COM S 578X: Optimization for Machine Learning

Lecture Note 7: Accelerated First-Order Methods

Jia (Kevin) Liu

Assistant Professor
Department of Computer Science
Iowa State University, Ames, Iowa, USA

Fall 2019

Outline

In this lecture:

- Heavy-ball method
- Conjugate gradient
- Nesterov optimal first-order method
- FISTA
- Barzilai-Borwein
- Adding simple constraints
- Extending to regularized optimization

First-Order Methods

- So far, you've seen gradient descent – The most natural first-order method
- GD has a sublinear $O(1/k)$ rate for $\mathcal{F}_L^{\infty,1}$ and a slow linear rate for $\mathcal{S}_{\mu,L}^{2,1}$. Can we do better?
- First-Order Method (Nesterov): An iterative method generates a sequence of test points $\{\mathbf{x}_k\}$ such that

$$\mathbf{x}_k \in \mathbf{x}_0 + \text{span}\{\nabla f(\mathbf{x}_0), \nabla f(\mathbf{x}_1), \ldots, \nabla f(\mathbf{x}_{k-1})\}, \quad k \geq 1$$

Theorem 1 (Nesterov)

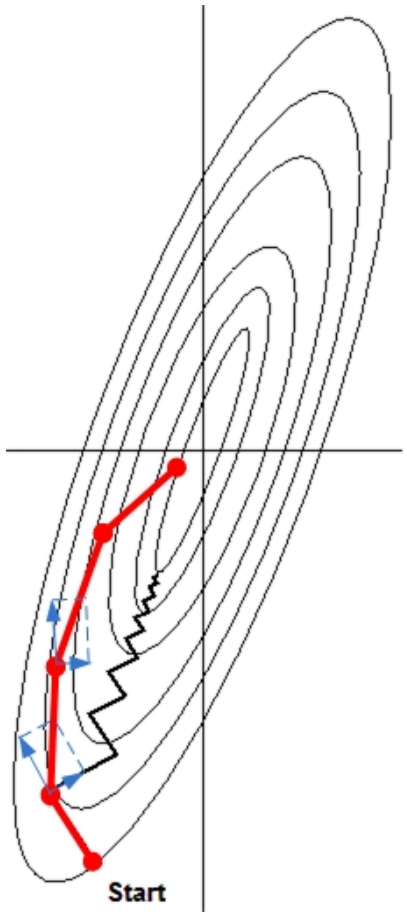
For any k, $1 \leq k \leq \frac{1}{2}(n-1)$, and any $\mathbf{x}_0 \in \mathbb{R}^n$, there exists a function $f \in \mathcal{F}_L^{\infty,1}$ such that for any first-order method, we have

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \geq \frac{3L\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{32(k+1)^2} = O(1/k^2)$$

- Pretty large gap btwn $O(1/k)$ and $O(1/k^2)$. So the answer is: Yes, we can!

Heavy-Ball

- Historical Perspective
  - First proposed by Polyak in 60s [Polyak '64]
  - Original goal: Accelerate gradient descent method
  - Inspired by 2nd-order ODE of heavy-body motion
  - Adopted in ML very early [Rumelhart, et al. '86]
- Basic Idea:
  - Search direction: Linear combination of current gradient (gravity) and previous search direction (momentum)
  - Also called two-step method and could be N-step in the general case



The Heavy-Ball Algorithm

Consider the unconstrained optimization problem, with f smooth and convex:

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$$

Denote the optimal value as $f^* = \min_{\mathbf{x}} f(\mathbf{x}^*)$ and an optimal solution as $\mathbf{x}^*$.

Heavy-Ball Method

Choose initial point $\mathbf{x}_0 \in \mathbb{R}^n$ and let $\mathbf{x}_{-1} = \mathbf{x}_0$. Repeat:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \nabla f(\mathbf{x}_k) + \beta_k(\mathbf{x}_k - \mathbf{x}_{k-1}), \quad k = 0, 1, 2, \ldots$$

$\beta_k \in [0, 1)$

GD

momentum.

Stop if some stopping criterion is satisfied.

Convergence Rate Analysis: Strongly Convex Case

Theorem 2 (Polyak)

If $f \in \mathcal{S}^{2,1}_{\mu,L}$, Heavy-Ball with constant step-size $s = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ and constant momentum coefficient $\beta = \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2$, satisfies:

$$\|\mathbf{x}_k - \mathbf{x}^*\|_2 \leq C \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \sqrt{2}\|\mathbf{x}_0 - \mathbf{x}^*\|_2,$$

where $\kappa \triangleq L/\mu$ and C is a constant.

HB Convergence Rate Analysis: Strongly Convex Case

Proof Sketch.

- Consider the new variable: $\mathbf{z}_k = [(\mathbf{x}_k - \mathbf{x}^*)^\top \quad (\mathbf{x}_{k-1} - \mathbf{x}^*)^\top]^\top$
- Under constant step-size s and momentum coefficient β, HB dynamics can be rewritten in terms of $\{\mathbf{z}_k\}$:

$$\|\mathbf{z}_{k+1}\|_2 = \left\| \begin{bmatrix} \mathbf{x}_{k+1} - \mathbf{x}^* \\ \mathbf{x}_k - \mathbf{x}^* \end{bmatrix} \right\|_2 = \left\| \begin{bmatrix} \mathbf{x}_k - s\nabla f(\mathbf{x}_k) + \beta(\mathbf{x}_k - \mathbf{x}_{k-1}) - \mathbf{x}^* \\ \mathbf{x}_k - \mathbf{x}^* \end{bmatrix} \right\|_2$$
$$= \left\| \begin{bmatrix} (1+\beta)\mathbf{I} & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}_k - \mathbf{x}^* \\ \mathbf{x}_{k-1} - \mathbf{x}^* \end{bmatrix} - s \begin{bmatrix} \nabla f(\mathbf{x}_k) \\ \mathbf{0} \end{bmatrix} \right\|_2 \tag{1}$$

