Efficient Distance-Adaptive Subgradient Methods

Anton Rodomanov (CISPA, Germany)

20 August 2025
Optimization Seminar at UCLouvain
Louvain-la-Neuve, Belgium

Outline

Basic Normalized Subgradient Method

Basic Normalized Subgradient Method (NSM)

Problem: $\min_{x \in \mathbb{R}^d} f(x)$, where f is a convex function.

NSM:

$$\boxed{x_{k+1} = x_k - h_k \frac{g_k}{\|g_k\|},} \quad g_k = f'(x_k), \qquad k \geq 0,$$

where $\|\cdot\|$ is the standard Euclidean norm, and $h_k > 0$ is an appropriately chosen sequence of step sizes, and $f'(x_k)$ is an arbitrary subgradient.

What Does NSM Actually Do?

Main property of minimizer: $\langle g(x), x - x^* \rangle \geq 0$, $\forall x \in \mathbb{R}^d$.

Localization set: After N steps, NSM "builds"

$$Q_N^+ := \{x : \langle g_k, x_k - x \rangle \geq 0, \ \ k = 0, \ldots, N\}$$

such that $x^* \in Q_N^+$.

Main result: For appropriately chosen step sizes,

$$\text{Size}(Q_N^+) := \max\{r \geq 0 : B(x^*, r) \subseteq Q_N^+\} \to 0.$$

Note:

$$\text{Size}(Q_N^+) = v_N^* := \min_{0 \leq k \leq N} v_k,$$

where v_k is the distance from x^* to the supporting hyperplane at x_k:

$$v_k := \frac{\langle g_k, x_k - x^* \rangle}{\|g_k\|} \geq 0$$

Standard Convergence Analysis

Main recurrence: Denote $\rho_k = \|x_k - x^*\|$. Then,

$$\frac{1}{2}\rho_{k+1}^2 = \frac{1}{2}\rho_k^2 - h_k v_k + \frac{1}{2}h_k^2.$$

Convergence rate: For $R := \|x_0 - x^*\|$, we get

$$v_N^* \leq \frac{1}{\sum_{k=0}^N h_k} \sum_{k=0}^N h_k v_k \leq \frac{R^2 + \sum_{k=0}^N h_k^2}{2\sum_{k=0}^N h_k}.$$

Choose $h_k \equiv h = \frac{R_0}{\sqrt{N+1}}$ and denote $\bar{R} = \frac{1}{2}(\frac{R^2}{R_0} + R_0)$. Then,

$$v_N^* \leq \frac{\bar{R}}{\sqrt{N+1}} \leq \Delta \quad \text{if} \quad N + 1 \geq \boxed{N(\Delta) := \frac{\bar{R}^2}{\Delta^2}}.$$

NB: The optimal choice is $R_0 = R$, giving us $\bar{R} = R$.

Estimating Quality of Approximate Solution

Growth function: $\omega(r) := \max_x\{f(x) - f^* : \|x - x^*\| \leq r\}$.

Main result (Lemma 3.2.1 in [Nesterov 2018]):

$$f(x) - f^* \leq \omega\big(v(x)\big),$$

where $v(x) := \frac{\langle f'(x), x - x^* \rangle}{\|f'(x)\|}$.

Corollary: For $x_N^* := \operatorname{argmin}\{f(x) : x \in \{x_0, \ldots, x_N\}\}$, we have

$$f(x_N^*) - f^* \leq \omega(v_N^*) \leq \omega(\Delta) \quad \text{whenever} \quad v_N^* \leq \Delta.$$

Growth Functions and Complexity Bounds

Nonsmooth Lipschitz Functions

Problem class: $|f(x) - f(y)| \leq L_0 \|x - y\|$.

Growth function:

$$\omega(r) \leq L_0 r \leq \varepsilon \quad \text{if} \quad r \leq \Delta(\varepsilon) := \frac{\varepsilon}{L_0}.$$

Complexity:

$$N(\Delta) = \frac{\bar{R}^2}{\Delta^2} \quad \Longrightarrow \quad \boxed{N(\Delta(\varepsilon)) = \frac{L_0^2 \bar{R}^2}{\varepsilon^2}.}$$

This is the standard complexity of the Subgradient Method on nonsmooth Lipschitz functions.

Lipschitz-Smooth Functions

Problem class: $\|\nabla f(x) - \nabla f(y)\| \leq L_1 \|x - y\|$.

Upper bound:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_1}{2} \|y - x\|^2.$$

Growth function: Assuming $\nabla f(x^*) = 0$, we get

$$\omega(r) \leq \frac{L_1}{2} r^2 \leq \varepsilon \quad \text{if} \quad r \leq \Delta(\varepsilon) := \sqrt{\frac{2\varepsilon}{L_1}}.$$

Complexity:

$$N(\Delta) = \frac{\bar{R}^2}{\Delta^2} \quad \Longrightarrow \quad \boxed{N(\Delta(\varepsilon)) = \frac{L_1 \bar{R}^2}{2\varepsilon}.}$$

This is the standard complexity of Gradient Descent for smooth functions.

Hölder-Smooth Functions

Problem class: $\|\nabla f(x) - \nabla f(y)\| \leq H_\nu \|x - y\|^\nu, \quad \nu \in [0, 1]$.

Upper bound:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{H_\nu}{1+\nu} \|y - x\|^{1+\nu}.$$

Growth function:

$$\omega(r) \leq \frac{H_\nu}{1+\nu} r^{1+\nu} \leq \varepsilon \quad \text{if} \quad r \leq \Delta(\varepsilon) := \left[\frac{(1+\nu)\varepsilon}{H_\nu} \right]^{\frac{1}{1+\nu}}.$$

Complexity:

$$N(\Delta) = \frac{\bar{R}^2}{\Delta^2} \quad \Longrightarrow \quad \boxed{N(\Delta(\varepsilon)) = \left[\frac{H_\nu}{(1+\nu)\varepsilon} \right]^{\frac{2}{1+\nu}} \bar{R}^2.}$$

This is the complexity of the Universal Gradient Method [Nesterov 2015].

