IE 598: Big Data Optimization **Fall 2016**

Lecture 9: Gradient Descent and Acceleration – September 20

Lecturer: Niao He *Scriber: Samantha Thrush*

In the previous lecture, we had introduced Gradient Descent for Smooth Convex Optimization problems. In this lecture, we will delve further in to the details of Gradient Descent for strongly convex problems and will discuss the mathematical and historical background of Accelerated Gradient Descent.

9.1 Review

Recall from last time that we considered the following unconstrained problem:

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

where f is L-smooth and convex.

The simplest gradient descent acts as follows:

$$x_{t+1} = x_t - \frac{1}{L}\nabla f(x_t)$$

which enjoys a $O(1/t)$ sublinear rate of convergence,

$$f(x_t) - f(x_*) \leq \frac{2L||x_0 - x_*||^2}{t}$$

The key to this convergence is based on the following observation:

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

Now we want to take this one step further and show that Gradient Descent achieves a linear convergence rate for strongly convex problems.

9.2 Strongly Convex Problems

Definition 9.1 *f is μ-strongly convex ($\mu \geq 0$) if $f(x) - \frac{\mu}{2}||x||_2^2$ is convex, where μ is the strong-convexity parameter.*

Proposition 9.2 *The following are all equivalent:*

(a) *f is continuously differentiable and μ-strongly convex*

(b) *f is continuously differentiable and $f(\alpha x+(1-\alpha)y) \leq \alpha f(x)+(1-\alpha)f(y)-\frac{\mu}{2}\alpha(1-\alpha)||x-y||_2^2, \forall \alpha \in [0,1]$*

(c) *$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}||x-y||_2^2$*

(d) *$\langle \nabla f(x) - \nabla f(y), x-y\rangle \geq \mu||x-y||_2^2$*

Remark 1. Recall the previous notes on equivalent characterization for L-smooth functions, one could observe that there exist some laws of symmetry between L-smoothness and μ-convexity. Notice that if $\mu = 0$ then properties (b), (c), (d) will reduce to a form that is applicable to general convex functions. If a function f is both L-smooth and μ-strongly convex, then usually we have $\mu \leq L$.

Remark 2. If f is further twice-differentiable, then f is μ-strongly convex if and only if $\nabla^2 f \succeq \mu I$.

Proof:

- **(a)** $\iff$ **(b)** Since $f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex, this implies that for any $x, y, \alpha \in [0, 1]$,

  $$f(\alpha x + (1-\alpha)y) - \frac{\mu}{2}\|\alpha x + (1-\alpha)y\|_2^2 \leq \alpha[f(x) - \frac{\mu}{2}\|x\|_2^2] + (1-\alpha)[f(y) - \frac{\mu}{2}\|y\|_2^2]$$

  Rearranging the terms, we will arrive at (b).

- **(b)** $\implies$ **(c)** From (b), we have $(1-\alpha)f(y) - (1-\alpha)f(x) \geq f(\alpha x + (1-\alpha)y) - f(x) + \frac{\mu}{2}\alpha(1-\alpha)\|x-y\|_2^2$. Hence,

  $$f(y) - f(x) \geq \frac{f(x + (1-\alpha)(y-x)) - f(x)}{1-\alpha} + \frac{\mu}{2}\alpha\|x-y\|_2^2, \forall \alpha \in [0, 1]$$

  Let α goes to one, thus leading to (c).

- **(c)** $\implies$ **(d)** From (c), we have

  $$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}||x-y||_2^2$$

  $$f(x) \geq f(y) - \nabla f(y)^T(x-y) + \frac{\mu}{2}||x-y||_2^2$$

  Adding these two equations together leads to (d).