- Note that

$$\nabla f(\mathbf{x}_k) = \nabla f(\mathbf{x}^*) + \int_0^1 \nabla^2 f(\mathbf{x}^* + \tau(\mathbf{x}_k - \mathbf{x}^*))(\mathbf{x}_k - \mathbf{x}^*) d\tau$$
$$= \underbrace{\left[\int_0^1 \nabla^2 f(\mathbf{x}^* + \tau(\mathbf{x}_k - \mathbf{x}^*)) d\tau \right]}_{\mathbf{B}(\mathbf{x}_k)} (\mathbf{x}_k - \mathbf{x}^*)$$

- Plugging this into (1) to obtain:

HB Convergence Rate Analysis: Strongly Convex Case

$$\mathbf{z}_{k+1} = \begin{bmatrix} \mathbf{x}_{k+1} - \mathbf{x}^* \\ \mathbf{x}_k - \mathbf{x}^* \end{bmatrix} = \begin{bmatrix} \mathbf{x}_k - s\nabla f(\mathbf{x}_k) + \beta(\mathbf{x}_k - \mathbf{x}_{k-1}) - \mathbf{x}^* \\ \mathbf{x}_k - \mathbf{x}^* \end{bmatrix}$$
$$= \underbrace{\begin{bmatrix} (1+\beta)\mathbf{I} - s\mathbf{B}(\mathbf{x}_k) & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}}_{\triangleq \mathbf{\Gamma}(\mathbf{x}_k)} \mathbf{z}_k \tag{2}$$

So, (2) implies $\|\mathbf{z}_k\|_2 = \|\prod_{i=0}^{k-1} \mathbf{\Gamma}(\mathbf{x}_i)\mathbf{z}_0\|_2 \leq \max_{\mathbf{x}_i, \forall i} \|\prod_{i=0}^{k-1} \mathbf{\Gamma}(\mathbf{x}_i)\mathbf{z}_0\|_2$.

- Noting the identical structure of $\mathbf{\Gamma}(\mathbf{x}_i)$ for i (implying index independence) and letting $\bar{\mathbf{\Gamma}} \triangleq \begin{bmatrix} (1+\beta)\mathbf{I} - s\mathbf{B}(\mathbf{x}^\triangle) & -\beta_k\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}$, where $\mathbf{x}^\triangle$ is a maximizer for the function $\|(\Gamma(\mathbf{x}))^k \mathbf{z}_0\|$. Then it can be shown that

$$\rho(\bar{\mathbf{\Gamma}}) \leq \sqrt{\beta},$$

where we take $\beta = \max\left\{|1 - \sqrt{s\mu}|, |1 - \sqrt{sL}|\right\}$.

HB Convergence Rate Analysis: Strongly Convex Case

- Plugging this into (2), we have:

$$\|\mathbf{z}_{k+1}\|_2 \leq \max\left\{|1-\sqrt{s\mu}|, |1-\sqrt{sL}|\right\}\|\mathbf{z}_k\|_2\,. \tag{3}$$

- If we let $s = \frac{4}{(\sqrt{\mu}+\sqrt{L})^2}$, we have:

$$(1-\sqrt{s\mu}) = (1-\sqrt{sL}) = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}} = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}.$$

- Substituting this back into (3) yields:

$$\|\mathbf{x}_k - \mathbf{x}^*\|_2 \leq \|\mathbf{z}_k\|_2 \leq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \sqrt{2}\|\mathbf{x}_0 - \mathbf{x}^*\|_2.$$

□

GD vs. HB under Strong Convexity

- Gradient Descent: Linear rate $\left(\frac{\kappa-1}{\kappa+1}\right)$
- Heavy-ball: Linear rate $\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)$

$\left(\frac{\kappa-1}{\kappa+1}\right) = 1 - \frac{2}{\kappa+1} \approx 1 - \frac{2}{\kappa}$

$\left(1-\frac{2}{\kappa}\right)^k \le \epsilon$. Take log on both sides:

$k \log\left(1-\frac{2}{\kappa}\right) \le \log \epsilon \Rightarrow k\left(-\log\left(1-\frac{2}{\kappa}\right)\right) \ge \log\frac{1}{\epsilon}$

$\Rightarrow k \ge \frac{+\log\epsilon}{+\log\left(1-\frac{2}{\kappa}\right)}$. Use $\log(1+x) \approx x$ if x is small.

$\Rightarrow k \ge \frac{-\log\epsilon}{-2/\kappa} = \frac{\kappa}{2}|\log\epsilon|$

Big difference! For $\|\mathbf{x}_k - \mathbf{x}^*\|$ to reach ϵ-accuracy, need k large enough so that

$$\left(\frac{\kappa-1}{\kappa+1}\right)^k \le \epsilon \quad \Rightarrow \quad k \ge \frac{\kappa}{2}|\log\epsilon| \qquad \text{(Gradient Descent)}$$

$$\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \le \epsilon \quad \Rightarrow \quad k \ge \frac{\sqrt{\kappa}}{2}|\log\epsilon| \qquad \text{(Heavy-Ball)}$$

A factor of $\sqrt{\kappa}$ difference; e.g., if $\kappa = 1000$ (not at all uncommon in ML problems), need approximately 30 times fewer iterations.