High-Order-Smooth Functions

Problem class: $\|D^p f(x) - D^p f(y)\| \leq L_p \|x - y\|, \quad p \geq 2.$

Upper bound:

$$f(y) \leq f(x) + \sum_{i=1}^{p} \frac{1}{i!} D^i f(x)[y - x]^i + \frac{L_p}{(p+1)!} \|y - x\|^{p+1}.$$

Growth function: Denoting $L_{i-1} := \|D^i f(x^*)\|$, $i = 2, \ldots, p$, we get

$$\omega(r) \leq \sum_{i=2}^{p+1} \frac{L_{i-1}}{i!} r^i \leq \varepsilon \quad \text{if} \quad r \leq \Delta(\varepsilon) := \min_{2 \leq i \leq p+1} \left[\frac{i! \varepsilon}{p L_{i-1}} \right]^{\frac{1}{i}}.$$

Complexity:

$$N(\Delta) = \frac{\bar{R}^2}{\Delta^2} \quad \Longrightarrow \quad \boxed{N(\Delta(\varepsilon)) = \max_{2 \leq i \leq p+1} \left[\frac{p L_{i-1}}{i! \varepsilon} \right]^{\frac{2}{i}} \bar{R}^2.}$$

Quasi-Self-Concordant (QSC) Functions [Bach 2010]

Problem class: $D^3 f(x)[u,u,v] \leq M\langle \nabla^2 f(x)u, u\rangle \|v\|$.

Upper bound: [Doikov 2023]

$$f(y) \leq f(x) + \langle \nabla f(x), y-x\rangle + \langle \nabla^2 f(x)(y-x), y-x\rangle \varphi(M\|y-x\|),$$

where $\varphi(t) := \frac{e^t - t - 1}{t^2}$.

Growth function: Denoting $L_1^* = \|\nabla^2 f(x^*)\|$, we get

$$\omega(r) \leq L_1^* r^2 \varphi(Mr) \leq \varepsilon \quad \text{if} \quad r \leq \Delta(\varepsilon) := \min\left\{\frac{1}{M}, \sqrt{\frac{\varepsilon}{(e-2)L_1^*}}\right\}.$$

Complexity:

$$N(\Delta) = \frac{\bar{R}^2}{\Delta^2} \quad \Longrightarrow \quad \boxed{N(\Delta(\varepsilon)) = \max\left\{M^2\bar{R}^2, (e-2)\frac{L_1^*\bar{R}^2}{\varepsilon}\right\}.}$$

(L_0, L_1)-Smooth Functions [Zhang et al. 2020]

Problem class: $\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|$.

Upper bound: [Vankov et al. 2025]

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_0 + L_1 \|\nabla f(x)\|}{L_1^2} \xi(L_1 \|y - x\|),$$

where $\xi(t) := e^t - t - 1$.

Growth function:

$$\omega(r) \leq \frac{L_0}{L_1^2} \xi(L_1 r) \leq \varepsilon \quad \text{if} \quad r \leq \min\left\{ \frac{1}{L_1}, \sqrt{\frac{4\varepsilon}{3L_0}} \right\}.$$

Complexity:

$$N(\Delta) = \frac{\bar{R}^2}{\Delta^2} \quad \Longrightarrow \quad \boxed{N(\Delta(\varepsilon)) = \max\left\{ L_1^2 \bar{R}^2, \frac{3L_0 \bar{R}^2}{4\varepsilon} \right\}.}$$

Distance Adaptation

Motivation

- The complexity $N(\Delta) = \frac{\bar{R}^2}{\Delta^2}$ of NSM depends on $\bar{R} = \frac{1}{2}(\frac{R^2}{R_0} + R_0)$, where $R = \|x_0 - x^*\|$ and R_0 is our "initial guess" of R.
- The best possible choice of R_0 would be $R_0 = R$ resulting in $\bar{R} = R$ and complexity $N^*(\Delta) = \frac{R^2}{\Delta^2}$. However, normally, we do not know R!
- NSM is too sensitive to the "wrong" choice of R_0:

$$\frac{N(\Delta)}{N^*(\Delta)} = \left[\frac{\bar{R}}{R}\right]^2 \simeq \left[\frac{R}{R_0} + \frac{R_0}{R}\right]^2 \simeq \left[\max\left\{\frac{R}{R_0}, \frac{R_0}{R}\right\}\right]^2.$$

  Overestimating or underestimating R in 10 (100) times, could slow the method in 100 (10000) times!

Can we make NSM more robust to the "wrong" choice of R_0?

Yes, using the recently proposed adaptive distance-estimation technique for stochastic AdaGrad-type methods [Ivgi et al. 2023]. This will reduce the "sensitivity" down to $\left[\max\{\log \frac{R}{R_0}, \frac{R_0}{R}\}\right]^2$.

Basic Distance-Adaptive NSM (DA-NSM)

NSM:

$$\boxed{x_{k+1} = x_k - h_k \frac{g_k}{\|g_k\|}}, \qquad k \geq 0.$$

Classical step sizes: $h_k = \frac{R_0}{\sqrt{N+1}}$.

Distance-adaptive step sizes:

$$\boxed{h_k = \frac{c_0 R_k}{\sqrt{N+1}},}$$

where $c_0 \in (0,1)$, and R_k is a distance-estimation (DE) sequence:

$$\boxed{R_{k+1} = \max\{R_k, r_{k+1}\}, \quad r_{k+1} = \|x_0 - x_{k+1}\|,} \qquad k \geq 0.$$

Main Result

Complexity:

$$\boxed{N(\Delta) = e^2 \frac{\bar{R}^2}{\Delta^2} \ln^2 \frac{eD}{R_0},}$$

where $D = \max\{\frac{2R}{1-c_0}, R_0\}$ $(\simeq R + R_0)$ and $\bar{R} = \frac{1}{c_0} R + \frac{c_0}{2} D$ $(\simeq R + R_0)$.

- If $R_0 \lesssim R$, then $D \simeq \bar{R} \simeq R$ and

$$N(\Delta) \simeq \frac{R^2}{\Delta^2} \ln_*^2 \frac{R}{R_0},$$

  $\ln_* t := \ln(et)$. In particular, if $R \simeq R_0$, then $N(\Delta) = N^*(\Delta) = \frac{R^2}{\Delta^2}$.
- If $R_0 \gtrsim R$, then $D \simeq \bar{R} \simeq R_0$ and

$$N(\Delta) \simeq \frac{R_0^2}{\Delta^2}.$$

Analysis I: Boundedness of DE Sequence

- We still have the main recurrence ($\rho_k \equiv \|x_k - x^*\|$):

$$\sum_{k=0}^{t} h_k v_k + \frac{1}{2}\rho_{t+1}^2 \leq \frac{1}{2}\rho_0^2 + \frac{1}{2}\sum_{k=0}^{t} h_k^2.$$

- Dropping $v_k \geq 0$, we get, for any $0 \leq t \leq N$ ($h_k \equiv \frac{c_0 R_k}{\sqrt{N+1}}$):

$$\rho_{t+1}^2 \leq R^2 + \sum_{k=0}^{t} h_k^2 \leq R^2 + c_0^2 \sum_{k=0}^{t} \frac{R_k^2}{N+1} \leq R^2 + c_0^2 R_t^2.$$