- **(d)** $\implies$ **(b)** Let $z = \alpha x + (1-\alpha)y$, from the fundamental theory of calculus, we can see the following:

  $$f(z) = f(x) + \textstyle\int_0^1 \nabla f(x + t(z-x))^T(z-x)dt = f(x) + (1-\alpha)\displaystyle\int_0^1 \nabla f(x + t(z-x))^T(y-x)dt$$

  $$f(z) = f(y) + \textstyle\int_0^1 \nabla f(y + t(z-y))^T(z-y)dt = f(y) - \alpha\displaystyle\int_0^1 \nabla f(x + t(z-x))^T(y-x)dt$$

  If we multiply this first integral equation by α and the second by $(1-\alpha)$ and add them together, we end up with

  $$\alpha f(x) + (1-\alpha)f(y) - f(z) = \alpha(1-\alpha)\int_0^1 [\nabla f(x + t(z-x)) - \nabla f(y + t(z-y))](x-y)dt$$

  From equation (b) above, we further have

  $$\alpha f(x) + (1-\alpha)f(y) - f(z) \geq \mu\alpha(1-\alpha)||x-y||^2 \int_0^1 (1-t)dt = \frac{1}{2}\mu\alpha(1-\alpha)||x-y||^2$$

  Hence, we have shown that f is μ-strongly convex.

■

Proposition 9.3 *Assume f is μ-strongly convex and differentiable. Let $x_* \in argmin_{x\in\mathbb{R}^n} f(x)$, the following must then be true for any $x \in \mathbb{R}^n$:*

(i) $f(x) - f(x_*) \geq \frac{\mu}{2}||x - x_*||_2^2$

(ii) $f(x) - f(x_*) \leq \frac{1}{2\mu}||\nabla f(x)||_2^2$

Proof:

(i) Following from equation (c) of Proposition 9.2, we have

$$f(x) \geq f(x_*) + \nabla f(x_*)^T(x - x*) + \frac{\mu}{2}||x - x_*||_2^2$$

Since x_* is the minimizer of $f(x)$, this means that $\nabla f(x_*) = 0$. Hence,

$$f(x) \geq f(x_*) + \frac{\mu}{2}||x - x_*||_2^2$$

(ii) First, we start with equation (c) from Proposition 9.2.

$$f(x_*) \geq f(x) + \nabla f(x)^T(x_* - x) + \frac{\mu}{2}||x - x_*||_2^2$$

We can then re-arrange this equation so that it has a similar form to equation (ii) in Proposition 9.3

$$f(x) - f(x_*) \leq \nabla f(x)^T(x - x_*) - \frac{\mu}{2}||x - x_*||_2^2$$

Therefore,

$$f(x) - f(x_*) \leq ||\nabla f(x)||_2 \cdot ||x - x_*||_2 - \frac{\mu}{2}||x - x_*||_2^2 \leq \max_{\alpha>0} ||\nabla f(x)||_2 \cdot \alpha - \frac{\mu}{2}\alpha^2 = \frac{1}{2\mu}||\nabla f(x)||_2^2$$

Thus, proving equation (ii) in Proposition 9.3.

■

9.3 Gradient Descent for Strongly Convex Problems

In the following, we consider unconstrained optimization problem

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

where f is L-smooth and μ-strongly convex. We try to analyze the convergence of the gradient descent

$$x_{t+1} = x_t - \gamma\nabla f(x_t), t = 0, 1, \ldots$$

under two different choices of stepsize.

9.3.1 Stepsize $\gamma = 1/L$

Theorem 9.4 *Let f be L-smooth and μ-strongly convex, with step size $\gamma = \frac{1}{L}$, Gradient Descent satisfies*

$$f(x_t) - f(x_*) \leq (1 - \frac{\mu}{L})^t [f(x_0) - f(x_*)]$$
$$||x_t - x_*||_2^2 \leq \frac{2}{\mu}(1 - \frac{\mu}{L})^t [f(x_0) - f(x_*)]$$

Remark: This implies a linear convergence of the Gradient Descent method for solving strongly convex problems. The term $\frac{L}{\mu}$, usually denoted as κ, is known as the condition number. The smaller the condition number is, the faster the convergence will be.

Proof: We know that the following is true due to properties of L-smoothness (which is responsible for the right-most term) and μ-strong convexity (which gives the left-most term):

$$2\mu[f(x_t) - f_*] \leq ||\nabla f(x_t)||_2^2 \leq 2L[f(x_t) - f(x_{t+1})]$$

Let $\epsilon_t = f(x_t) - f(x_*)$, then we have $2\mu\epsilon_t \leq 2L \cdot (\epsilon_t - \epsilon_{t+1})$.

