Analysis of Smoothing Parameter Scheduling for Non-Smooth Convex Optimization

수리과학부 2022-15173 멍흐첼맥

December 16, 2025

Abstract

First-order methods for non-smooth convex optimization, such as subgradient methods, suffer from slow convergence rates of $\mathcal{O}(1/\sqrt{k})$. Nesterov's smoothing technique offers an improved rate of $\mathcal{O}(1/k)$ by approximating the non-smooth objective function with a smoothing function involving a smoothing parameter mu. However, the standard approach relies on a fixed mu, which entails a trade-off between approximation accuracy and convergence speed. In this paper, we propose and analyze the performance of Nesterov's method when the smoothing parameter μ is scheduled to decrease adaptively during iterations. We apply this approach to the Sparse Signa Recovery problem (Lasso) and demonstrate through numerical experiments that an appropriate decay schedule of μ achieves both faster convergence and higher solution precision compared to the fixed parameter approach.

1 서론

Lasso 회귀, 서포트 벡터 머신(SVM)등 많은 머신러닝 문제는 비매끄러운(non-smooth) 볼록 최적화 문제로 귀결된다. 이러한 문제를 해결하기 위한 가장 직관적이고 일반적인 방법은 서브그레디언트 방법(subgradient method)이다. 그러나 서브그레디언트 방법은 이론적으로 $\mathcal{O}(1/\sqrt{k})$의 느린 수렴률을 가지며, 높은 정확도의 해를 얻기 위해 많은 반복 횟수를 요구한다는 단점이 있다. 이는 대규모 문제에서 계산 효율성 측면에서 심각한 제약으로 작용한다.

이러한 한계를 극복하기 위해 Nesterov [1]는 비매끄러운 볼록함수를 매끄러운 함수로 근사하는 smoothing 기법을 제안했다. 이 방법은 목적함수 $f(x)$를 smoothing 파라미터 μ에 의해 Lipschitz 연속인 그레디언트를 가지는 함수 $f_\mu(x)$로 근사함으로써,

가속 경사 하강법(Accelerated Gradient Descent)을 적용할 수 있게 된다. 그 결과, $\mathcal{O}(1/k^2)$의 빠른 수렴률을 보장할 수 있다.

기존 연구에서는 목표 정확도 ε에 따라 고정된 smoothing 파라미터 μ를 설정하는 방식을 사용한다. 그러니 이러한 고정 파라미터 방식은 근사 차와 최적화 속도 사이의 trade-off를 내포하고 있다. 즉, μ를 크게 설정하면 빠른 수렴을 얻을 수 있으나 근사 오차가 커지고, 반대로 μ를 작게 설정하면 정확도는 향상되지만 최적화 과정이 느려진다.

본 보고서에서는 이러한 한계를 극복하기 위해 iteration k에 따라 smoothing 파라미터를 $\mu_k = 1/k^\alpha$의 형태로 감소시키는 scheduling 기법을 제안한다. 제안한 방법에 대해 가속 경사 하강법 하에서의 이론적 수렴률을 분석하고, 적절한 α의 선택을 통해 전체 수렴률이 $\mathcal{O}(1/k)$임을 보인다. 또한, 실제 문제에 적용시켜 검증한다.

2 배경 이론

정의역 $\mathcal{X}$가 유계 닫힌 볼록 집합인, 다음과 같이 표현되는 함수 볼록함수 $f(x)$를 고려하자.

$$f(x) := \max_{u \in \mathcal{U}} \{\langle Ax, u\rangle - \phi(u)\} \tag{1}$$

($\mathcal{U}$는 유계 닫힌 볼록 집합, ϕ는 볼록함수이다.) Nesterov는 다음과 같은 근사 f_μ를 제시한다.

$$f_\mu(x) := \max_{u \in \mathcal{U}} \{\langle Ax, u\rangle - \phi(u) - \mu d(u)\} \tag{2}$$

여기서 함수 $d(u)$는 σ-strongly 볼록함수이다. 그러면 다음이 성립한다.