- Therefore,

$$r_{t+1} \equiv \|x_0 - x_{t+1}\| \leq R + \rho_{t+1} \leq 2R + c_0 R_t$$

and

$$R_{t+1} \equiv \max\{R_t, r_{t+1}\} \leq \max\{R_t, 2R + c_0 R_t\}.$$

- This implies

$$R_{t+1} \leq D \equiv \max\left\{\frac{2R}{1-c_0}, R_0\right\}, \qquad \rho_{t+1} \leq R + c_0 D.$$

Analysis II: Final Convergence Rate

- Now we estimate ($r_{t+1} \equiv \|x_0 - x_{t+1}\|$)

$$\frac{1}{2}(\rho_0^2 - \rho_{t+1}^2) = \frac{1}{2}(\rho_0 - \rho_{t+1})(\rho_0 + \rho_{t+1}) \leq |\rho_0 - \rho_{t+1}|\rho_0 \leq r_{t+1}R.$$

- This gives us ($h_k \equiv \frac{c_0 R_k}{\sqrt{N+1}}$, $\bar{R} \equiv \frac{1}{c_0}R + \frac{c_0}{2}D$):

$$v_t^* \leq \frac{1}{\sum_{k=0}^t h_k} \sum_{k=0}^t h_k v_k \leq \frac{r_{t+1}R + \frac{1}{2}\sum_{k=0}^t h_k^2}{\sum_{k=0}^t h_k} \leq \frac{R_{t+1}}{\sum_{k=0}^t R_k}\sqrt{N+1}\,\bar{R}.$$

- For any nondecreasing and bounded (by D) sequence, we have

$$\min_{0 \leq t \leq N} \frac{R_{t+1}}{\sum_{k=0}^t R_k} \leq \frac{\xi_N}{N+1}, \qquad \xi_N := \left(\frac{D}{R_0}\right)^{\frac{1}{N+1}} \ln \frac{eD}{R_0}.$$

- Hence, $v_N^* \leq \frac{\xi_N \bar{R}}{\sqrt{N+1}}$, where $\xi_N \leq e \ln \frac{eD}{R_0}$ for $N+1 \geq \ln \frac{D}{R_0}$.

Normalization for Constrained Problems

Constrained Problems

From now on, we consider a more general problem:

$$\min_{x \in Q} f(x),$$

where $Q \subseteq \mathbb{R}^d$ is a simple closed convex set (possibly unbounded).

For such a problem, the Subgradient Method (SM) must be modified:

$$x_0 \in Q, \qquad x_{k+1} = T(x_k, \lambda_k), \quad k \geq 0,$$

where

$$\boxed{T(x, \lambda) \equiv \operatorname*{argmin}_{y \in Q} \left\{ \lambda \langle g(x), y - x \rangle + \frac{1}{2} \|y - x\|^2 \right\}} = \pi_Q(x - \lambda g(x)),$$

and $\pi_Q(u) = \operatorname{argmin}_{y \in Q} \|y - u\|$ is the Euclidean projection onto Q.

Is the old normalization $\lambda_k = \frac{h_k}{\|g_k\|}$ still the "correct" idea?

Indirect Control of Step Size [Nesterov 2024]

Scaling function: $\varphi_x(\lambda) := \max\limits_{y \in Q} \left\{ \lambda \langle g(x), x - y \rangle - \frac{1}{2}\|x - y\|^2 \right\} \geq 0.$

- $\varphi_x(\cdot)$ is a continuously differentiable increasing convex function.
- $\varphi_x(0) = 0$, $\varphi_x'(\lambda) = \langle g(x), x - T \rangle \geq 0$, where $T \equiv T(x, \lambda)$.
- $\varphi_x(\lambda) \leq \frac{1}{2}\lambda^2 \|g(x)\|^2$. If $Q = \mathbb{R}^d$, then $\varphi_x(\lambda) = \frac{1}{2}\lambda^2 \|g(x)\|^2$.

Main recurrence for SM ($\rho_k \equiv \|x_k - x^*\|$):

$$\sum_{k=0}^{t} \lambda_k \langle g_k, x_k - x^* \rangle + \frac{1}{2}\rho_{t+1}^2 \leq \frac{1}{2}\rho_0^2 + \sum_{k=0}^{t} \varphi_{x_k}(\lambda_k).$$

Generalized Normalized Step Size:

$$\boxed{\lambda \equiv \lambda_x(h): \qquad \varphi_x(\lambda) \leq \frac{1}{2}h^2 \quad \text{and} \quad \lambda \geq \frac{h}{\|g(x)\|}.}$$

Choosing $\lambda_k = \lambda_{x_k}(h_k)$ in SM results in the same recurrence as before:

$$\sum_{k=0}^{t} h_k v_k + \frac{1}{2}\rho_{t+1}^2 \leq \frac{1}{2}\rho_0^2 + \frac{1}{2}\sum_{k=0}^{t} h_k^2.$$

Generalized Normalized Step Size: Examples

Classical Normalized Step Size: $\boxed{\lambda_x^{\mathrm{cl}}(h) := \dfrac{h}{\|g(x)\|}.}$

Set-Scaled Normalized Step Size: [Nesterov 2024]

$$\boxed{\lambda_x^{\mathrm{SS}}(h) := \text{solution of equation } \varphi_x(\lambda) = \frac{1}{2}h^2.}$$

- Can be computed efficiently by either a Newton-type method or a direct formula (e.g., when Q is a Euclidean ball) with no extra queries to the oracle.
- For $\lambda \equiv \lambda_x^{\mathrm{SS}}(h)$, $T \equiv T(x,\lambda)$, we have $\lambda \geq \frac{h\|x-T\|}{\langle g(x), x-T\rangle} \geq \frac{h}{\|g(x)\|}$.

In what follows, we denote

$$T_h(x) := T\big(x, \lambda_x(h)\big),$$

where $\lambda_x(\cdot)$ is a generalized normalized step size (e.g., any of the above).

Short-Step NSM (SS-NSM) for Constrained Problems

SS-NSM:

$$x_0 \in Q, \qquad \boxed{x_{k+1} = T_{h_k}(x_k)}, \quad k \geq 0.$$

We can still apply the same short-step-size strategies:

- Classical: $h_k = \frac{R_0}{\sqrt{N+1}}$.
- DA: $h_k = \frac{c_0 R_k}{\sqrt{N+1}}$, where R_k is a DE sequence based on x_k.

This results in the same complexity guarantees after N steps as before.

Long-Step NSM (LS-NSM)

Motivation

The previously considered NSM

$$\boxed{x_{k+1} = T_{h_k}(x_k)}, \qquad k \geq 0,$$

was using **short step sizes**:

$$h_k \equiv h = \frac{c_0 R_k}{\sqrt{N+1}}, \qquad k \geq 0.$$

- For large N, the step sizes at the beginning of the process are very small. This hinders empirical performance of the algorithm.
- The algorithm does not provide any meaningful guarantee before it makes exactly N steps.
- After it makes N steps, it is impossible to easily continue the optimization process without restarting the method.