Re-arranging this equation leads to

$$\epsilon_{t+1} \leq (1 - \frac{\mu}{L})\epsilon_t$$

Hence, by induction, we have

$$\epsilon_t \leq (1 - \frac{\mu}{L})^t \epsilon_0, \forall t \geq 1.$$

Invoking the fact that $f(x) - f(x_*) \geq \frac{\mu}{2}\|x - x_*\|_2^2$, we further have

$$||x_t - x_*||_2^2 \leq \frac{2}{\mu}(1 - \frac{\mu}{L})^t [f(x_0) - f(x_*)].$$

■

9.3.2 Stepsize $\gamma = 2/(\mu + L)$

Proposition 9.5 *If f is μ-strongly convex and L-smooth, then the following is true:*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu}{\mu + L}||x - y||^2 + \frac{1}{\mu + L}||\nabla f(x) - \nabla f(y)||^2$$

Proof: Let us first consider the case when $\mu = L$, from the μ-strongly convexity and L-smoothness of f, we have respectively,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu||x - y||_2^2$$
$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{L}||\nabla f(x) - \nabla f(y)||_2^2$$

Adding these two equations together leads to

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle > \frac{\mu}{2}||x - y||^2 + \frac{1}{2L}||\nabla f(x) - \nabla f(y)||^2$$

as expected.

Now, let us explore the case of $\mu < L$. Define the function $\phi(x) = f(x) - \frac{\mu}{2}||x||_2^2$. So $\phi(x)$ is convex and $(L-\mu)$-smooth, which is equivalent to

$$\langle \nabla\phi(x) - \nabla\phi(y), x - y\rangle \geq \frac{1}{L-\mu}||\nabla\phi(x) - \nabla\phi(y)||_2^2$$

Plugging in the definition of $\phi(x)$ gives

$$\langle [\nabla f(x) - \nabla f(y)] - \mu[x-y], x - y\rangle \geq \frac{1}{L-\mu}||[\nabla f(x) - \nabla f(y)] - \mu[x-y]||_2^2$$

Expanding the norms and rearranging the terms will then lead to the conclusion.

■

Theorem 9.6 *Let f be L-smooth and μ-strongly convex, with step size $\gamma = \frac{2}{\mu+L}$, Gradient Descent satisfies*

$$f(x_t) - f(x_*) \leq \frac{L}{2}\left(\frac{\kappa-1}{\kappa+1}\right)^{2t}||x_0 - x_*||^2$$
$$||x_t - x_*||_2^2 \leq \left(\frac{\kappa-1}{\kappa+1}\right)^{2t}||x_0 - x_*||^2$$

As stated before, $\kappa = \frac{L}{\mu}$.

Proof: The proof of the above proposition is straightforward:

$$\begin{aligned}
||x_{t+1} - x_*||_2^2 &= ||x_t - \frac{2}{\mu+L}\nabla f(x_t) - x_*||_2^2 \\
&= ||x_t - x_*||^2 - \frac{4}{\mu+L}\langle \nabla f(x_t), x_t - x_*\rangle + \frac{4||\nabla f(x_t)||_2^2}{(\mu+L)^2} \\
&\leq ||x_t - x_*||^2 - \frac{4}{\mu+L}\left[\frac{\mu L}{\mu+L}||x_t - x_*||^2 + \frac{1}{\mu+L}||\nabla f(x_t)||_2^2\right] + \frac{4||\nabla f(x_t)||_2^2}{(\mu+L)^2} \\
&\leq \left[1 - \frac{4\mu L}{(\mu+L)^2}\right]||x_t - x_*||_2^2 \\
&\leq \left(\frac{\mu-L}{\mu+L}\right)^2||x_t - x_*||_2^2 \\
&= \left(\frac{\kappa-1}{\kappa+1}\right)^2||x_t - x_*||_2^2
\end{aligned}$$

By induction, we have

$$||x_t - x_*||_2^2 \leq \left(\frac{\kappa-1}{\kappa+1}\right)^{2t}||x_0 - x_*||^2.$$

Furthermore, by L-smoothness, we have $f(x_t) - f(x_*) \leq \frac{L}{2}\|x_t - x_*\|_2^2$, therefore,

$$f(x_t) - f(x_*) \leq \frac{L}{2}\left(\frac{\kappa-1}{\kappa+1}\right)^{2t}||x_0 - x_*||^2.$$

■

Discussion. In the subsections above, we have seen that the Gradient Descent exhibit slightly different linear convergences under two different choices of step-sizes:

Before: $\gamma = \frac{1}{L}$ which gave a rate of order $\left(1 - \frac{\mu}{L}\right)^t = \left(1 - \frac{1}{\kappa}\right)^t \leq e^{-t/\kappa}$.