Lemma 2.1. 함수 $f_\mu(x)$는 다음을 만족하는 매끄러운 볼록함수이며, $u_\mu(x)$를 (2)의 최적해라고 할 때 다음이 성립한다.

$$\nabla f_\mu(x) = A^\top u_\mu(x) \tag{3}$$

또한 ∇f는 *Lipchitz* 연속이며, 파라미터는 다음과 같이 구해진다.

$$L_\mu = \frac{\|A\|^2}{\mu\sigma} \tag{4}$$

Lemma 2.2. $0 \leq \mu' \leq \mu$이고, 함수 d가 *bounded*일 때, 즉 임의의 u에 대해 $d(u) \leq D^2$를 만족시킬 때, 다음이 성립한다.

$$f_\mu(x) \leq f_{\mu'}(x) \leq f_\mu(x) + (\mu - \mu')D^2 \tag{5}$$

3 알고리즘 개요

본 보고서에서 제안하는 방법은 각 반복 k에 대해 smoothing 파라미터가 $\mu_k = 1/k^\alpha$와 같이 변화하며, (단, $0 \leq \alpha < 2$) 가속 경사 하강법을 적용시킬 때, 함수 $f_{\mu_k}(x)$에 대한 기울기를 계산하여 업데이트할 것이다. 즉,

$$\begin{aligned} x_{k+1} &= y_k - \frac{1}{L_k}\nabla f_{\mu_k}(y_k) \\ y_{k+1} &= (1-\gamma_k)x_{k+1} + \gamma_k x_k \end{aligned} \tag{6}$$

여기서 γ_k는 λ_k에 의해 다음과 같이 정의된다:

$$\gamma_k = \frac{1-\lambda_k}{\lambda_{k+1}}, \quad \lambda_{k+1} = \frac{1+\sqrt{1+4\lambda_k^2}}{2} \tag{7}$$

4 수렴률 분석

앞으로는 편의성을 위해 $f_{\mu_k}(x)$, L_{μ_k}를 각각 $f_k(x)$, L_k로 쓴다.

우선 λ_k에 대한 간단한 성질을 소개한다.

Lemma 4.1. (7)에서 정의된 수열 $\{\lambda_k\}_{k=0}^\infty$는

$$\frac{k}{2} < \lambda_k < k$$

이며, 다음 점화식을 만족시킨다.

$$\lambda_{k+1}(\lambda_{k+1} - 1) = \lambda_k^2 \tag{8}$$

이제 증명에 필요한 기초적인 부등식들을 유도한다.

Lemma 4.2. 업데이트 규칙 (6)를 따를 때, 다음이 식이 성립한다.

$$f_k(x_{k+1}) \leq f_k(y_k) - \frac{1}{2L_k} \|\nabla f_k(y_k)\|^2 \tag{9}$$

Proof. f_k의 Lipschitz 조건에 의해

$$\begin{aligned} f_k(x_{k+1}) &\leq f_k(y_k) + \langle \nabla f_k(y_k), x_{k+1} - y_k \rangle + \frac{L_k}{2} \|x_{k+1} - y_k\|^2 \\ &= f_k(y_k) - \frac{1}{L_k} \|\nabla f_k(y_k)\|^2 + \frac{1}{2L_k} \|\nabla f_k(y_k)\|^2 \\ &= f_k(y_k) - \frac{1}{2L_k} \|\nabla f_k(y_k)\|^2 \end{aligned} \tag{10}$$

이 되어 증명된다. □

Lemma 4.3. 업데이트 규칙 (6)를 따를 때, 임의의 $x \in \mathcal{X}$에 대해 다음 식이 성립한다.

$$f_k(x_{k+1}) \leq f_k(x) + \frac{L_k}{2} \left(\|y_k - x\|^2 - \|x_{k+1} - x\|^2 \right) \tag{11}$$

Proof. Lemma 4.2와 함수 f_k의 볼록성에 의해

$$\begin{aligned} f_k(x_{k+1}) &\leq f_k(y_k) - \frac{1}{2L_k} \|\nabla f_k(y_k)\|^2 \\ &\leq f_k(x) + \langle \nabla f_k(y_k), y_k - x \rangle - \frac{1}{2L_k} \|\nabla f_k(y_k)\|^2 \\ &= f_k(x) - \frac{L_k}{2} \left(\frac{1}{L_k^2} \|\nabla f_k(y_k)\|^2 - \frac{2}{L_k} \langle \nabla f_k(y_k), y_k - x \rangle \right) \\ &= f_k(x) - \frac{L_k}{2} \left(\left\| \frac{1}{L_k} \nabla f_k(y_k) - y_k + x \right\|^2 - \|y_k - x\|^2 \right) \\ &= f_k(x) + \frac{L_k}{2} \left(\|y_k - x\|^2 - \|x_{k+1} - x\|^2 \right) \end{aligned} \tag{12}$$

이 되어 증명된다. □

이제 f의 최적값을 x^*라고 할 때, 새로운 수열 $\{v_k\}_{k=1}^{\infty}$를

$$v_k := \lambda_{k-1} x_k + (1 - \lambda_{k-1}) x_{k-1} - x^* \tag{13}$$

으로 정의하면, 다음을 증명할 수 있다.