To fix these drawbacks, we would like to use **long step sizes**:

$$h_k \approx \frac{c_0 R_k}{\sqrt{k+1}}.$$

Classical Theory

Classical long step sizes: $\boxed{h_k = \frac{R_0}{\sqrt{k+1}},} \qquad k \geq 0.$

Recall: $v_N^* \leq \frac{R^2 + \sum_{k=0}^N h_k^2}{2\sum_{k=0}^N h_k}.$

We have $\sum_{k=0}^N h_k \geq R_0\sqrt{N+1}$ and $\sum_{k=0}^N h_k^2 \leq R_0^2 \ln[e(N+1)]$.

Denoting $\bar{R} = \frac{1}{2}(\frac{R^2}{R_0} + R_0)$, we thus get

$$v_N^* \leq \frac{\frac{R^2}{R_0} + R_0 \ln[e(N+1)]}{2\sqrt{N+1}} \leq \frac{\bar{R}\ln[e(N+1)]}{\sqrt{N+1}}.$$

Complexity: $\boxed{N(\Delta) = \frac{4e^2}{(e-1)^2}\frac{\bar{R}^2}{\Delta^2}\ln_+^2 \frac{2e^{\frac{1}{2}}\bar{R}}{\Delta}.}$

CF: For the short-step method, we had $N(\Delta) = \frac{\bar{R}^2}{\Delta^2}$.

Towards DA Step Sizes

Unfortunately, the "classical choice" does not work with DA step sizes

$$h_k = \frac{c_0 R_k}{\sqrt{k+1}}, \qquad k \geq 0,$$

because $\rho_t := \|x_t - x^*\|$ may become unbounded. Recall:

$$\rho_{t+1}^2 \leq R^2 + \sum_{k=0}^{t} h_k^2 = R^2 + c_0^2 \sum_{k=0}^{t} \frac{R_k^2}{k+1}.$$

Even if $R_k \leq \bar{R}$ for all $k \geq 0$, the sum in the right-hand side diverges!

To fix this, we need a scaling sequence γ_k such that

$$\gamma_k \approx \frac{1}{c_0}\sqrt{k+1} \quad \text{and} \quad \sum_{k=0}^{\infty} \frac{1}{\gamma_k^2} = c_0^2 < 1.$$

This is the sequence (c.f. [Ivgi et al. 2023])

$$\boxed{\gamma_k = \frac{1}{c_0}\sqrt{2(k+1)\log[e(k+1)]},} \qquad k \geq 0.$$

Distance-Adaptive LS-NSM (DA-LS-NSM)

Choosing now

$$h_k = \frac{R_k}{\gamma_k}, \quad \gamma_k = \frac{1}{c_0}\sqrt{2(k+1)}\log[e(k+1)], \qquad k \geq 0,$$

we get the following complexity result

$$\boxed{N(\Delta) = \frac{4e^4}{(e-1)^2}\frac{\bar{D}^2}{\Delta^2}\ln^2\frac{eD}{R_0}\ln_+^2\frac{2e^{\frac{3}{2}}\bar{D}\ln\frac{eD}{R_0}}{\Delta}},$$

$D = \max\{\frac{2R}{1-c_0}, R_0\}$ $(\simeq R + R_0)$, $\bar{D} = \frac{2^{\frac{3}{2}}}{c_0}R + (2^{\frac{3}{2}} + \frac{c_0}{\sqrt{2}})D$ $(\simeq R + R_0)$.

NB: For DA-SS-NSM, we had $N(\Delta) \simeq \frac{\bar{D}^2}{\Delta^2}\ln^2\frac{eD}{R_0}$.

To remove the red logarithmic factor while still using "long step sizes", we need more advanced algorithms...

Normalized Dual Subgradient Method (NDSM)

Normalized Dual Subgradient Method (NDSM)

Dual Averaging [Nesterov 2009]: Choose $x_0 \in Q$ and iterate:

$$\boxed{x_{t+1} = \operatorname*{argmin}_{x \in Q} \left\{ \sum_{k=0}^{t} a_k \langle g(x_k), x - x_k \rangle + \frac{\gamma_t}{2} \|x - x_0\|^2 \right\}, \qquad t \geq 0,}$$

where $a_k \geq 0$ is a sequence of weights and γ_t is a nondecreasing sequence of scaling coefficients:

$$0 < \gamma_t \leq \gamma_{t+1}, \qquad t \geq 0.$$

Normalized weights:

$$\boxed{a_k = \gamma_{k-1} \lambda_{x_k} \left(\frac{\alpha_k}{\gamma_{k-1}} \right),} \qquad k \geq 0,$$

where $\alpha_k > 0$ is an auxiliary control sequence and $\gamma_{-1} := \gamma_0$.

Classical Theory

Recurrence: $\sum_{k=0}^{t} \alpha_k v_k + \frac{\gamma_t}{2}\rho_{t+1}^2 \leq \frac{\gamma_t}{2}\rho_0^2 + \frac{1}{2}\sum_{k=0}^{t} \frac{\alpha_k^2}{\gamma_{k-1}}$ $(\rho_k = \|x_k - x^*\|)$.

CF: For NSM, we had $\sum_{k=0}^{t} h_k v_k + \frac{1}{2}\rho_{t+1}^2 \leq \frac{1}{2}\rho_0^2 + \frac{1}{2}\sum_{k=0}^{t} h_k^2$.

Gap bound: $v_t^* \leq \dfrac{\gamma_t R^2 + \sum_{k=0}^{t} \frac{\alpha_k^2}{\gamma_{k-1}}}{2\sum_{k=0}^{t} \alpha_k}$.

Classical choice of coefficients: $\gamma_{-1} = 1$ and

$$\boxed{\alpha_t \equiv R_0, \qquad \gamma_t = \sum_{k=0}^{t} \frac{1}{\gamma_{k-1}} \quad (\Rightarrow \gamma_t \simeq \sqrt{t+1}\,)}, \qquad t \geq 0.$$

This results in $v_N^* \lesssim \frac{\bar{R}}{\sqrt{N+1}}$, where $\bar{R} = \frac{1}{2}(\frac{R^2}{R_0} + R_0)$.

Complexity: $\boxed{N(\Delta) \simeq \dfrac{\bar{R}^2}{\Delta^2}.}$ Same as in SS-NSM!

