cases} \{g'(x)\} & \text{If } g(x) > h(x) \\ \in [g'(x), h'(x)] & \text{If } g(x) = h(x) \\ \{h'(x)\} & \text{If } g(x) < h(x) \end{cases}$$

Example



$$f(x) = |x| \qquad \delta f(x) = \begin{cases} \{-1\} & \text{If } x < 0 \\ [-1, 1] & \text{If } x = 0 \\ \{1\} & \text{If } x > 0 \end{cases}$$

Subgradients

Behaves very much like a gradient;

- $\partial(\alpha f) = \alpha \partial f$ for $\alpha \geq 0$
- $\partial(f + g) = \partial f + \partial g$
- For convex f, if $g(x) = f(Ax + b)$, $\partial g(x) = A^T \partial f(Ax + b)$

ℓ_1 norm

$$f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} \underbrace{|x_i|}_{:= f_i(\boldsymbol{x})}$$

since

$$\partial f_i(\boldsymbol{x}) = \begin{cases} \mathrm{sgn}(x_i)\boldsymbol{e}_i, & \text{if } x_i \neq 0 \\ [-1, 1] \cdot \boldsymbol{e}_i, & \text{if } x_i = 0 \end{cases}$$

we have

$$\sum_{i: x_i \neq 0} \mathrm{sgn}(x_i)\boldsymbol{e}_i \in \partial f(\boldsymbol{x})$$

Lets talk about Lasso

$$\hat{\beta}_{LASSO} = \min_{\beta}(\boldsymbol{y} - \boldsymbol{x}\beta)^{\top}(\boldsymbol{y} - \boldsymbol{x}\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1}$$

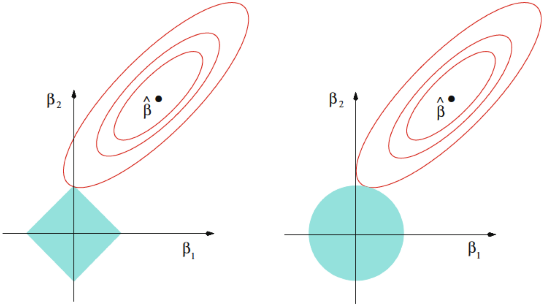


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function.*

Alternative formulation

$$
\begin{aligned}
\hat{\beta}_{ridge} &= \min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{x}\beta)^\top (\boldsymbol{y} - \boldsymbol{x}\beta) \qquad \text{Subject to } \boldsymbol{\beta}^\top \boldsymbol{\beta} \leq \tau^2 && (2) \\
\hat{\beta}_{lasso} &= \min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{x}\beta)^\top (\boldsymbol{y} - \boldsymbol{x}\beta) \qquad \text{Subject to } \|\boldsymbol{\beta}\|_1 \leq \tau && (3)
\end{aligned}
$$

Uniqueness - OLS (Thanks to Piaoping Jiang for asking this in class)

- So what happens to linear regression when $p > n$ or $rank(X) < p$?
- There are many solutions,
  - You can just add a vector lying in the null space of X to a solution to get another
  - In particular, you can always find a variable which has +ve sign on solution and -ve sign on another.
  - This makes interpreting a solution rather difficult.

Uniqueness - Lasso

- So the question is, what happens in Lasso, when X is rank deficient.
  - For a fixed λ, can one lasso solution have a positive i^{th} coefficient, and another have a negative i^{th} coefficient?
  - Must any two lasso solutions, at the same value of λ, necessarily share the same support, and differ only in their estimates of the nonzero coefficient values? Or can different lasso solutions exhibit different active sets?

Uniqueness - Lasso

- Q0. When does Lasso have non-unique solutions?
  - If the elements of X are drawn from a continuous probability distribution, then the lasso returns a unique solution with probability one over the distribution of X, regardless of the sizes of n and p.
  - So, the only time you have to worry about non-uniqueness, is when X is discrete.

Uniqueness - Lasso

- Q1. For a fixed λ, can one lasso solution have a positive i^{th} coefficient, and another have a negative i^{th} coefficient?
  - The short answer is no, and you can prove this. So, unlike OLS, lasso solutions do not suffer from sign inconsistencies.
- Q2. Can there be different supports for the same λ?
  - Unfortunately yes. But you can compute upper and lower bounds for the lasso coefficients, and deal with this.