Lemma 4.4. 업데이트 규칙 (6)를 따를 때, f의 최적값을 x^*라고 하면 다음이 식이 성립한다.

$$\lambda_k^2(f_k(x_{k+1}) - f_k(x^*)) - \lambda_{k-1}^2(f_k(x_k) - f_k(x^*)) \leq \frac{L_k}{2}\left(\|v_k\|^2 - \|v_{k+1}\|^2\right) \tag{14}$$

Proof. Lemma 4.3의 양변에 λ_k^2을 곱하면

$$\lambda_k^2(f_k(x_{k+1}) - f_k(x)) \leq \frac{L_k}{2}\left(\|\lambda_k(y_k - x)\|^2 - \|\lambda_k(x_{k+1} - x)\|^2\right) \tag{15}$$

이 되고, 여기서

$$x = \left(1 - \frac{1}{\lambda_k}\right)x_k + \frac{1}{\lambda_k}x^* \tag{16}$$

를 대입하자. 그러면 좌변은 f_k의 볼록성과 Lemma 4.1 의해

$$\begin{aligned}
\lambda_k^2(f_k(x_{k+1}) - f_k(x)) &\geq \lambda_k^2\left(f_k(x_{k+1}) - \left(1 - \frac{1}{\lambda_k}\right)f_k(x_k) - \frac{1}{\lambda_k}f_k(x^*)\right) \\
&= \lambda_k^2 f_k(x_{k+1}) - \lambda_k(\lambda_k - 1)f_k(x_k) - \lambda_k f_k(x^*) \\
&= \lambda_k^2 f_k(x_{k+1}) - \lambda_{k-1}^2 f_k(x_k) - (\lambda_k^2 - \lambda_{k-1}^2)f_k(x^*) \\
&= \lambda_k^2(f_k(x_{k+1}) - f_k(x^*)) - \lambda_{k-1}^2(f_k(x_k) - f_k(x^*))
\end{aligned} \tag{17}$$

이고, 다른 한편 우변은

$$\begin{aligned}
RHS &= \frac{L_k}{2}\left(\|\lambda_k(y_k - x)\|^2 - \|\lambda_k(x_{k+1} - x)\|^2\right) \\
&= \frac{L_k}{2}\left(\|\lambda_k y_k - (\lambda_k - 1)x_k - x^*\|^2 - \|\lambda_k x_{k+1} + (1 - \lambda_k)x_k - x^*\|^2\right) \\
&= \frac{L_k}{2}\left(\|\lambda_k((1 - \gamma_k)x_k + \gamma_k x_{k-1}) - (\lambda_k - 1)x_k - x^*\|^2 - \|v_{k+1}\|^2\right) \\
&= \frac{L_k}{2}\left(\|\lambda_{k-1}x_k + (1 - \lambda_{k-1})x_{k-1} - x^*\|^2 - \|v_{k+1}\|^2\right) \\
&= \frac{L_k}{2}\left(\|v_k\|^2 - \|v_{k+1}\|^2\right)
\end{aligned} \tag{18}$$

이므로 우리가 원하는 관계식이 성립한다. □

Lemma 4.5. 업데이트 규칙 (6)를 따를 때, 어떤 상수 $R > 0$이 존재하여 다음 부등

식을 만족시킨다.

$$\lambda_T^2(f_T(x_{T+1}) - f_T(x^*)) \leq 2R^2 L_T + D^2 \sum_{k=1}^{T-1} \lambda_k^2(\mu_k - \mu_{k+1}) + \mathcal{O}(1) \tag{19}$$