Convergence Rate Analysis: Weakly Convex Case

Theorem 3 (Ghadimi-Feyzmahdavian-Johansson)

If $f \in \mathcal{F}_L^{1,1}$, Heavy-Ball with $\beta_k = \frac{k}{k+2}$ and $\alpha_k = \frac{\alpha_0}{k+2}$, where $\alpha_0 \in (0, 1/L]$, satisfies:

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2\alpha_0(k+1)} = O(1/k),$$

Remark:

- HB has essentially the same sublinear rate as GD in the weakly convex case.

Conjugate Gradient

Definition 4 (The Notion of Conjugacy)

Let $\mathbf{H} \in \mathbb{R}^{n \times n}$ be symmetric. The vectors $\mathbf{d}_1, \ldots, \mathbf{d}_n$ are called $\mathbf{H}$-conjugate (or simply conjugate) if they are linearly independent and if $\mathbf{d}_i^\top \mathbf{H} \mathbf{d}_j = 0$ for all $i \neq j$.

Conjugacy is significant in quadratic functions minimization (often arise in ML):

- Let $f(\mathbf{x}) = \mathbf{c}^\top \mathbf{x} + \frac{1}{2}\mathbf{x}^\top \mathbf{H}\mathbf{x}$. Suppose $\mathbf{d}_1, \ldots, \mathbf{d}_n$ are $\mathbf{H}$-conjugate. Then:

Pre lin. indep. of $\underline{d}_1 \cdots \underline{d}_n$: Given starting pt. $\underline{x}_1$. Any pt. $\underline{x}$ can written as: $\underline{x} = \underline{x}_1 + \sum_{i=1}^{n} s_i \underline{d}_i$. Therefore, $f(\underline{x}) = \underline{c}^T(\underline{x}_1 + \sum_{j=1}^{n} s_j \underline{d}_j) + \frac{1}{2}(\underline{x}_1 + \sum_{j=1}^{n} s_j \underline{d}_j)^T \underline{H}(\underline{x}_1 + \sum_{j=1}^{n} s_j \underline{d}_j)$.

$= \underline{c}^T\underline{x}_1 + \sum_{j=1}^{n} s_j \underline{c}^T\underline{d}_j + \frac{1}{2}\left[\underline{x}_1^T \underline{H}\, \underline{x}_1 + 2\underline{x}_1^T \underline{H}\left(\sum_{j=1}^{n} s_j \underline{d}_j\right) + \sum_{j=1}^{n}\sum_{i=1}^{n} s_i s_j \underline{d}_i^T \underline{H}\, \underline{d}_j\right]$

$= \underline{c}^T\underline{x}_1 + \sum_{j=1}^{n} s_j \underline{c}^T\underline{d}_j + \frac{1}{2}\left[\underline{x}_1^T \underline{H}\, \underline{x}_1 + 2\underline{x}_1^T \underline{H}\left(\sum_{j=1}^{n} s_j \underline{d}_j\right) + \sum_{i=1}^{n} s_i^2 \underline{d}_i^T \underline{H}\, \underline{d}_i\right]$ ← $\underline{H}$-conjugacy

$\Rightarrow \sum_{j=1}^{n}\left[\underline{c}^T\underline{x}_1 + s_j \underline{c}^T\underline{d}_j + \frac{1}{2}(\underline{x}_1 + s_j \underline{d}_j)^T \underline{H}(\underline{x}_1 + s_j \underline{d}_j)\right] \triangleq F(\underline{x})$ ← $F(\underline{x})$ can be minimized by minimizing each term in $\sum$. Solving for $s_j \Rightarrow s_j^* = -[\underline{c}^T\underline{d}_j + \underline{x}_1^T \underline{H}\, \underline{d}_j]/\underline{d}_j^T \underline{H}\, \underline{d}_j$

(add $(n-1)$ copies of $\frac{1}{2}\underline{x}_1^T \underline{H}\, \underline{x}_1$ & $\underline{c}^T\underline{x}_1$)

- Note: Conjugate directions are not unique

① We can start from $\underline{x}_1$, along ANY order of $\underline{d}_1 \cdots \underline{d}_n$

② We compute s_j^* in parallel and combine together. (Distr. implementation).

Conjugate Directions: Geometric Interpretation

$\min -12x_2 + 4x_1^2 + 4x_2^2 - 4x_1x_2$.