Now: $\gamma = \frac{2}{\mu+L}$ which gives a rate of order $\left(1 - \frac{2}{\kappa+1}\right)^{2t} \leq e^{-4t/(\kappa+1)}$, which is better than $e^{-t/\kappa}$.

9.4 Accelerated Gradient Descent

9.4.1 Overview

Question: is Gradient Descent the best algorithm to solve smooth convex optimization problems? Is it possible to obtain a even faster rate of convergence?
It was shown in Nesterov's seminal work in 1983 that the answer is yes, and a better (indeed optimal) convergence rate can be achieved through Accelerated Gradient Descent algorithm. Here is a brief summary.

Problem type	GD Convergence Rate	AGD Convergence Rate
f is L-smooth and convex	$\mathcal{O}\left(\frac{LD^2}{t}\right)$	$\mathcal{O}\left(\frac{LD^2}{t^2}\right)$
f is L-smooth and μ-strongly convex	$\mathcal{O}\left(\left(\frac{\kappa-1}{\kappa+1}\right)^{2t}\right)$	$\mathcal{O}\left(\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2t}\right)$

9.4.2 Historical Note of Accelerated Gradient Descent

The following will detail the history and evolution of the study of accelerated gradient descent:

- [YN83] In 1983, Nesterov created the first accelerated gradient descent scheme for smooth functions.
- [YN88] In 1988, Nesterov published another, more general acceleration scheme for smooth functions.
- [YN04] In 2004, Nesterov combined the acceleration scheme with smoothing techniques to address non-smooth problems.
- [YN07] In 2007, Nesterov further extended the accelerated gradient descent to solve composite problems (problems that have both smooth and non-smooth parts).
- [BT08] In 2008, Beck and Teboulle established another simple and popular acceleration algorithm, FISTA (Fast Iterative Shrinkage-Thresholding Algorithm), designed to solve composite problems.
- [TY09] In 2009, Tseng performed unified analysis of both smooth and non-smooth problems under accelerated gradient descent.
- Since then, this leads to an interesting "acceleration phenomenon" and the acceleration idea has been applied to many other first-order algorithms,

Despite of its significance, the mechanism behind this algorithm has long been conceived as pure "algebraic trick" and remains mysterious. Many interesting interpretations have been recently developed from different viewpoints, see more details from the following references.

- **ODE perspective:**

  - [SB15] Su, Boyd, and Candes, 2015."A differential equation for modeling Nesterovs accelerated gradient method: theory and insights."
  - [LR14] Lessard, Racht, Packward, 2014."Analysis and Design of Optimization Algorithms via Integral Quadratic Constraints."
  - [WW16] Wibisono, Wilson, and Jordan, 2016. "A Variational Perspective on Accelerated Methods in Optimization."

- **Geometry perspective:**
  - [AZ14] Allen-Zhu and Orrchia, 2014. "Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent."
  - [BL15] Bubeck, Lee and Singh, 2015. "A geometric alternative to Nesterovs accelerated gradient descent."

- **Game theory perspective:**
  - [LZ15] Lan and Zhou, 2015. "An optimal randomized incremental gradient method."

9.4.3 Accelerated Gradient Descent

We will look at one of Nesterov's Optimal Gradient Scheme for the unconstrained smooth convex optimization

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

where f is L-smooth and μ-strongly convex ($\mu \geq 0$).

The Accelerated Gradient Descent works as follows: let $y_0 = x_0$, for $t = 0, 1, 2, ...$, update

$$x_{t+1} = y_t - \frac{1}{L}\nabla f(y_t)$$
$$y_{t+1} = x_{t+1} + \beta_t(x_{t+1} - x_t), \text{with } \beta_t = \frac{\alpha_t(1-\alpha_t)}{\alpha_t^2 + \alpha_{t+1}}$$

where α_{t+1}^2 is defined in the following way:

$$\alpha_{t+1}^2 = (1-\alpha_{t+1})\alpha_t^2 + \frac{\mu}{L}\alpha_{t+1}$$

The above scheme is indeed very general and flexible. We list a few simple variants below under particular choices of parameters.