Optimality condition

- For differentiable f

$$f(x^*) = \min_x f(x) \leftrightarrow \nabla f(x^*) = 0$$

- For convex f that may not be differentiable,

$$f(x^*) = \min_x f(x) \leftrightarrow 0 \in \delta f'(x^*)$$

- Just plug into the definition of a subgradient!

$$f(y) \geq f(x^*) + 0^T (y - x^*) = f(x^*)$$

Soft thresholding

- Consider the easier problem

$$x = \arg\min \frac{1}{2}\|y - x\|^2 + \lambda\|x\|_1$$

- Show that the soft thresholding operator $x^* = S_\lambda(y)$ is the solution to this.

$$S_\lambda(y) = \begin{cases} y_i - \lambda & \text{if } y_i > \lambda \\ 0 & \text{if } y_i \in [-\lambda, \lambda] \\ y_i + \lambda & \text{if } y_i < -\lambda \end{cases}$$

Sub-gradient method

- $\beta_{k+1} = \beta_k - \alpha_k g_k$
- Here g_k is any subgradient at the β_k
- Note that subgradient direction is not always a direction of descent
- So we do

$$f(\beta_k^{best}) = \min_{i=1,\ldots,k} f(\beta_i)$$

- We can choose it as
  - Fixed, i.e. $\alpha_k = \alpha$
  - Or diminishing such that $\sum_k \alpha_k^2 < \infty, \sum_k \alpha_k = \infty$

Convergence

- Assume that f is convex, and Lipschitz continuous with some constant $G > 0$

$$|f(x) - f(y)| \leq G\|x - y\|_2 \text{ for all } x, y$$

Theorem
For a fixed step size α,

$$\lim_{k\to\infty} f(x_{best}^{(k)}) \leq f^* + G^2\alpha/2$$

Theorem
For diminishing step size α_k with $\sum_k \alpha_k \to \infty$ and $\sum_k \alpha_k^2 < \infty$,

$$\lim_{k\to\infty} f(x_{best}^{(k)}) = f^*$$

Regularized logistic regression

- Let $\{x_i, y_i\}_{i=1}^{n}$ with $x_i \in \Re^p$ and $y_i \in \{0, 1\}$
- The logistic regression loss is:

$$f(\beta) = \sum_i \left(-y_i x_i^T \beta + \log(1 + \exp(x_i^T \beta)) \right)$$

- With lasso regularization we have:

$$\hat{\beta} = \arg\min f(\beta) + \lambda \|\beta\|_1$$

- So, use

$$\Delta_\beta = \underbrace{-\sum_i (y_i - p_i(\beta)) x_i}_{\textit{gradient}} + \underbrace{\partial \|\beta\|_1}_{\textit{subgradient}}$$

Convergence

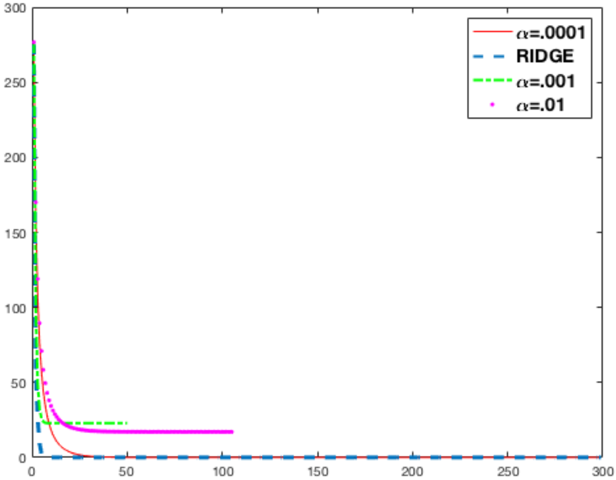


Figure 3: $f - f^*$ on X axis, iterations on Y axis. Logistic regression in two dimensions, $\lambda = 1$

Convergence

- Gradient descent takes $1/\epsilon$ time to converge, whereas subgradient descent with variable step-size takes $1/\epsilon^2$ time to converge.

Theorem
For any $k \leq n-1$ and starting point $\beta^{(0)}$, there is a function such that any non-smooth first order method satisfies:

$$f(\beta^{(k)}) - f^* \geq \frac{G\|\beta^{(0)} - \beta^*\|}{2(1+\sqrt{k+1})}$$

- So it seems like we cant really improve on sub-gradient methods.

Acknowledgment

Y. Chen's large scale Optimization class at Princeton and Hastie and Tibshirani's book, Ryan Tibshirani's class, "The Lasso Problem and Uniqueness", R. J. Tibshirani.