$H = \begin{bmatrix} 8 & -4 \\ -4 & 8 \end{bmatrix} > 0$

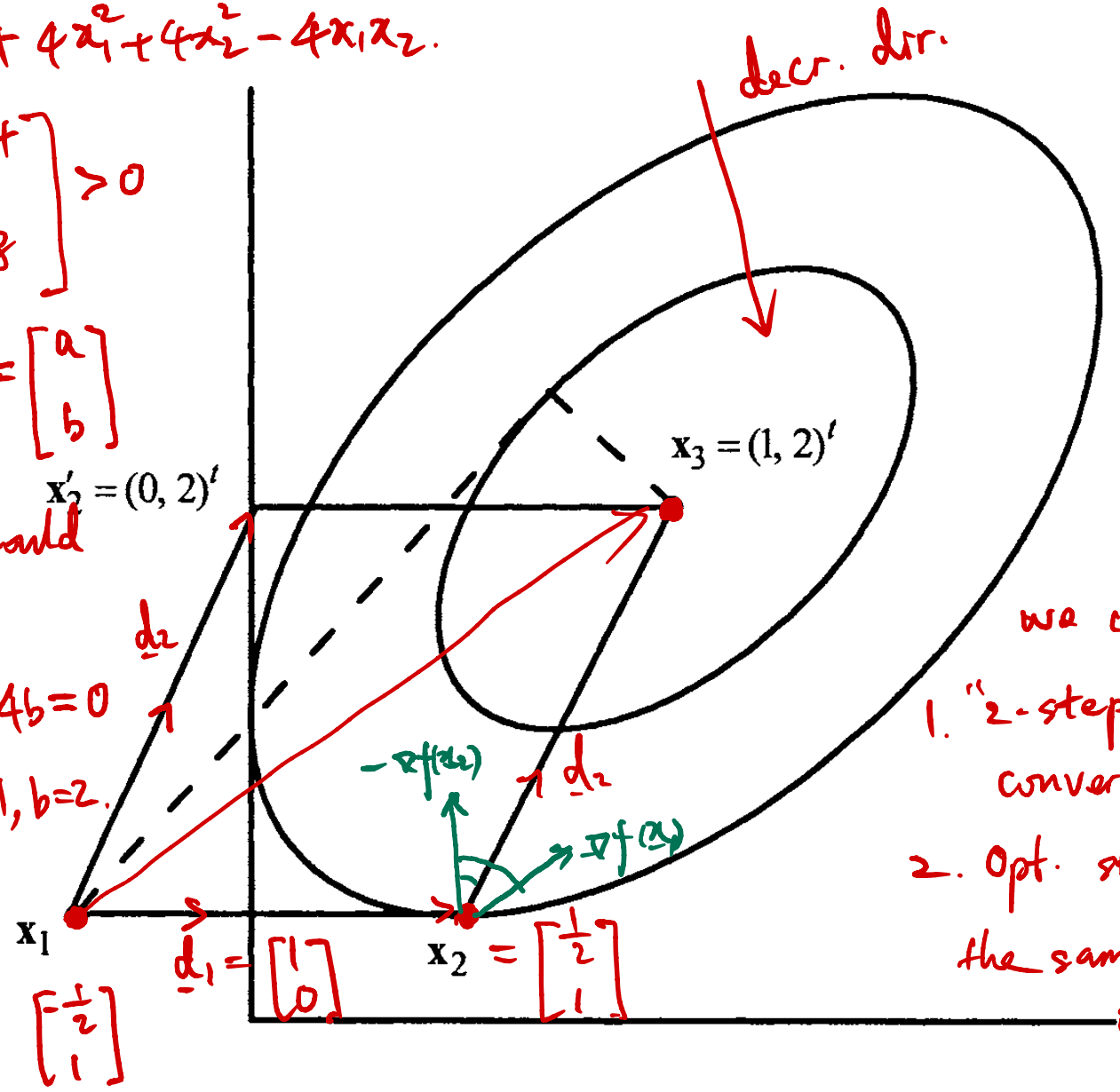
$d_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $d_2 = \begin{bmatrix} a \\ b \end{bmatrix}$

d_1 and d_2 should satisfy:

$d_1^T H d_2 = 8a - 4b = 0$

Let's pick $a=1, b=2$.

i.e., $d_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$



we can see:

1. "2-step" exact convergence.
2. Opt. step-size are the same regardless of the path choice.

Quadratic Minimization w/ Conjugacy: Finite Convergence

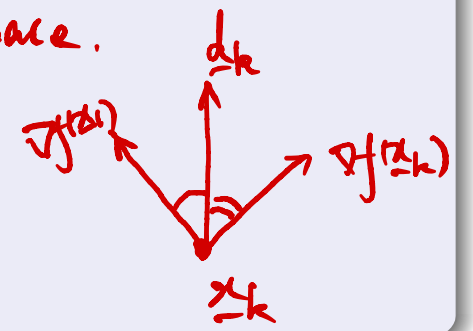
Exact LS.

Theorem 5 (Expanding subspace property)