- **Heavy-ball method:** This method is named the heavy-ball method as it has a momentum term in the x_{t+1} term ($\beta_t(x_t - x_{t-1})$), which allows for the equation to imitate the motion of a heavy ball rolling down a hill.
$$x_{t+1} = x_t - \alpha_t\nabla f(x_t) + \beta_t(x_t - x_{t-1})$$
where α_t and β_t are set to $\alpha_t = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}, \beta_t = \frac{(\sqrt{L}-\sqrt{\mu})^2}{(\sqrt{L}+\sqrt{\mu})^2}$.

- **Nesterov'83:-**
$$\begin{cases} x_{t+1} = y_t - \frac{1}{L}\nabla f(y_t) \\ y_{t+1} = x_{t+1} + \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}(x_{t+1} - x_t) \end{cases}$$

- **FISTA:**

$$\begin{cases} x_{t+1} = y_t - \frac{1}{L}\nabla f(y_t) \\ y_{t+1} = x_{t+1} + \frac{\lambda_t - 1}{\lambda_{t+1}}(x_{t+1} - x_t) \end{cases}$$

  where λ is defined in the following way:

$$\lambda_0 = 0, \lambda_{t+1} = \frac{1 + \sqrt{1 + 4\lambda_t^2}}{2}, \forall t \geq 0$$

- **Nesterov'88:**

$$\begin{cases} x_{t+1} = y_t - \frac{1}{L}\nabla f(y_t) \\ y_{t+1} = x_{t+1} + \frac{t}{t+3}(x_{t+1} - x_t) \end{cases}$$

The first two variants are widely used for strongly convex problems and the latter two are used for general smooth convex problems. In the following, we are going to analyze the convergence rate for the FISTA algorithm.

Theorem 9.7 *Let f be L-smooth and convex, then the solution x_t from the FISTA algorithm satisfies:*

$$f(x_t) - f(x_*) \leq \frac{2L \cdot ||x_0 - x_*||_2^2}{t^2}$$

Proof: We start with the following fact:

$$f(x_{t+1}) - f(x) = f(x_{t+1}) - f(y_t) + f(y_t) - f(x)$$

Using the facts that $f(x_{t+1}) - f(y_t) \leq -\frac{1}{2L}\|\nabla f(y_t)\|_2^2$ and $f(y_t) - f(x) \geq \nabla f(y_t)^T(y_t - x)$, we have

$$f(x_{t+1}) - f(x) \leq -\frac{1}{2L}||\nabla f(y_t)||_2^2 + \nabla f(y_t)^T(y_t - x)$$

Hence, by definition of $x_{t+1} = y_t - \frac{1}{L}\nabla f(y_t)$, we have

$$f(x_{t+1}) - f(x) \leq -\frac{L}{2}||x_{t+1} - y_t||_2^2 - L(x_{t+1} - y_t)^T(y_t - x)$$

Let $x = x_t$ and $x = x_*$, respectively, and multiply the resulting equation with two separate factors:

$$[f(x_{t+1}) - f(x_t)](\lambda_t - 1) \leq \left[-\frac{L}{2}||x_{t+1} - y_t||_2^2 - L(x_{t+1} - y_t)^T(y_t - x_t)\right](\lambda_t - 1)$$

$$[f(x_{t+1}) - f(x_*)] \leq \left[-\frac{L}{2}||x_{t+1} - y_t||_2^2 - L(x_{t+1} - y_t)^T(y_t - x_*)\right]$$

Note that $\lambda_t \geq 1, \forall t \geq 1$. Adding these two equations together, we arrive at

$$\lambda_t f(x_{t+1}) - (\lambda_t - 1)f(x_t) - f(x_*) \leq -\frac{\lambda_t L}{2}||x_{t+1} - y_t||_2^2 - L(x_{t+1} - y_t)^T(\lambda_t y_t + (\lambda_t - 1)x_t - x_*)$$

Let $\epsilon_t = f(x_t) - f(x_*)$, this leads to :

$$\lambda_t \epsilon_{t+1} - (\lambda_t - 1)\epsilon_t \leq -\frac{\lambda_t L}{2}||x_{t+1} - y_t||_2^2 - L(x_{t+1} - y_t)^T(\lambda_t y_t + (\lambda_t - 1)x_t - x_*)$$