Distance-Adaptive NDSM (DA-NDSM)

DA coefficients: $\gamma_{-1} = \frac{1}{c_0}$ and

$$\boxed{\alpha_t = R_t, \qquad \gamma_t = \frac{1}{c_0^2} \sum_{k=0}^{t} \frac{1}{\gamma_{k-1}} \quad (\Rightarrow \gamma_t \simeq \sqrt{t+1}\,),} \qquad t \geq 0,$$

where $c_0 \in (0,1)$ is an absolute constant, and R_t is a DE sequence for x_t:

$$\boxed{R_{t+1} = \max\{R_t, r_{t+1}\}, \quad r_{t+1} = \|x_0 - x_{t+1}\|,} \qquad t \geq 0.$$

Complexity:

$$\boxed{N(\Delta) = e^2 \frac{\bar{D}^2}{\Delta^2} \ln^2 \frac{eD}{R_0},}$$

where $D = \max\{\frac{2R}{1-c_0}, R_0\}$ $(\simeq R + R_0)$, $\bar{D} = \sqrt{3}(\frac{1}{c_0} R + \frac{c_0}{2} D)$ $(\simeq R + R_0)$.

NB: This is the same complexity as in DA-SS-NSM.

NDSM with Extra Averaging (NDSM-EA)

NDSM with Extra Avearaging (NDSM-EA)

DSM-EA [Nesterov and Shikhman 2015]: Choose $z_0 \in Q$ and iterate:

$$\boxed{\begin{aligned} x_t &= \frac{\gamma_{t-1}}{\gamma_t} z_t + \left(1 - \frac{\gamma_{t-1}}{\gamma_t}\right) z_0, \\ z_{t+1} &= \operatorname*{argmin}_{x \in Q} \left\{ \sum_{k=0}^{t} a_k \langle g(x_k), x - x_k \rangle + \frac{\gamma_t}{2} \|x - z_0\|^2 \right\}, \end{aligned}} \qquad t \geq 0.$$

Normalized weights:

$$\boxed{a_k = \gamma_k \lambda_{x_k}\left(\frac{\alpha_k}{\gamma_k}\right),} \qquad k \geq 0,$$

where $\alpha_k > 0$ is an auxiliary control sequence.

CF: For NDSM, we had $x_t = z_t$ and $a_k = \gamma_{k-1} \lambda_{x_k}(\frac{\alpha_k}{\gamma_{k-1}})$.

Classical Theory

Main recurrence: $\sum_{k=0}^{t} \alpha_k v_k + \frac{\gamma_t}{2}\rho_{t+1}^2 \leq \frac{\gamma_t}{2}\rho_0^2 + \frac{1}{2}\sum_{k=0}^{t} \frac{\alpha_k^2}{\gamma_k}$, where $\rho_k = \|z_k - x^*\|$ and $v_k = \frac{\langle g(x_k), x_k - x^* \rangle}{\|g(x_k)\|}$.

CF: For NDSM, we had $\sum_{k=0}^{t} \alpha_k v_k + \frac{\gamma_t}{2}\rho_{t+1}^2 \leq \frac{\gamma_t}{2}\rho_0^2 + \frac{1}{2}\sum_{k=0}^{t} \frac{\alpha_k^2}{\gamma_{k-1}}$.

Gap bound: $v_t^* \leq \frac{\gamma_t R^2 + \sum_{k=0}^{t} \frac{\alpha_k^2}{\gamma_k}}{2\sum_{k=0}^{t} \alpha_k}$.

Choice of coefficients: Almost as before:

$$\boxed{\alpha_t = R_0, \qquad \gamma_t = \sum_{k=0}^{t} \frac{1}{\gamma_k} \quad \left(\iff \gamma_t \simeq \sqrt{t+1} \right),} \qquad t \geq 0.$$

Complexity: $\boxed{N(\Delta) \simeq \frac{\bar{R}^2}{\Delta^2}.}$ Same as in NDSM.

The advantage of NDSM-EA over NDSM arises when α_t is not a well-chosen predefined sequence (next slide).

Distance-Adaptive NDSM-EA (DA-NDSM-EA)

Gap bound: $v_t^* \leq \dfrac{\gamma_t r_{t+1} R + \frac{1}{2}\sum_{k=0}^{t} \frac{\alpha_k^2}{\gamma_k}}{\sum_{k=0}^{t} \alpha_k}.$

DA coefficients:

$$\boxed{\alpha_t = R_t^p, \quad \gamma_t = \frac{1}{c_0^2} \sum_{k=0}^{t} \frac{\alpha_k^2}{\gamma_k R_k^2} \quad \left(\iff \gamma_t \simeq \sqrt{\sum_{k=0}^{t} R_k^{2(p-1)}} \right), \quad t \geq 0,}$$

where $p \geq 1$, $c_0 \in (0, 1)$, and R_t is a DE-sequence for z_t.

NB:

- We now set α_t to a power of R_t (inspired by [Khaled et al. 2023]).
- The most interesting cases are $p = 1$ and $p = 2$.

Efficiency of DA-NDSM-EA

Gap bound: $v_N^* \leq \delta_N^* \bar{R}$, where $\delta_N^* := \min_{0 \leq t \leq N} \frac{R_{t+1} \sqrt{\sum_{k=0}^t R_k^{2(p-1)}}}{\sum_{k=0}^t R_k^p}$, and $D = \max\{\frac{2R}{1-c_0}, R_0\}$ $(\simeq R + R_0)$, $\bar{R} = \frac{\sqrt{2}}{c_0} R + \frac{c_0}{\sqrt{2}} D$ $(\simeq R + R_0)$.

Option $p = 1$**:** Then, $\delta_N^* = \min_{0 \leq t \leq N} \frac{R_{t+1}}{\sum_{k=0}^t R_k} \sqrt{t+1}$, $\gamma_t \simeq \sqrt{t+1}$,

$$N(\Delta) = e^2 \frac{\bar{R}^2}{\Delta^2} \ln^2 \frac{eD}{R_0}.$$

Same as in DA-NDSM.

Option $p = 2$**:** Then, $\delta_N^* = \min_{0 \leq t \leq N} \sqrt{\frac{R_{t+1}^2}{\sum_{k=0}^t R_k^2}}$, $\gamma_t \simeq \sqrt{\sum_{k=0}^t R_k^2}$,

$$\boxed{N(\Delta) = e^2 \frac{\bar{R}^2}{\Delta^2} \ln \frac{eD^2}{R_0^2}.}$$

The logarithmic factor now has power 1 instead of 2.