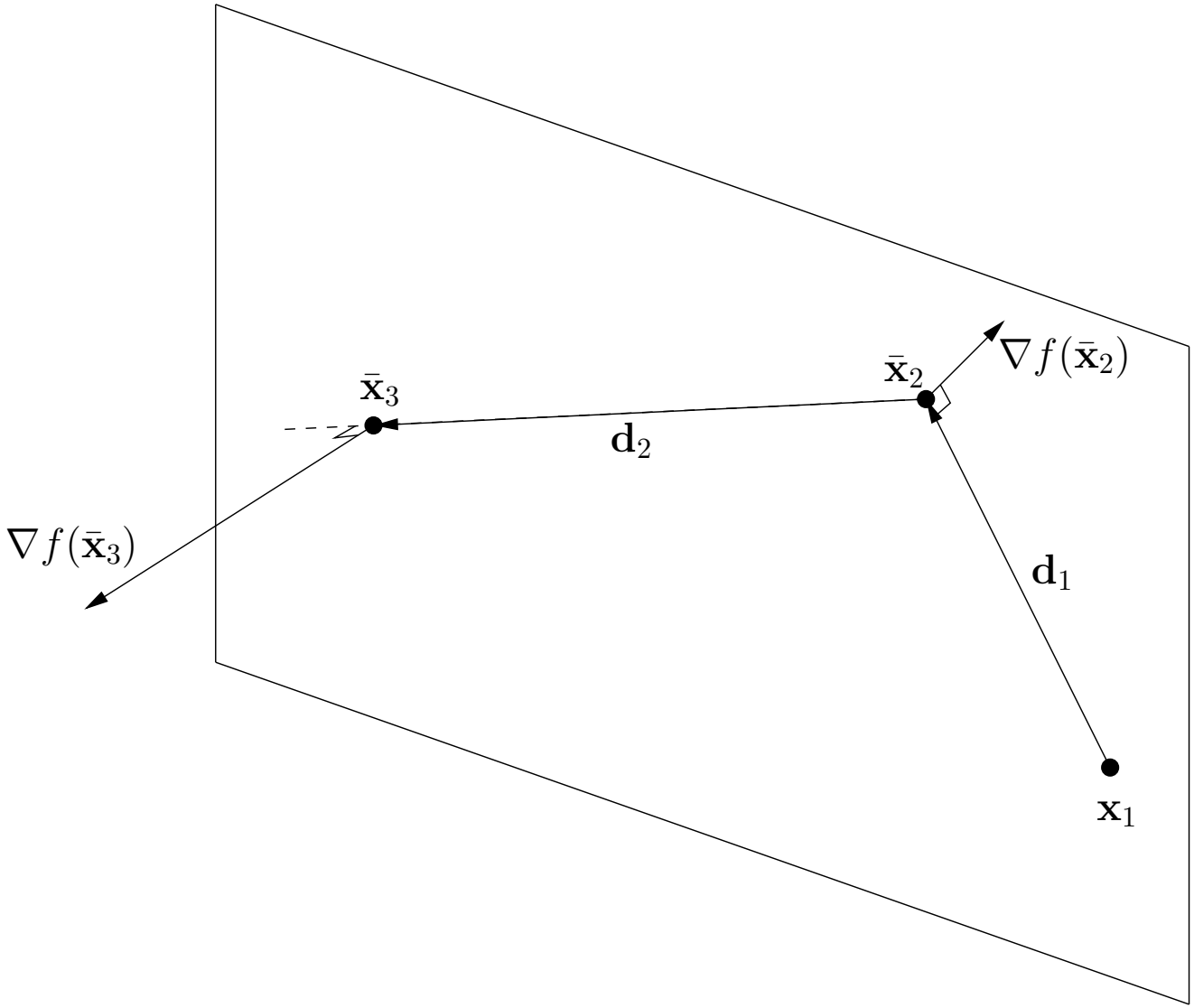
Let $f(\mathbf{x}) = \mathbf{c}^\top \mathbf{x} + \frac{1}{2}\mathbf{x}^\top \mathbf{H}\mathbf{x}$, *where* $\mathbf{H} \in \mathbb{R}^{n\times n}$ *is symmetric. Let* $\mathbf{d}_1, \ldots, \mathbf{d}_n$ *be* $\mathbf{H}$*-conjugate. Let* $\mathbf{x}_1$ *be arbitrary starting point. Let* $s_k = \arg\min_{s\in\mathbb{R}} f(\mathbf{x}_k + s\mathbf{d}_k)$, $k = 1, \ldots, n$, *and let* $\mathbf{x}_{k+1} = \mathbf{x}_k + s_k\mathbf{d}_k$. *Then, for* $k = 1, \ldots, n$, *we must have:*

- $\nabla f(\mathbf{x}_{k+1})^\top \mathbf{d}_j = 0$, *for* $j = 1, \ldots, k$. ← expanding subspace.
- $\nabla f(\mathbf{x}_1)^\top \mathbf{d}_k = \nabla f(\mathbf{x}_k)^\top \mathbf{d}_k$.
- $\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} \left\{ f(\mathbf{x}) | \mathbf{x} - \mathbf{x}_1 \in \text{span}\{\mathbf{d}_1, \ldots, \mathbf{d}_k\} \right\}$
- $\mathbf{x}_{n+1}$ *is a minimizing point of* $f(\mathbf{x})$ *over* $\mathbb{R}^n$

K. cone.



Geometric Interpretation



Conjugate Gradient (Standard Version)

1. Initial point $\mathbf{x}_1$, $\mathbf{y}_1 = \mathbf{x}_1$, $\mathbf{d}_1 = -\nabla f(\mathbf{y}_1)$, $k = j = 1$.
2. If $\|\nabla f(\mathbf{y}_j)\| \le \epsilon$, stop. Otherwise, let $s_j = \arg\min_{s \ge 0} f(\mathbf{y}_j + s\mathbf{d}_j)$. Let $\mathbf{y}_{j+1} = \mathbf{y}_j + s_j \mathbf{d}_j$. If $j = n$, go to Step 4
3. Let $\mathbf{d}_{j+1} = -\nabla f(\mathbf{y}_{j+1}) + \gamma_j \mathbf{d}_j$. Let $j = j + 1$ and go to Step 2.
4. Let $\mathbf{x}_{k+1} = \mathbf{y}_1 = \mathbf{y}_{n+1}$, $\mathbf{d}_1 = -\nabla f(\mathbf{y}_1)$. Let $j = 1$, $k = k + 1$. Go to Step 2.