Multiplying both sides by λ_t:

$$\lambda_t^2\epsilon_{t+1} - \lambda_t(\lambda_t - 1)\epsilon_t \leq -\frac{L}{2}\left[\lambda_t^2||x_{t+1} - y_t||_2^2 + 2\lambda_t L(x_{t+1} - y_t)^T(\lambda_t y_t + (\lambda_t - 1)x_t - x_*)\right]$$

By definition of λ_t, we know the following is true,

$$\lambda_{t+1}^2 - \lambda_{t+1} = \lambda_t^2$$

Rearranging terms, we have

$$\lambda_t^2\epsilon_{t+1} - \lambda_{t-1}^2\epsilon_t \leq -\frac{L}{2}(||\lambda_t x_{t+1} - (\lambda_t - 1)x_t - x_*||^2 - ||\lambda_t y_t - (\lambda_t - 1)x_t - x_*||^2)$$

Invoking the definition of y_{t+1}, one can show that

$$\lambda_t x_{t+1} - (\lambda_t - 1)x_t - x_* = \lambda_{t+1}y_{t+1} - (\lambda_{t+1} - 1)x_{t+1} - x_*$$

Hence, defining $u_0 = x_0, u_t = \lambda_t y_t - (\lambda_t - 1)x_t - x_*, \forall t \geq 1$, the above equation simplifies to

$$\lambda_t^2\epsilon_{t+1} - \lambda_{t-1}^2\epsilon_t \leq -\frac{L}{2}(||u_{t+1}||^2 - ||u_t||^2)$$

Therefore, by induction, we have

$$\lambda_{t-1}^2\epsilon_t \leq \frac{L}{2}||u_0||^2$$

From this, we realize the following:

$$\epsilon_t \leq \frac{L||x_0||^2}{2\lambda_{t-1}^2}$$

By simple induction, we can show that

$$\lambda_{t-1} \geq \frac{t}{2}, \forall t \geq 1$$

Therefore:

$$\epsilon_t \leq \frac{2L||x_0||^2}{t^2}$$

■

References

[AZ14] Z. Allen-Zhu and L. Orecchia, "Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent", *ArXiv*, 2014, arXiv:1407.1537v4.

[BT08] A. Beck, and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems", *SIAM journal on imaging sciences*, 2008, vol. 2, num1, pp. 183-202.

[BL15] S. Bubeck, Y. T. Lee, and M. Singh, "A geometric alternative to Nesterov's accelerated gradient descent", *ArXiv*, 2015, arXiv:1506:08187v1.

[LZ15] G. Lan and Y. Zhou, "An optimal randomized incremental gradient method", *ArXiv*, 2015, arXiv:1507.02000v3.

[LR14] L. Lessard, B. Recht, and A. Packard, "Analysis and Design of Optimization Algorithms via Integral Quadratic Constraints", *Arxiv*, 2014, arXiv: 1408.3595v7.

[YN83] Y. Nesterov, "A method of solving a convex programming problem with convergence rate O (1/k2)," *Soviet Mathematics Doklady*, 1983, vol. 27 num. 2, pp. 372–376.

[YN88] Y. Nesterov and A. Nemirovsky "A general approach to polynomial-time algorithms design for convex programming," *Report, Central Economical and Mathematical Institute, USSR Academy of Sciences, Moscow*, 1988.

[YNBK] Y. Nesterov, "Introductory Lectures on Convex Optimization: A Basic Course", *Kluwer-Academic*, 2003.

[YN04] Y. Nesterov, "Introductory Lectures on Convex Optimization. Applied Optimization" , *Kluwer Academic Publishers, Boston*, 2004, vol. 87.

[YN07] Y. Nesterov, "Gradient methods for minimizing composite objective function", *UCL*, 2007.

[SB15] W. Su, S. Boyd, and E.J. Candes , "A differential equation for modeling Nesterovs accelerated gradient method: theory and insights", *ArXiv*, 2015, arXiv:1503.01243.

[TY09] P. Tseng and S. Yun "A coordinate gradient descent method for nonsmooth separable minimization", *Mathematical Programming*, 2009, vol. 117, pp. 387-423.

[WW16] A. Wibisono, A. Wilson, and M. Jordan, "A Variational Perspective on Accelerated Methods in Optimization", *ArXiv*, 2016, arXiv:1603.04245.