Acceleration

Lecturer: Pradeep Ravikumar
Co-instructor: Aarti Singh

Convex Optimization 10-725/36-725

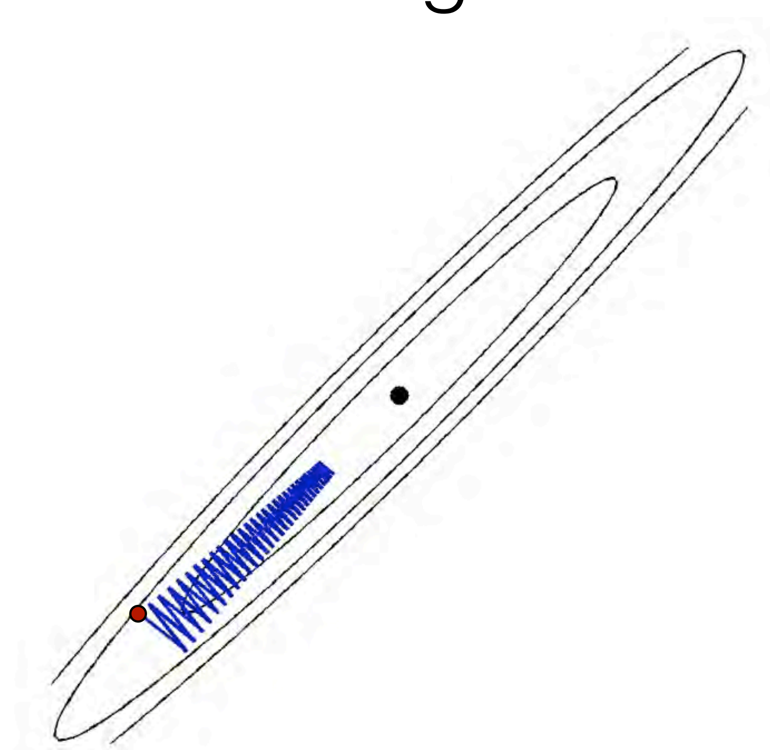
Based on slides from Recht, Tibshirani

Gradient Descent

- Recall Gradient Descent:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

- One caveat is that it relies too much on local information to decide direction, and hence might be too slow



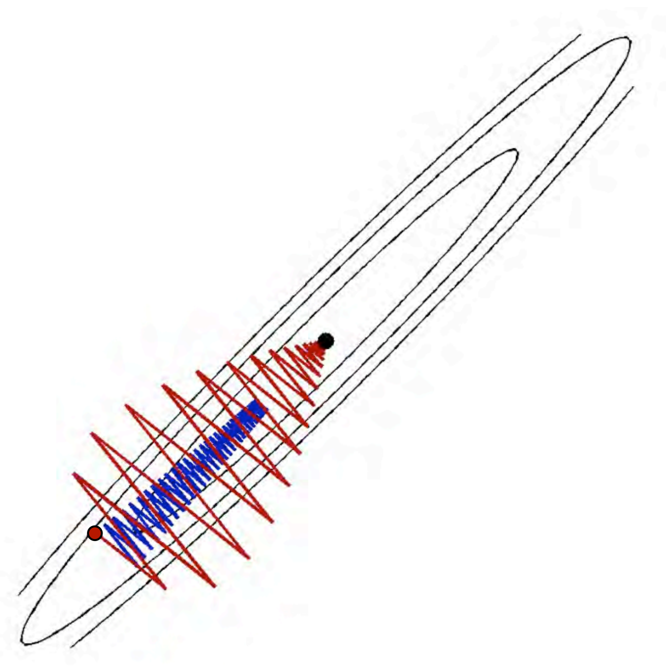
Gradient Descent

- Recall Gradient Descent:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

- One caveat is that it relies too much on local information to decide direction, and hence might be too slow

- With an additional "momentum" term, it might be less slow



Heavy Ball Method

- Gradient Descent + Momentum:

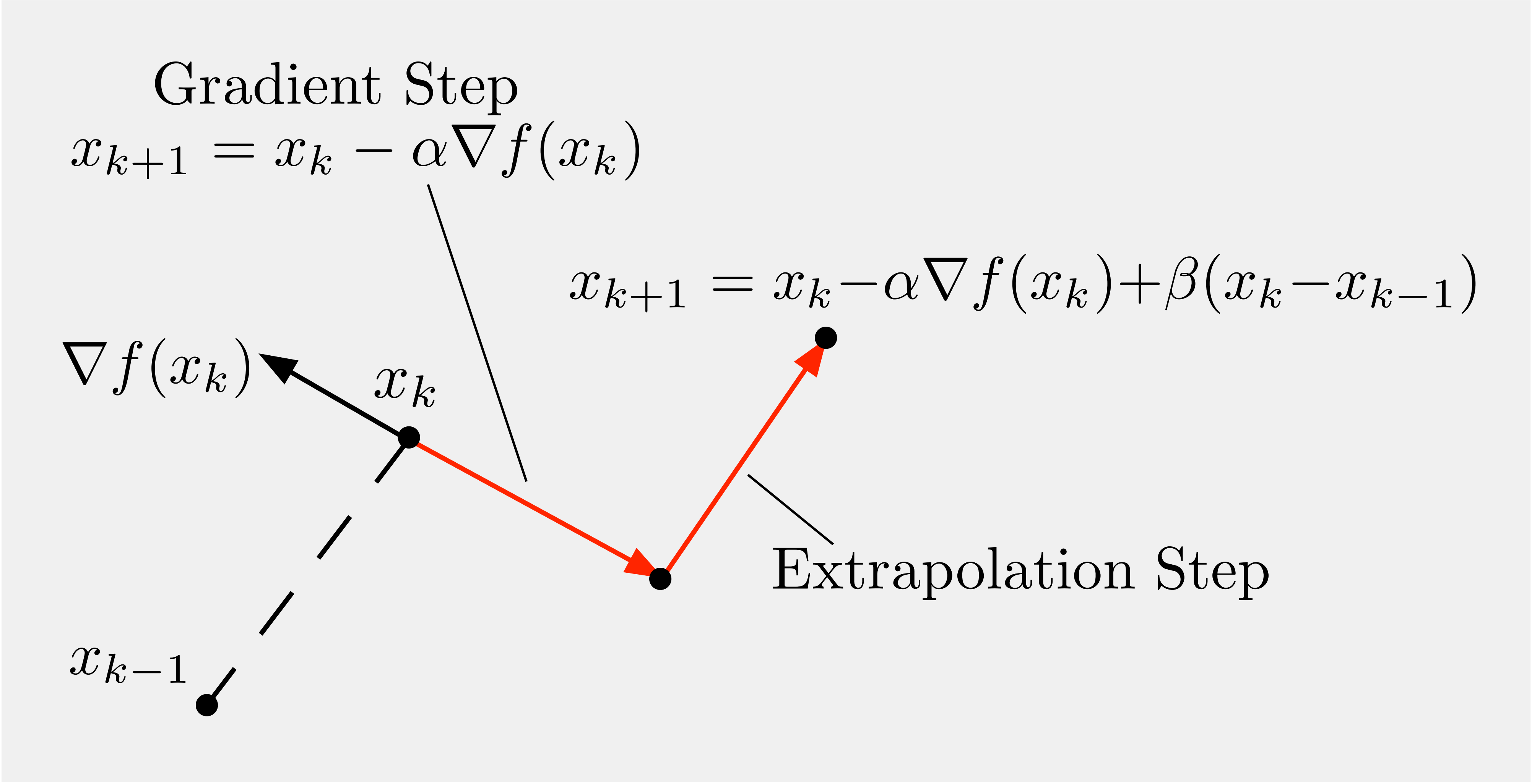
$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) + \beta_k (x_k - x_{k-1})$$

  - When f is quadratic, this is the Chebyshev Iterative Method
  - Momentum prevents oscillation due to local-driven i.e. gradient direction

- Can be re-written as a purely descent-type method:

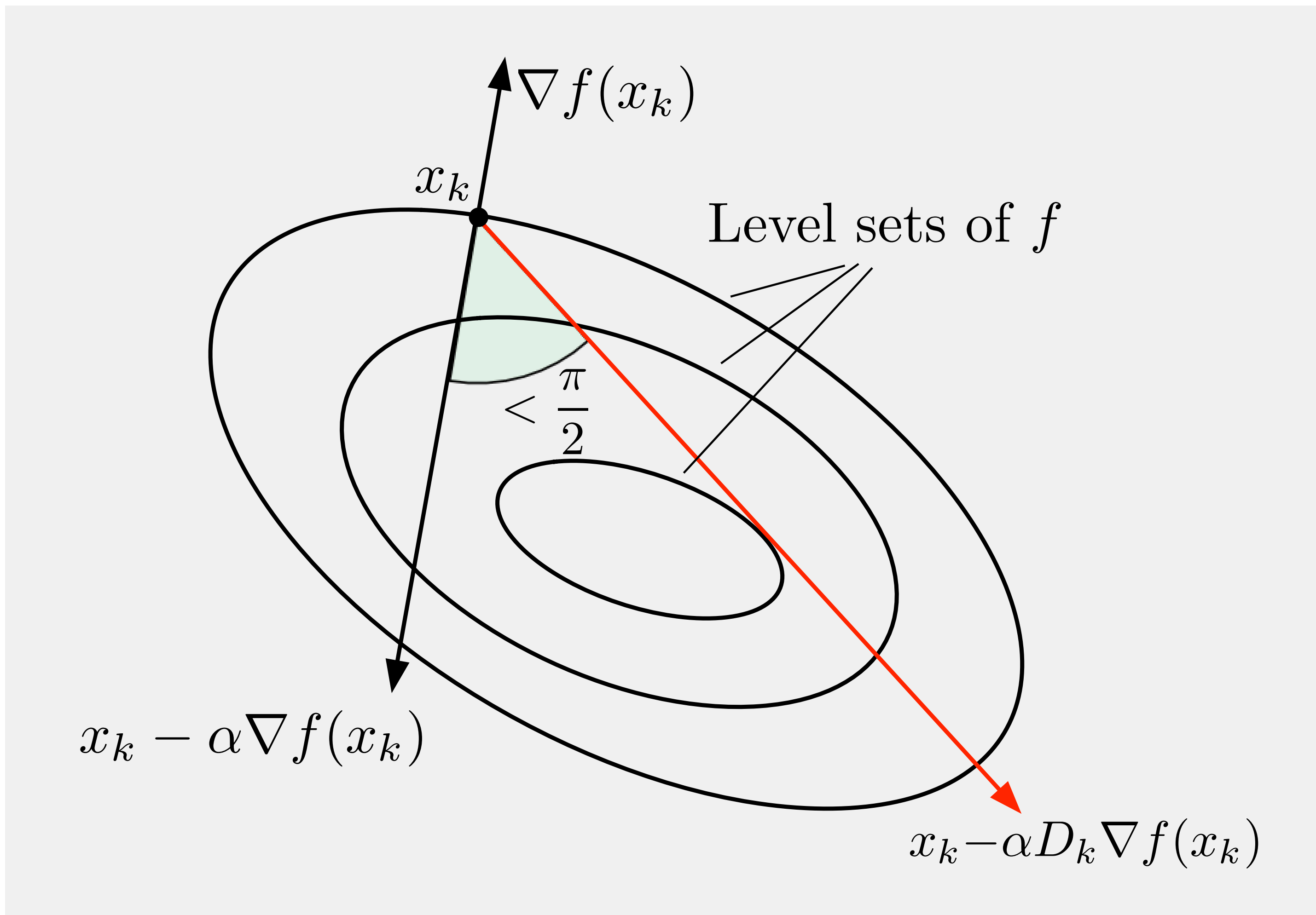
$$p_k = -\nabla f(x_k) + \beta_k p_{k-1}$$
$$x_{k+1} = x_k + \alpha_k p_k$$

Heavy Ball



Gradient Step
$x_{k+1} = x_k - \alpha \nabla f(x_k)$
$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$
$\nabla f(x_k)$
x_k
Extrapolation Step
x_{k-1}

Need not be a descent direction



Convergence Analysis

- Consider m-strongly convex functions, with L-Lipshitz gradients

  Let $\kappa := L/m$ be the condition number.