- Can be viewed as heavy-ball, with $\beta_j = \frac{s_j \gamma_j}{s_{j-1}}$. But CG can be implemented without requiring knowledge (or estimation) of L and μ
  - Choose s_j to (approximately) minimize f along $\mathbf{d}_j$. Variants of choosing γ_j:

$$\text{Fletcher-Reeves: } \gamma_j^{\text{FR}} = \frac{\|\nabla f(\mathbf{y}_{j+1})\|^2}{\|\nabla f(\mathbf{y}_j)\|^2}$$

$$\text{Polak-Rebiere: } \gamma_j^{\text{PR}} = \frac{\nabla f(\mathbf{y}_{j+1})^\top (\nabla f(\mathbf{y}_{j+1}) - \nabla f(\mathbf{y}_j))}{\|\nabla f(\mathbf{y}_j)\|^2}$$

$$\text{Hestenes-Stiefel: } \gamma_j^{\text{HS}} = \frac{s_j \nabla f(\mathbf{y}_{k+1})^\top (\nabla f(\mathbf{y}_{j+1}) - \nabla f(\mathbf{y}_j))}{(\mathbf{y}_{j+1} - \mathbf{y}_j)^\top (\nabla f(\mathbf{y}_{j+1}) - \nabla f(\mathbf{y}_j))}$$

  - All equivalent if f is convex quadratic & exact line search is used ([BSS Ch.8])

Conjugate Gradient

- Nonlinear CG: Variants include Fletcher-Reeves, Polak-Ribiere, Hestnes-Stiefel, etc.
- Restarting periodically with $\mathbf{d}_1 = -\nabla f(\mathbf{x}_k)$ (e.g., every n iterations, or when $\mathbf{d}_j$ fails to be a descent direction)
- For quadratic $f(\cdot)$, convergence analysis is based on eigenvalues of $\mathbf{A}$ and Chebyshev polynomials, min-max arguments. Get:
  - Finite termination is as many as iterations as there are distinct eigenvalues
  - Asymptotic linear convergence with rate approximately $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ (like heavy-ball, see [Necedal and Wright, Ch.5])
  - Similar sublinear $O(1/k)$ rate for $f \in \mathcal{F}_L^{1,1}$

Can we close the gap between $O(1/k)$ and $O(1/k^2)$ for $\mathcal{F}_L^{1,1}$?

Nesterov Accelerated First-Order Method

- Consider an unconstrained convex optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}), \text{ where } f(\cdot) \text{ has } L\text{-Lipschitz cont. gradient.}$$

- Nesterov's AGD method [Nesterov '83]:
  1. Choose initial $\mathbf{x}_0$. Let $\mathbf{y}_0 = \mathbf{x}_0$. Choose $\alpha_0 \in (0,1)$.
  2. Iteration $k \geq 0$: a) Compute $f'(\mathbf{y}_k)$ and let:

$$\mathbf{x}_{k+1} = \mathbf{y}_k - (1/L)f'(\mathbf{y}_k). \quad \text{(regular gradient step)}$$

b) Compute $\alpha_{k+1} \in (0,1)$ by solving $\alpha_{k+1}^2 + [\alpha_k^2 - (1/\kappa)]\alpha_{k+1} - \alpha_k^2 = 0$.

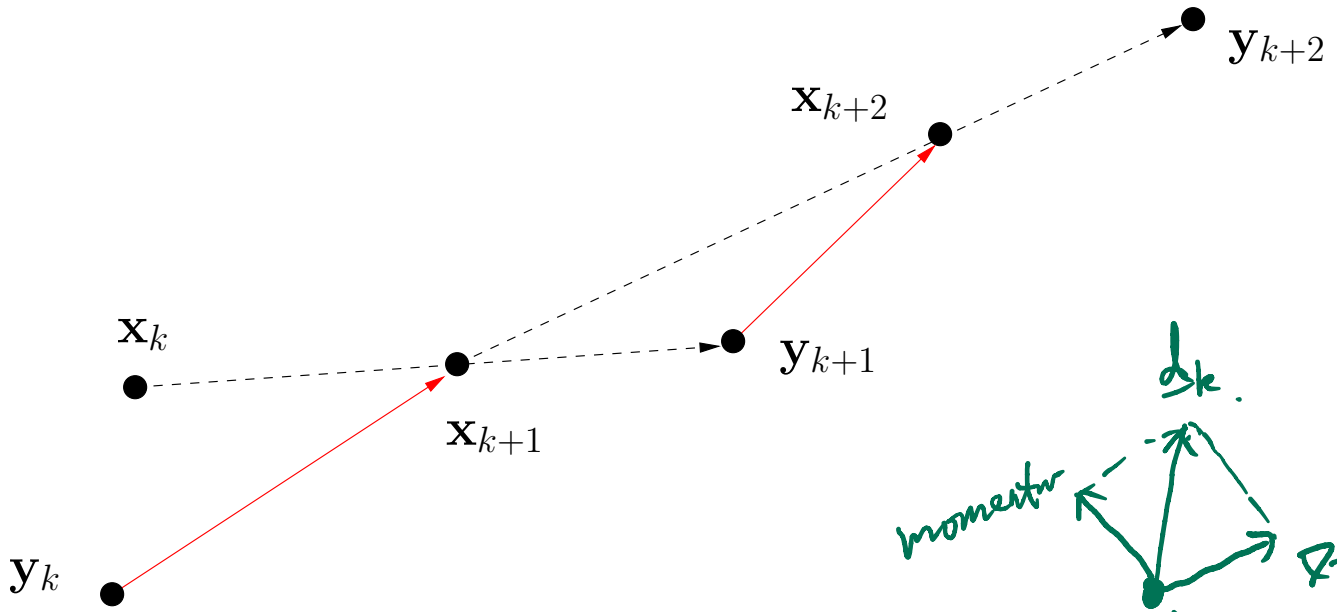
$$\text{Let } \beta_k = \frac{\alpha_k(1-\alpha_{k+1})}{(\alpha_k^2 + \alpha_{k+1})}, \text{ and}$$