(Primal) NSM-EA

(Primal) NSM-EA

The previous NDSM-EA has a simple primal counterpart:

NSM-EA [Nesterov 2024]: Choose $z_0 \in Q$ and iterate:

$$\boxed{\begin{aligned} x_k &= \frac{\gamma_{k-1}}{\gamma_k} z_k + \left(1 - \frac{\gamma_{k-1}}{\gamma_k}\right) z_0, \\ z_{k+1} &= T_{h_k}(x_k), \quad h_k := \frac{\alpha_k}{\gamma_k}, \end{aligned}} \qquad k \geq 0.$$

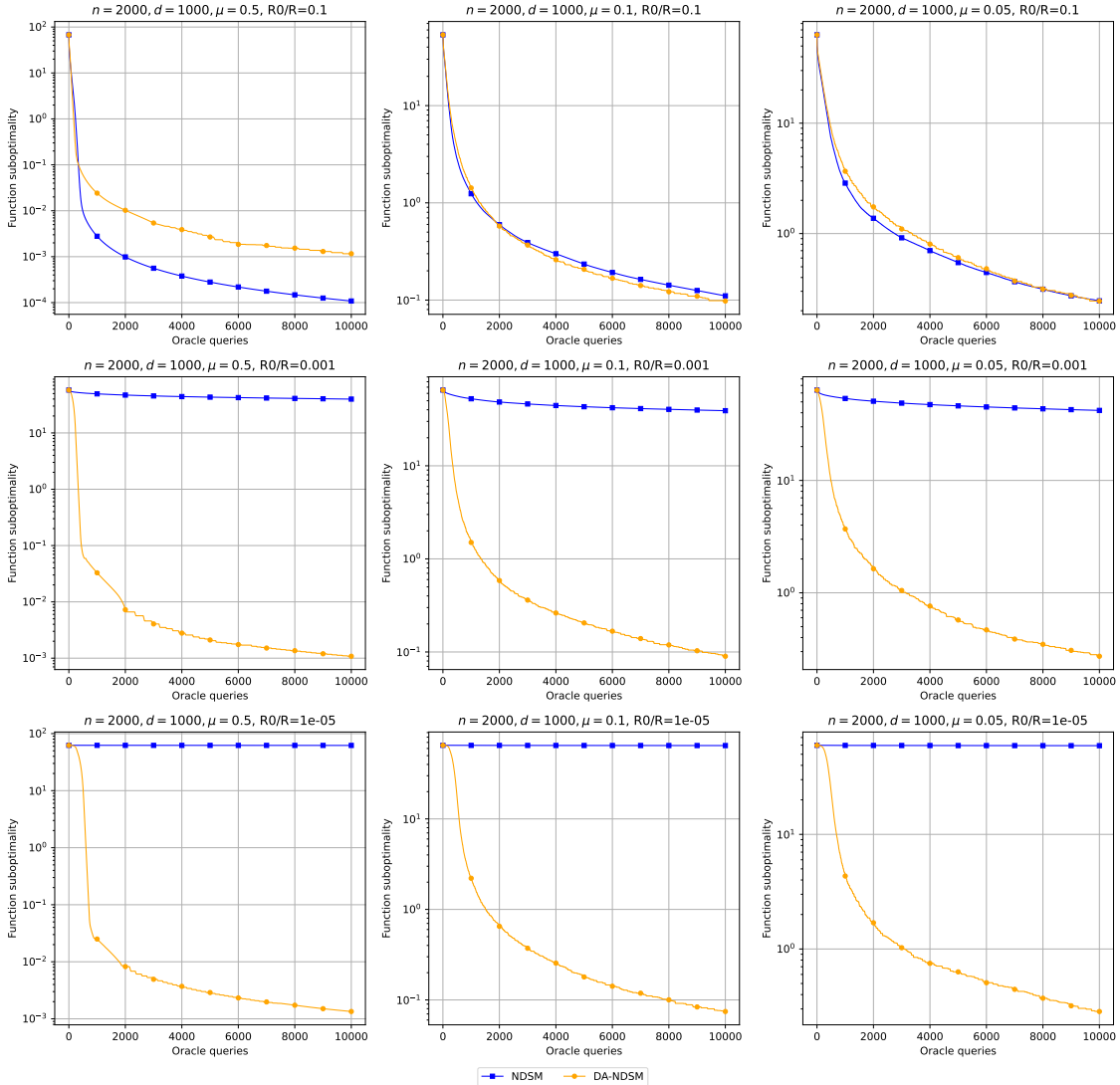
($T_{h_k}(x_k)$ is the normalized subgradient step from x_k.)

- Shares the same main recurrence as NDSM-EA.
- Hence, the same choices of weights α_k and scaling coefficients γ_k gives identical efficiency bounds, both in the classical and DA versions.