- Gradient descent with optimal step size has linear convergence with rate:

$$\|x_k - x^*\|_2 \leq \left(1 - \frac{2}{\kappa + 1}\right)^k \|x_0 - x^*\|_2$$

- Heavy Ball with optimal step sizes has linear convergence with rate:

$$\|x_k - x^*\|_2 \leq \left(1 - \frac{2}{\sqrt{\kappa} + 1}\right)^k \|x_0 - x^*\|_2$$

- Seemingly similar, but the square root makes a huge difference!

Convergence Analysis

- To yield $\|x_k - x^*\|_2 \leq \epsilon \|x_0 - x^*\|_2$, we need:

$$k > \frac{\kappa}{2}\log(1/\epsilon) \quad \text{for gradient descent}$$
$$k > \frac{\sqrt{\kappa}}{2}\log(1/\epsilon) \quad \text{for heavy ball}$$

- A factor of $\sqrt{\kappa}$ difference entails that if $\kappa = 100$, heavy ball needs 10 times fewer steps (i.e. is 10 times faster)

Recall: Conjugate Gradients

- Has similar form to heavy ball:

$$p_k = -\nabla f(x_k) + \beta_k \, p_{k-1}$$
$$x_{k+1} = x_k + \alpha_k \, p_k$$

  - Choose \beta_k to ensure p_k is conjugate to {p_1, …,p_{k-1}}

  - Choose \alpha_k by line search

- PRO:

  - Systematic approach to select parameters in heavy ball

- CON:

  - Does not achieve better rate than heavy ball, and convergence rates not completely resolved

  - Most ideal for quadratic rather than general functions

Optimality of Heavy Ball

- For strongly convex functions with Lipshitz gradient, rate of heavy ball is optimal

$$f(x) = x_1^2 + \sum_{i=1}^{n-1}(x_i - x_{i+1})^2 + x_n^2 - 2x_1 + \mu\|x\|_2^2$$

$$\boxed{\mu I \succeq \nabla^2 f(x) \succeq (4+\mu) I}$$

$$\boxed{\kappa \approx 1 + \frac{4}{\mu}}$$

- start at x[0] = e_1.
- after k steps, x[j] = 0 for j>k+1

- norm of the optimal solution on the unseen coordinates tends to $(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^{2k}$

Optimality of Heavy Ball

- For strongly convex functions with Lipshitz gradient, rate of heavy ball is optimal
- For convex functions with Lipshitz gradient, optimality unclear