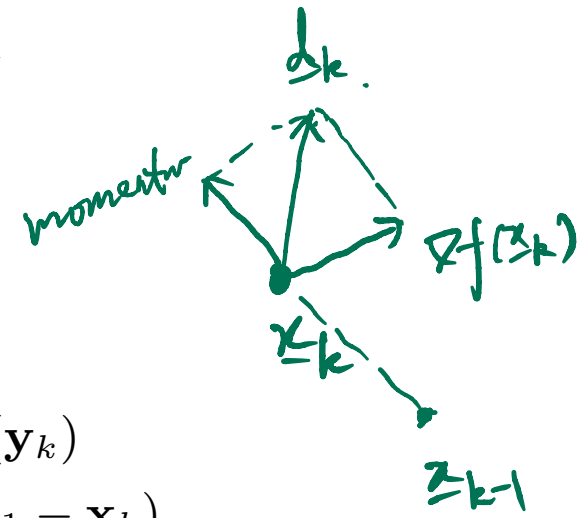
$$\mathbf{y}_{k+1} = \mathbf{x}_{k+1} + \beta_k(\mathbf{x}_{k+1} - \mathbf{x}_k). \quad \text{("slide" in dir. of } \mathbf{x})$$

- Still works for weakly convex ($\kappa = \infty$)

$y_k - \frac{1}{L}\nabla f(y_k)$.

Nesterov's Accelerated First-Order Method





- In simplest form (constant step-sizes):

$$\mathbf{x}_{k+1} = \mathbf{y}_k - (1/L) f'(\mathbf{y}_k)$$
$$\mathbf{y}_{k+1} = \mathbf{x}_{k+1} + \beta(\mathbf{x}_{k+1} - \mathbf{x}_k).$$

- Nesterov's AGD achieves:
  - Strongly convex: Linear convergence rate insensitive to $\kappa \triangleq L/\mu$
  - Weakly convex: Sublinearly $O(1/t^2)$ (order-optimal).

[Su, Boyd, Candes, NIPS'16] : ODE interpretation.

Convergence Performance of Nesterov's Method

Theorem 6

The Nesterov accelerated first-order method with $\alpha[0] \geq 1/\sqrt{\kappa}$ *satisfies:*

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq c_1 \min\left\{\left(1 - \frac{1}{\sqrt{\kappa}}\right)^k, \frac{4L}{(\sqrt{L} + c_2 k)^2}\right\},$$

where constants c_1 *and* c_2 *depend on* $\mathbf{x}_0$, $\alpha[0]$, *and* L.

Remark:

- Linear convergence rate similar to Heavy-Ball if $f \in \mathcal{S}^{2,1}_{\mu,L}$
- Sublinear $O(1/k^2)$ if $f \in \mathcal{F}^{1,1}_{L}$
- In the special case where $\alpha[0] = 1/\sqrt{\kappa}$, this scheme yields:

$$\alpha[k] = \frac{1}{\sqrt{\kappa}}, \quad \beta[k] = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}$$

FISTA

- Beck and Teboulle proposed a similar momentum-based algorithm:
  1. Choose initial $\mathbf{x}_0$. Let $\mathbf{y}_1 = \mathbf{x}[0]$. $t_1 = 1$
  2. Iteration $t \geq 0$: Do the following computations:

$$\mathbf{x}_k \leftarrow y_k - \frac{1}{L}\nabla f(\mathbf{y}_k);$$

$$t_{k+1} \leftarrow \frac{1}{2}\left(1 + \sqrt{1 + 4t_k^2}\right)$$

$$\mathbf{y}_{k+1} \leftarrow \mathbf{x}_k + \frac{t_k - 1}{t_{k+1}}(\mathbf{x}_k - \mathbf{x}_{k-1})$$

$y - t\nabla f(y_k)$.

difference.

- For weakly convex f, converges with $f(\mathbf{x}_k) - f(\mathbf{x}^*) = O(1/k^2)$
- When L is unknown, increase an estimate of L until it's big enough

Backtracking.

Nonmonotone Gradient Method: Barzilai-Borwein

- Barzilai and Borwein (BB) proposed an unusual choice of s_k
- Allows f to increase (sometimes a lot) on some steps: non-monotone

$$\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \nabla f(\mathbf{x}_k), \qquad s_k = \arg\min_s \|\mathbf{r}_k - s\mathbf{q}_k\|^2,$$

where $\mathbf{r}_k = \mathbf{x}_k - \mathbf{x}_{k-1}$ and $\mathbf{q}_k \triangleq \nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k-1})$