Experiments

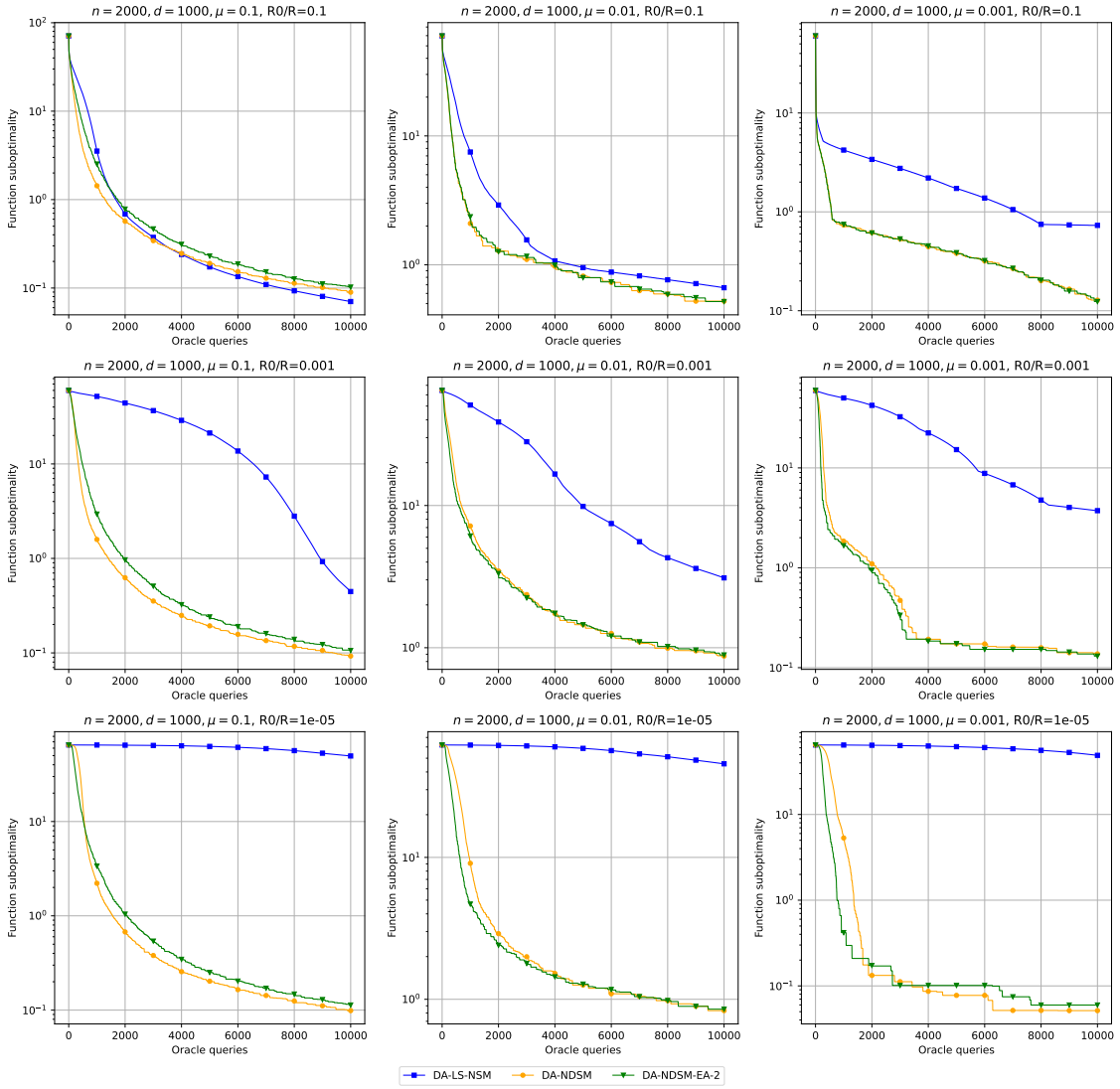
Standard vs DA Methods

Softmax: $\min_{x \in \mathbb{R}^d} \left\{ f(x) = \mu \ln \sum_{i=1}^{n} e^{[\langle a_i, x \rangle - b_i]/\mu} \right\}$.



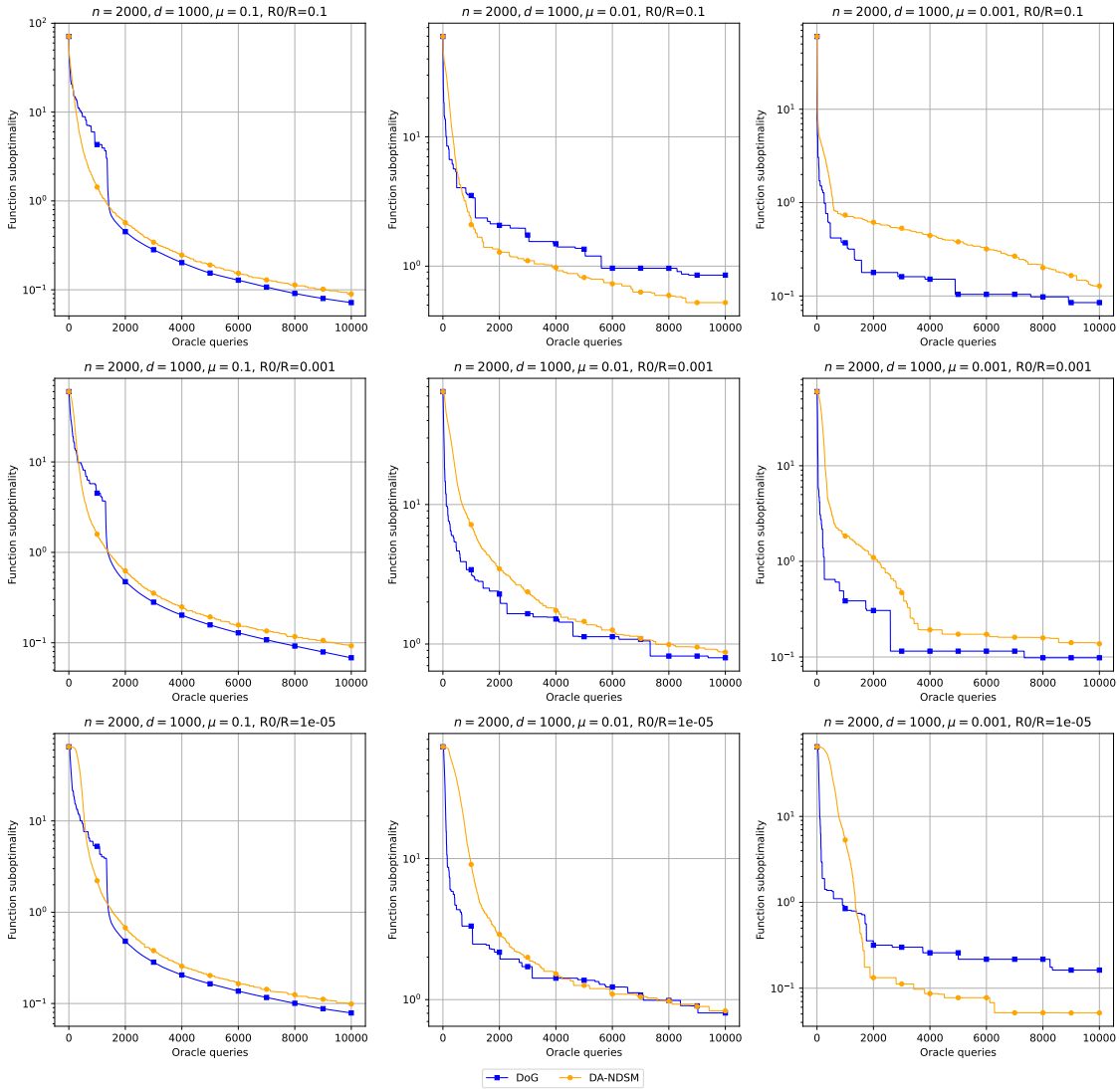
Standard methods are too sensitive to the "wrong" choice of R_0!

DA Normalized Methods (Softmax)



- DA-LS-NSM is slower because of the logarithmic safety factor in the step size.
- Not shown: DA-NDSM-EA-1 $\simeq$ DA-NDSM, DA-NSM-p $\simeq$ DA-NDSM-EA-p.