Proof. Lemma 4.4를 $k = 1$부터 T까지 전부 다 더한 뒤 $\mu_k \geq \mu_{k+1}$라는 사실로부터 Lemma 2.2를 적용시켜

$$\begin{aligned}
LHS &= \sum_{k=1}^{T} \left(\lambda_k^2(f_k(x_{k+1}) - f_k(x^*)) - \lambda_{k-1}^2(f_k(x_k) - f_k(x^*))\right) \\
&= \lambda_T^2(f_T(x_{T+1}) - f_T(x^*)) + \sum_{k=1}^{T-1} \lambda_k^2 \left(f_k(x_{k+1}) - f_{k+1}(x_{k+1}) + f_{k+1}(x^*) - f_k(x^*)\right) \\
&\geq \lambda_T^2(f_T(x_{T+1}) - f_T(x^*)) + \sum_{k=1}^{T-1} \lambda_k^2(f_k(x_{k+1}) - f_{k+1}(x_{k+1})) \\
&\geq \lambda_T^2(f_T(x_{T+1}) - f_T(x^*)) + \sum_{k=1}^{T-1} \lambda_k^2 D^2(\mu_{k+1} - \mu_k)
\end{aligned} \tag{20}$$

이고, 한편, $\mathcal{X}$는 유계이므로 $\|x\| \leq R$을 만족하는 상수 $R > 0$이 존재한다. 따라서 우변은

$$\begin{aligned}
RHS &= \sum_{k=1}^{T} \frac{L_k}{2} \left(\|v_k\|^2 - \|v_{k+1}\|^2\right) \\
&\leq -\frac{L_T}{2} \|v_T\|^2 + \sum_{k=1}^{T-1} \frac{\|v_k\|^2}{2}(L_{k+1} - L_k) + \mathcal{O}(1) \\
&\leq \sum_{k=1}^{T-1} \frac{\lambda_{k-1} \|x_k - x^*\|^2 + (1 - \lambda_{k-1}) \|x_{k-1} - x^*\|^2}{2}(L_{k+1} - L_k) + \mathcal{O}(1) \\
&\leq \sum_{k=1}^{T-1} \frac{4\lambda_{k-1}R^2 + 4(1 - \lambda_{k-1})R^2}{2}(L_{k+1} - L_k) + \mathcal{O}(1) \\
&\leq 2R^2 \sum_{k=1}^{T-1}(L_{k+1} - L_k) + \mathcal{O}(1) \\
&= 2R^2 L_T + \mathcal{O}(1)
\end{aligned} \tag{21}$$

이 되어 증명된다. □

최종적으로, 수렴률이 다음과 같음을 증명할 수 있다.

Theorem 4.6. 업데이트 규칙 (6)를 따를 때, 다음 식이 성립한다.

$$f(x_{T+1}) - f(x^*) \leq D^2 T^{-\alpha} + \frac{8R^2 \|A\|^2}{\sigma} T^{\alpha-2} + \frac{4D^2\alpha}{2-\alpha} T^{-\alpha} + \mathcal{O}(T^{-2}) \tag{22}$$

Proof. Lemma 2.2와 Lemma 4.5를 적용한 후, Lemma 4.1와 평균값 정리를 적용하고, 마지막으로 $y = x^{1-\alpha}$의 넓이를 비교하면

$$\begin{aligned}
f(x_{T+1}) - f(x^*) &= (f(x_{T+1}) - f_T(x_{T+1})) + (f_T(x_{T+1}) - f_T(x^*)) + (f_T(x^*) - f(x^*)) \\
&\leq \mu_T D^2 + \frac{2R^2 L_T}{\lambda_T^2} + \frac{D^2}{\lambda_T^2} \sum_{k=1}^{T-1} \lambda_k^2 (\mu_k - \mu_{k+1}) + \mathcal{O}(T^{-2}) \\
&\leq D^2 T^{-\alpha} + \frac{8R^2 \|A\|^2}{\sigma} T^{\alpha-2} + \frac{4D^2}{T^2} \sum_{k=1}^{T-1} k^2 \left(\frac{1}{k^\alpha} - \frac{1}{(k+1)^\alpha} \right) + \mathcal{O}(T^{-2}) \\
&\leq D^2 T^{-\alpha} + \frac{8R^2 \|A\|^2}{\sigma} T^{\alpha-2} + \frac{4D^2}{T^2} \sum_{k=1}^{T-1} k^2 \cdot \alpha k^{-\alpha-1} + \mathcal{O}(T^{-2}) \\
&= D^2 T^{-\alpha} + \frac{8R^2 \|A\|^2}{\sigma} T^{\alpha-2} + \frac{4D^2\alpha}{T^2} \sum_{k=1}^{T-1} k^{1-\alpha} + \mathcal{O}(T^{-2}) \\
&= D^2 T^{-\alpha} + \frac{8R^2 \|A\|^2}{\sigma} T^{\alpha-2} + \frac{4D^2\alpha}{T^2} \sum_{k=2}^{T-1} k^{1-\alpha} + \mathcal{O}(T^{-2}) \\
&\leq D^2 T^{-\alpha} + \frac{8R^2 \|A\|^2}{\sigma} T^{\alpha-2} + \frac{4D^2\alpha}{T^2} \int_1^T x^{1-\alpha}\, dx + \mathcal{O}(T^{-2}) \\
&= D^2 T^{-\alpha} + \frac{8R^2 \|A\|^2}{\sigma} T^{\alpha-2} + \frac{4D^2\alpha}{(2-\alpha)T^2} T^{2-\alpha} + \mathcal{O}(T^{-2}) \\
&= D^2 T^{-\alpha} + \frac{8R^2 \|A\|^2}{\sigma} T^{\alpha-2} + \frac{4D^2\alpha}{(2-\alpha)} T^{-\alpha} + \mathcal{O}(T^{-2})
\end{aligned} \tag{23}$$