- Explicitly, we have

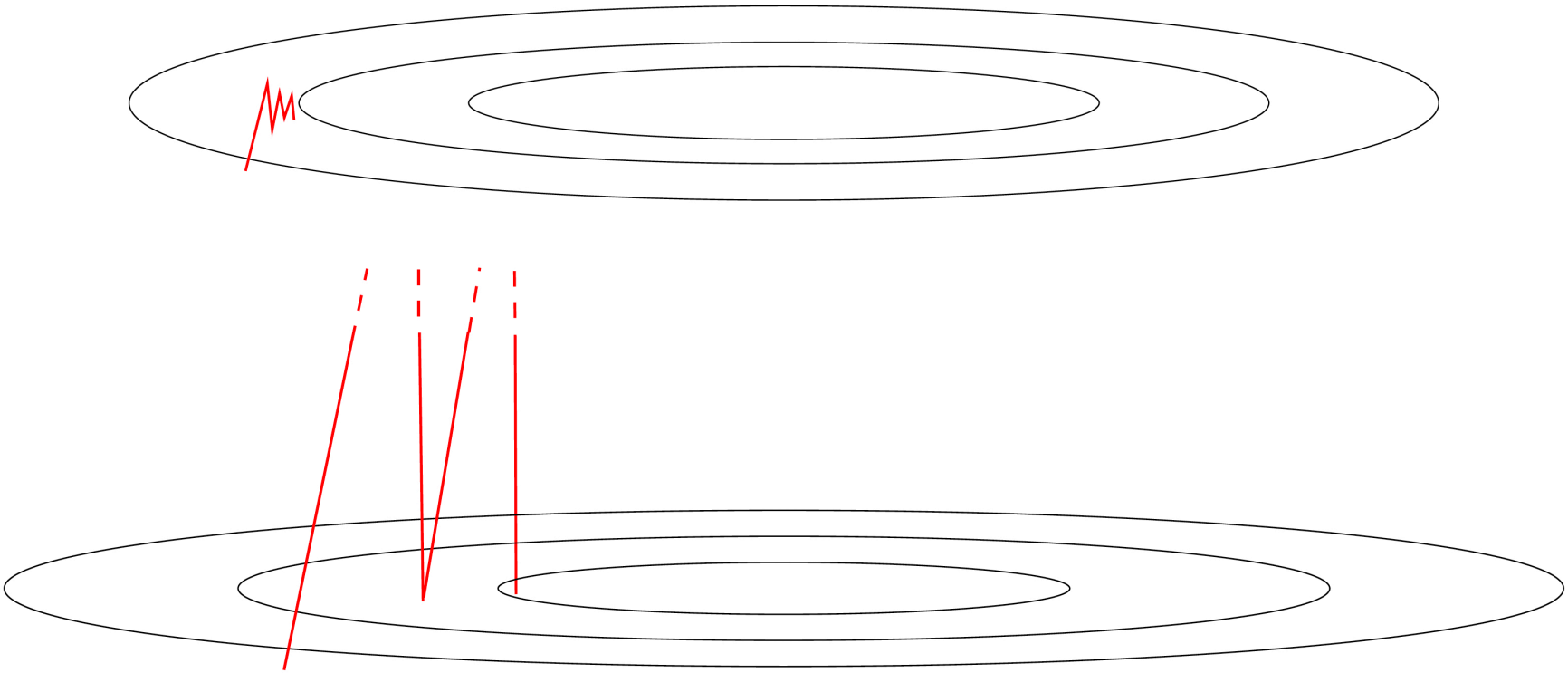
$$s_k = \frac{\mathbf{r}_k^\top \mathbf{q}_k}{\mathbf{q}_k^\top \mathbf{q}_k}$$

Note that for $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x}$, we have

$$s_k = \frac{\mathbf{r}_k^\top \mathbf{A}\mathbf{r}_k}{\mathbf{r}_k^\top \mathbf{A}^2\mathbf{r}_k} \in \left[\frac{1}{L}, \frac{1}{\mu}\right]$$

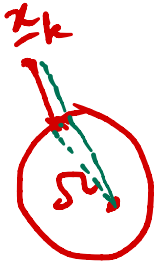
- BB can be viewed as quasi-Newton, with the Hessian approximated by $s_k^{-1}\mathbf{I}$

Comparison: BB vs Gradient Descent



Extension to Simple Constraints

- Constraint set Ω: a (relatively simple) closed convex set
- Some algorithms and theory stay largely the same, if we can involve the constraint $\mathbf{x} \in \Omega$ explicitly in the subproblems
- Example: Nesterov's constant step scheme requires just one calculation to be changed from the unconstrained version as follows:
  1. Choose initial $\mathbf{x}_0$. Let $\mathbf{y}_0 = \mathbf{x}_0$. Choose $\alpha_0 \in (0, 1)$.
  2. Iteration $k \geq 0$: a) Compute $f'(\mathbf{y}_k)$ and let:



$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{y} \in \Omega} \{\|\mathbf{y} - (\mathbf{y}_k - (1/L) f'(\mathbf{y}_k))\|\}. \quad \text{(projected gradient step)}$$

b) Compute $\alpha_{k+1} \in (0, 1)$ by solving $\alpha_{k+1}^2 + [\alpha_k^2 - (1/\kappa)]\alpha_{k+1} - \alpha_k^2 = 0$.

$$\text{Let } \beta_k = \frac{\alpha_k(1 - \alpha_{k+1})}{(\alpha_k^2 + \alpha_{k+1})}, \text{ and}$$

$$\mathbf{y}_{k+1} = \mathbf{x}_{k+1} + \beta_k(\mathbf{x}_{k+1} - \mathbf{x}_k). \quad \text{("slide" in dir. of } \mathbf{x})$$

- Convergence theory is unchanged.

Extension to Regularized Optimization

- Consider the following optimization with regularization:

$$\min_{\mathbf{x}} f(\mathbf{x}) + \tau\psi(\mathbf{x}),$$

  where $f \in \mathcal{F}_{1,1}^{L}$ and ψ is convex but usually nonsmooth

- Often, we only need to change the update to:

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} \frac{1}{2s_k}\|\mathbf{x} - (\mathbf{x}_k + s_k\mathbf{d}_k)\|_2^2 + \tau\psi(\mathbf{x}),$$

  where $\mathbf{d}_k$ could be a scaled gradient descent step, or deflected gradient (e.g., heavy-ball, Nesterov, etc.), while s_k is the step size

- This is also referred to as shrinkage/thresholding step. More on how to solve the above problem later when we discuss sparse/regularized optimization

Next Class

Subgradient Method