Nesterov's Optimal Method

$$p_k = -\nabla f(x_k + \beta_k (x_k - x_{k-1})) + \beta_k p_{k-1}$$

$$x_{k+1} = x_k + \alpha_k p_k$$

Nesterov, 1983, 2004

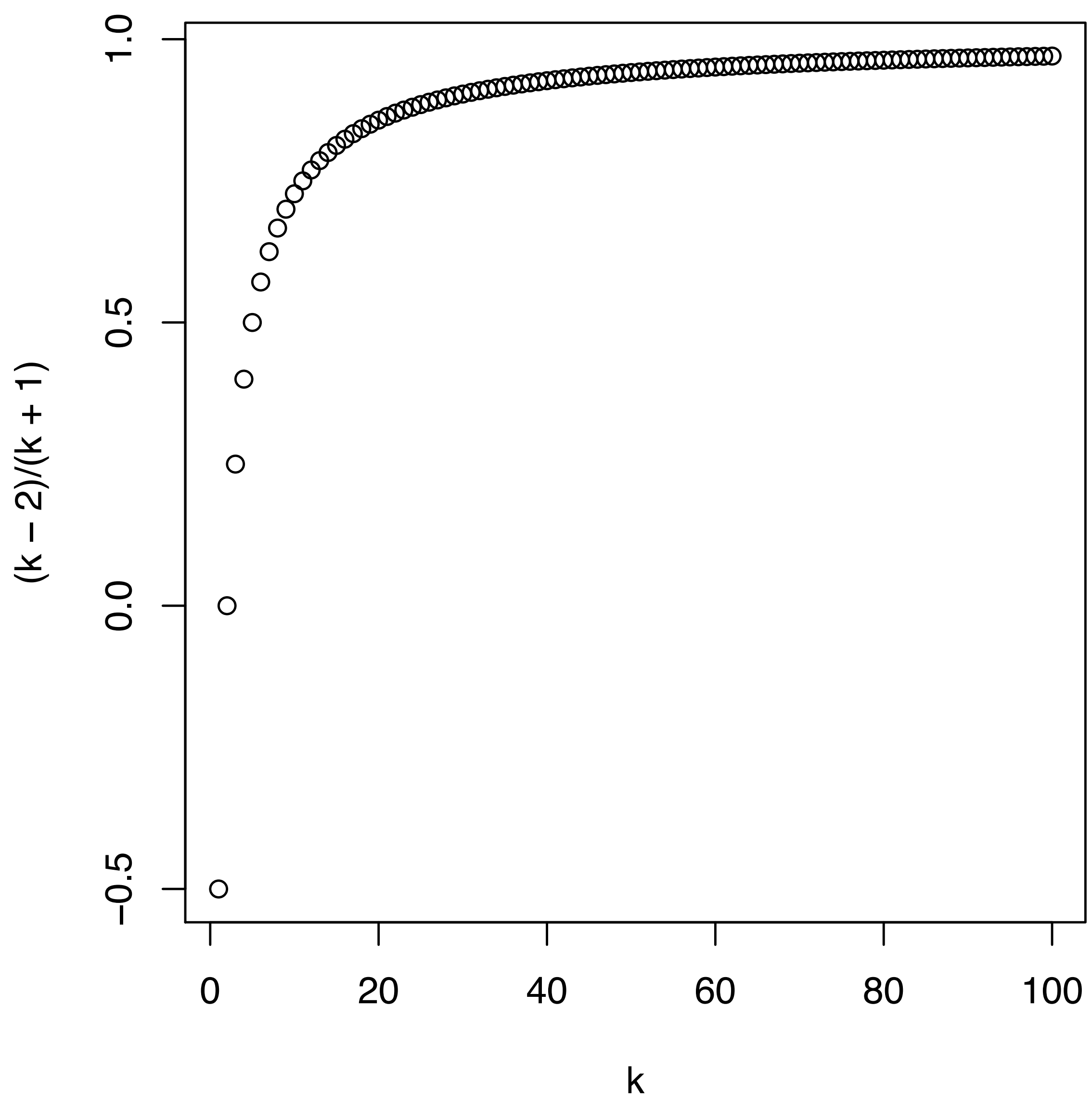
- Heavy Ball, but interchanging order of computing momentum and gradient terms
- Compute momentum and then compute gradient
- Standard settings of parameters:

$$\alpha_k = \frac{1}{L}$$

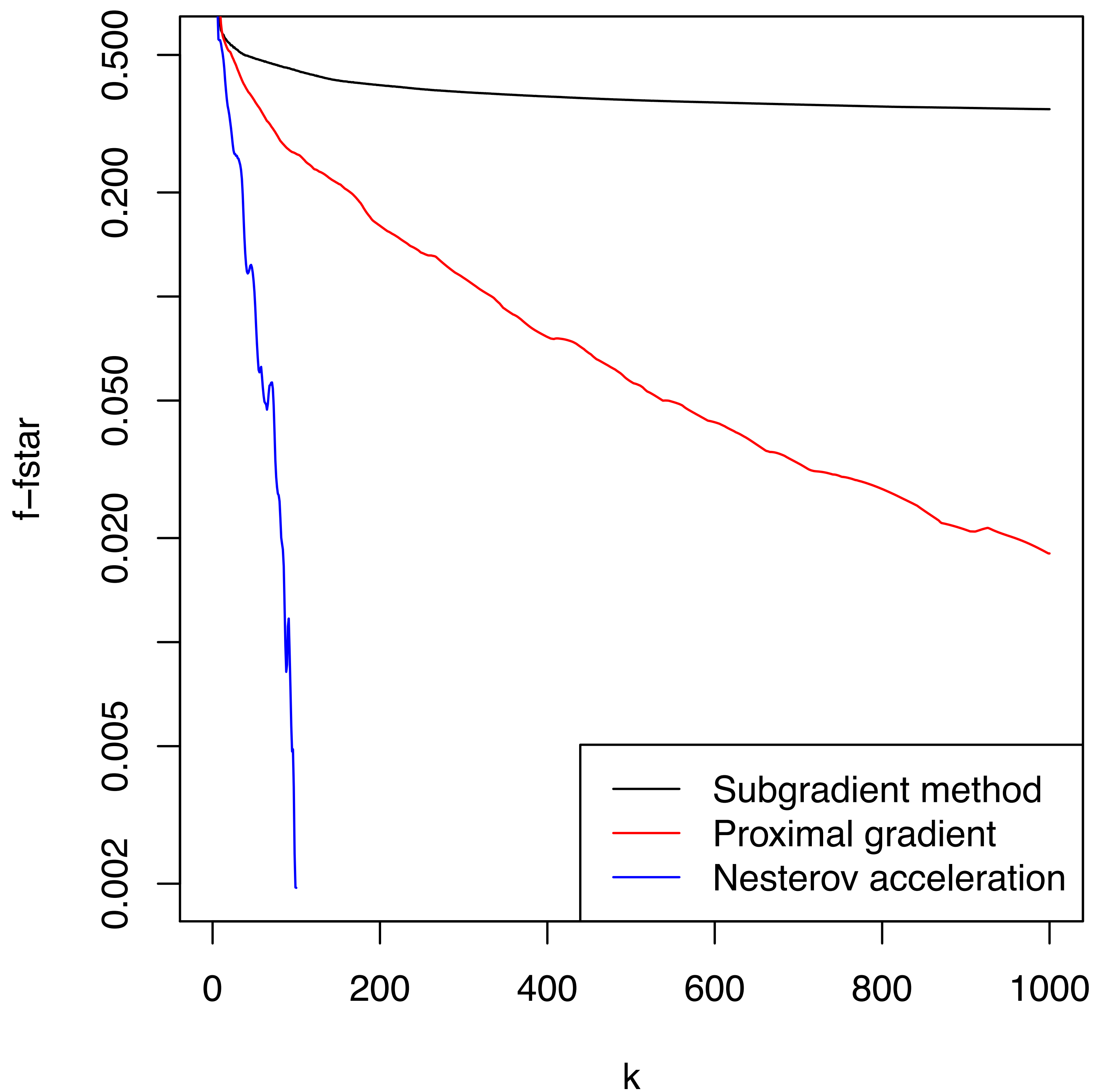
$$\beta_k = \frac{k-2}{k+1}$$

Nesterov Momentum Weights

Momentum weights:



Acceleration can really help



- Accelerated gradient is not strictly a descent method

- Notice the “Nesterov Ripples”

Nesterov's Optimal Method

$$p_k = -\nabla f(x_k + \beta_k (x_k - x_{k-1})) + \beta_k p_{k-1}$$
$$x_{k+1} = x_k + \alpha_k p_k$$

Nesterov, 1983, 2004

- Heavy Ball, but interchanging order of computing momentum and gradient terms

- Compute momentum and then compute gradient

- Standard settings of parameters:

$$\alpha_k = \frac{1}{L}$$
$$\beta_k = \frac{k-2}{k+1}$$

Line Search also achieves optimal rate modulo log factors

Convergence Analysis

- Consider convex functions, with L-Lipshitz gradients

- Gradient descent with optimal step size has convergence with rate:

$$f(x_k) - f(x^*) \leq \frac{2L\left\|x_0 - x^*\right\|_2^2}{k+4}$$

- Nesterov's Optimal Method has convergence with rate:

$$f(x_k) - f(x^*) \leq \frac{2L\left\|x_0 - x^*\right\|_2^2}{(k+2)^2}$$

- Seemingly similar, the square makes a huge difference!

Convergence Analysis

- To yield $f(x_k) - f(x^*) < \epsilon$, we need:

$$
\begin{aligned}
k &> \frac{2L\left\|x_0 - x^*\right\|_2^2}{\epsilon} - 4 \quad \text{for gradient descent} \\
k &> \frac{2L\left\|x_0 - x^*\right\|_2^2}{\sqrt{\epsilon}} - 2 \quad \text{for Nesterov's optimal method}
\end{aligned}
$$

- A factor of $\sqrt{\epsilon}$ difference entails that if $\epsilon = 10^{-4}$, optimal method needs 100 times fewer steps (i.e. is 100 times faster)