가 되어 증명된다. □

Corollary 4.7. *Theorm* 4.6에서 $\alpha = 1$로 선택하면 수렴률이 $\mathcal{O}(1/T)$이 됨을 알 수 있다.

5 응용

5.1 LASSO 회귀

LASSO 회귀 문제는 다음과 같은 형태를 가진다.

$$\underset{x}{\text{minimize}}\ \frac{1}{2}\|Ax-b\|^2+\lambda\|x\|_1 \tag{24}$$

특히, $\|\cdot\|_1$은

$$\|x\|_1=\max_{\|u\|_\infty\le\lambda}\langle x,u\rangle \tag{25}$$

와 같이 표현되므로, 우리의 초기 세팅과 잘 맞는다. 만약 페널티함수를 $d(u)=\|u\|^2/2$로 잡으면 으로 잡으면, $\sigma=1$이고,

$$u_\mu(x)=\underset{\|u\|_\infty\le\lambda}{\text{argmax}}\,\langle x,u\rangle-\frac{\mu}{2}\|u\|^2 \tag{26}$$

를 만족시킨다. 우변 함수는 매끄러운 함수이므로 미분하면

$$u_\mu(x)=\underset{\|\cdot\|_\infty\le\lambda}{\text{proj}}\left(\frac{x}{\mu}\right)\implies (u_\mu(x))_i=\text{sign}(x_i)\min\left(\frac{\|x_i\|}{\mu},\lambda\right) \tag{27}$$

를 얻는다. 여기서 L_1 정규화를 가지는 미분가능한 모든 문제는 이와 같이 풀 수 있음을 알 수 있다.

5.2 서포트 벡터 머신

Bias 항이 없는 서포트 베터 머신은 다음과 같다.

$$\underset{w}{\text{minimize}}\ \sum_{i=1}^{N}\max\{0,1-y_i(w^\intercal x_i)\}+\frac{\lambda}{2}\|x\|^2 \tag{28}$$

여기서 비매끄러운 항인 hinge 손실은 다음과 같이 표현된다.

$$h(w)=\max_{u\in[0,1]^N}\sum_{i=1}^{N}u_i(1-y_iw^\intercal x_i) \tag{29}$$

이제 $d(u) = \|u\|^2/2$로 잡으면, 우리가 풀고자 하는 문제는

$$\max_{u\in[0,1]^N} \sum_{i=1}^{N} u_i - w^\intercal \sum_{i=1}^{N} u_i y_i x_i - \frac{\mu}{2}\|u\|^2 \tag{30}$$

따라서 smoothing된 문제의 해는 우변 함수를 u에 대해 미분하여 얻을 수 있다. 제약조건 $[0,1]^N$을 고려한 해는 다음과 같다.

$$\mathbf{1} - \mathrm{diag}(y)X^\intercal w - \mu u = \mathbf{0} \implies (u_\mu(w))_i = \mathrm{clip}\left(\frac{1 - y_i w^\intercal x_i}{\mu}, 0, 1\right) \tag{31}$$

6 실험

본 섹션에서는 앞서 제안한 가속 경사 하강법의 성능을 검증하기 위해, 대표적인 비매끄러운 최적화 문제의 LASSO 회귀에 적용하여 수치 실험을 수행한다. [1] 특히, smoothing 파라미터 α가 수렴 속도에 미치는 영향을 분석하여 이론적으로 유도된 최적 감쇠율 $\alpha = 1$의 타당성을 보인다.