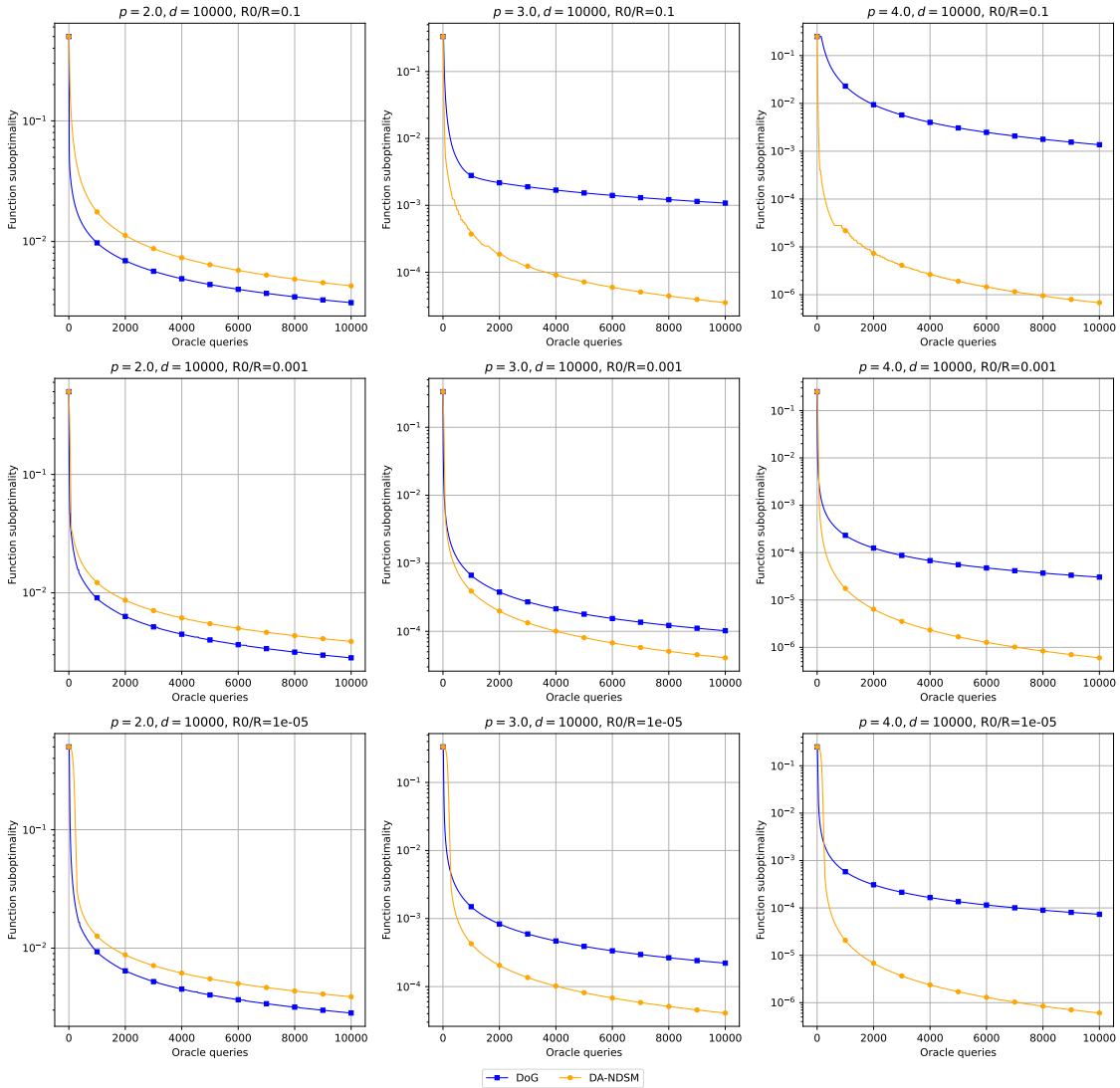
Normalized vs AdaGrad-type Methods (Softmax)



- DoG is a foundational DA AdaGrad-type method from (Ivgi et al. 2023); we show the authors' heuristic version (faster in practice, but without a convergence guarantee).
- DA-NDSM performs similarly to DoG, sometimes slightly slower.

Normalized vs AdaGrad-type Methods – II

Problem: $\min_{x \in \mathbb{R}^d} \left\{ f(x) = \frac{1}{p} \sum_{i=1}^{d-1} |x_i - x_{i+1}|^p + \frac{1}{p} |x_d|^p \right\}$, $p \geq 2$.



AdaGrad-type methods adapt poorly to higher-order smoothness (unlike normalized methods)?

Conclusions

Conclusions

- NSMs are universal algorithms, suitable for many problem classes.
- Their performance critically depends on a "good" estimate R_0 of the distance $R = \|x_0 - x^*\|$.
- The DA technique is a simple but effective way to relax this requirement: use the maximum of $r_{k+1} = \|x_0 - x_{k+1}\|$ instead of R_0. Originally proposed for a basic AdaGrad-type method, it also works well for many other subgradient algorithms, including normalization-based methods.
- DA extends to problems with functional constraints, handled via a simple switching strategy in the separating-oracle framework.
- Open question: Can DA be applied to non-Euclidean settings using Bregman mappings?

Thank you!

References I


F. Bach. Self-concordant analysis for logistic regression. **Electronic Journal of Statistics**, 4(none), 2010. DOI: 10.1214/09-ejs521.


N. Doikov. Minimizing Quasi-Self-Concordant Functions by Gradient Regularization of Newton Method. **arXiv preprint arXiv:2308.14742**, 2023.


M. Ivgi, O. Hinder, and Y. Carmon. DoG is SGD's Best Friend: A Parameter-Free Dynamic Step Size Schedule. In **Proceedings of the 40th International Conference on Machine Learning**. International Conference on Machine Learning, pages 14465–14499. PMLR, 2023.


A. Khaled, K. Mishchenko, and C. Jin. DoWG Unleashed: An Efficient Universal Parameter-Free Gradient Descent Method. **Advances in Neural Information Processing Systems**, 36:6748–6769, 2023.

References II

- Y. Nesterov. Universal gradient methods for convex optimization problems. **Mathematical Programming**, 152:381–404, 2015. DOI: 10.1007/s10107-014-0790-0.
- Y. Nesterov. Primal-dual subgradient methods for convex problems. **Mathematical Programming**, 120(1):221–259, 2009. DOI: 10.1007/s10107-007-0149-x.
- Y. Nesterov. **Lectures on Convex Optimization**. Volume 137. Springer, 2nd edition, 2018.
- Y. Nesterov. Primal Subgradient Methods with Predefined Step Sizes. **Journal of Optimization Theory and Applications**, 2024. DOI: 10.1007/s10957-024-02456-9.
- Y. Nesterov and V. Shikhman. Quasi-monotone Subgradient Methods for Nonsmooth Convex Minimization. **Journal of Optimization Theory and Applications**, 165(3):917–940, 2015. DOI: 10.1007/s10957-014-0677-5.

References III

D. Vankov, A. Rodomanov, A. Nedich, L. Sankar, and S. Stich. Optimizing (L_0, L_1)-Smooth Functions by Gradient Methods. In **International Conference on Representation Learning**, volume 2025, pages 15953–15979, 2025.

J. Zhang, T. He, S. Sra, and A. Jadbabaie. Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity. In **8th International Conference on Learning Representations**, 2020.