LASSO 문제의 설정은 다음과 같다. $m = 100$, $n = 50$ 크기의 무작위 행렬 $A \in \mathbb{R}^{m\times n}$과 해 $x_{\text{true}} \in \mathbb{R}^n$를 생성하고, 정규화 계수는 $\lambda = 0.5$로 설정하였다. 알고리즘은 $T_{\max} = 1000$회 반복하였고 목적함수의 최적값 f^*는 CVXPY를 사용하여 계산하였다.

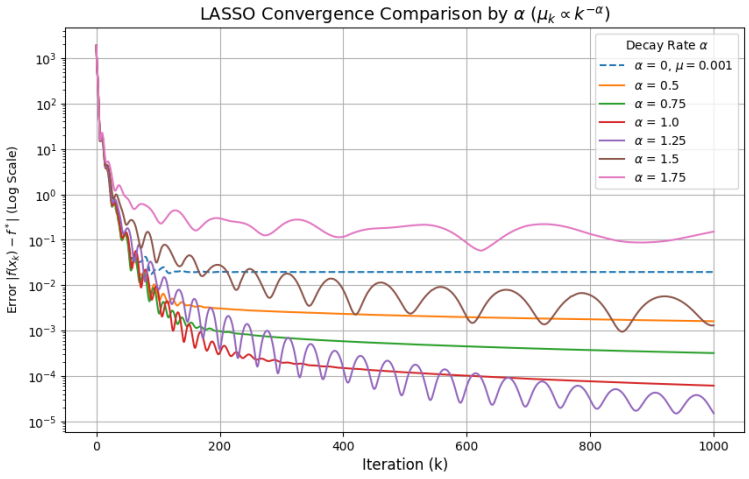

Figure 1: α 값에 따른 LASSO 회귀의 수렴 곡선

$\alpha = 0$인 경우 초기에 빠르게 감소하지만, $k \approx 150$부터 오차가 10^{-2} 수준에서 더

[1] https://github.com/Oioioi-Baka/2025fa_optimziation_theory

이상 감소하지 않고 유지된다. 이는 μ가 상수일 때 발생하는 bias가 반복횟수가 늘어나도 사라지지 않기 때문이다.

반면, $0 < \alpha \leq 1$의 경우 적당히 빠르며 안정적이며, 특히 $\alpha = 1$일 때 오차가 가장 빠르게 감소하여 $k = 1000$에서 약 10^{-4} 수준에 도달한다.

$1 < \alpha \leq 2$의 경우 초기에 심한 진동이 관찰된다. 특히 $\alpha = 1.75$일 때는 오차가 거의 감소하지 않는 모습을 보이는데, 이는 μ_k가 너무 급격히 감소함에 따라 smoothing된 함수의 Lipschitz 상수가 빠르게 증가하여 최적화 과정의 안정성을 해치는 것으로 분석된다.

7 결론

본 보고서에서는 비매끄러운 볼록최적화 문제를 효율적으로 해결하기 위해 Nesterov의 smoothing 기법에 scheduling 기법을 적용시켜 제안하였고, 그 수렴성을 분석하였다. 기존의 고정된 smoothng 파라미터 방식은 bias와 수렴 속도 간의 trade-off로 인해 정밀한 해를 얻는 데 한계가 있었다. 이를 극복하기 위해 본 연구에서는 반복횟수에 따라 smoothing 파라미터를 $\mu_k = 1/k^\alpha$의 형태로 감소시키는 기법을 도입하였다.

이론적 분석을 통해, 제안된 알고리즘의 오차는 $\mathcal{O}(T^{-\alpha}) + \mathcal{O}(T^{\alpha-2})$임을 보였다. 이를 통해 두 오차 항이 균형을 이루는 최적의 α 값이 1임을 보였으며, 이때 알고리즘은 $\mathcal{O}(1/T)$의 수렴률을 달성한다.

References

[1] Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1), 127-152.

[2] Y. Nesterov, "A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$," *Doklady Akademii Nauk SSSR,* vol. 269, no. 3, pp. 543–547, 1983.

[3] Boyd, S., & Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.