Mathematics of Machine Learning - Summer School

Lecture 6
Mirror Descent

June 30, 2021

Patrick Rebeschini
Department of Statistics, University of Oxford

Recap: Subgradient Descent for Lipschitz Functions

Goal: $\boxed{\min_{x \in \mathcal{C}} f(x)}$ with f convex and $\mathcal{C}$ convex

Projected Subgradient Descent—Lipschitz (Theorem 9.3)

- Function f is γ-Lipschitz
- Assume $\|x_1 - x^\star\|_2 \leq b$

Then, projected subgradient descent with $\eta_s \equiv \eta = \frac{b}{\gamma\sqrt{t}}$ satisfies

$$\boxed{f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^\star) \leq \frac{\gamma b}{\sqrt{t}}}$$

- **Optimal rate.** Lower bound is $\Omega\left(\frac{\gamma a}{1+\sqrt{t}}\right)$ where $a := \max_{x \in \mathcal{C}} \|x\|_2$
- **Dimension-free rate if both the function f and the constraint set $\mathcal{C}$ "behave nicely" with the dimension d (i.e., γ, b do not depend on d)**

It does not always happen...

Subgradient Descent with Euclidean Geometry

Risk minimization:

$$\begin{aligned} &\underset{w}{\text{minimize}} && r(w) = \mathbf{E}\varphi(w^\top XY) \\ &\text{subject to} && \|w\|_2 \leq c_2^{\mathcal{W}} \end{aligned} \implies \text{Let } w^\star \text{ be a minimizer}$$

Empirical risk minimization:

$$\begin{aligned} &\underset{w}{\text{minimize}} && R(w) = \frac{1}{n}\sum_{i=1}^{n}\varphi(w^\top X_i Y_i) \\ &\text{subject to} && \|w\|_2 \leq c_2^{\mathcal{W}} \end{aligned} \implies \text{Let } W^\star \text{ be a minimizer}$$

$$\boxed{r(\overline{W}_t) - r(w^\star) \leq \underbrace{R(\overline{W}_t) - R(W^\star)}_{\texttt{Optimization}} + \underbrace{\sup_{w\in\mathcal{W}}\{r(w) - R(w)\} + \sup_{w\in\mathcal{W}}\{R(w) - r(w)\}}_{\texttt{Statistics}}}$$

$$\boxed{\mathbf{E}\,\texttt{Statistics} \leq \frac{4c_2^{\mathcal{X}}c_2^{\mathcal{W}}\gamma_\varphi}{\sqrt{n}}} \qquad \boxed{\texttt{Optimization} \leq \frac{2c_2^{\mathcal{X}}c_2^{\mathcal{W}}\gamma_\varphi}{\sqrt{t}}}$$

Principled approach: Enough to run algorithm for $t \sim n$ time steps (ONLY BASED ON UPPER BOUNDS!)

Subgradient Descent with Non-Euclidean Geometry

Risk minimization:

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & r(w) = \mathbf{E}\varphi(w^\top XY) \\ \text{subject to} \quad & w \in \Delta_d \end{aligned} \implies \text{Let } w^\star \text{ be a minimizer}$$

Empirical risk minimization:

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & R(w) = \frac{1}{n}\sum_{i=1}^{n} \varphi(w^\top X_i Y_i) \\ \text{subject to} \quad & w \in \Delta_d \end{aligned} \implies \text{Let } W^\star \text{ be a minimizer}$$

$$\boxed{r(\overline{W}_t) - r(w^\star) \leq \underbrace{R(\overline{W}_t) - R(W^\star)}_{\texttt{Optimization}} + \underbrace{\sup_{w \in \mathcal{W}}\{r(w) - R(w)\} + \sup_{w \in \mathcal{W}}\{R(w) - r(w)\}}_{\texttt{Statistics}}}$$

$$\boxed{\mathbf{E}\,\texttt{Statistics} \leq 4 c_\infty^{\mathcal{X}} c_1^{\mathcal{W}} \gamma_\varphi \sqrt{\frac{2 \log d}{n}}} \qquad \boxed{\texttt{Optimization} \leq 2 c_\infty^{\mathcal{X}} c_1^{\mathcal{W}} \gamma_\varphi \sqrt{\frac{d}{t}}}$$

Not same rate with respect to the dimension d

Different Geometry

- **Problem:** Using Cauchy-Schwarz's, R has Lipschitz constant proport. to $\sqrt{d}$:

$$
\begin{aligned}
|R(w) - R(u)| &\leq \frac{1}{n} \sum_{i=1}^{n} |\varphi(w^\top X_i Y_i) - \varphi(u^\top X_i Y_i)| \leq \frac{\gamma_\varphi}{n} \sum_{i=1}^{n} |Y_i (w-u)^\top X_i| \\
&\leq \gamma_\varphi \|w-u\|_2 \max_{i \in [n]} \|X_i\|_2 \leq \sqrt{d} c_\infty^{\mathcal{X}} \gamma_\varphi \|w-u\|_2
\end{aligned}
$$

  as we have $\|x\|_2 \leq \sqrt{d} \|x\|_\infty$ (a sharp inequality)

- **Intuition:** To get $\sqrt{\log d}$ for `Statistics` term, we used Hölder's (Lecture 3)
- **Idea:** Use Hölder's inequality also for `Optimization` term:

$$
\begin{aligned}
|R(w) - R(u)| &\leq \frac{1}{n} \sum_{i=1}^{n} |\varphi(w^\top X_i Y_i) - \varphi(u^\top X_i Y_i)| \leq \frac{\gamma_\varphi}{n} \sum_{i=1}^{n} |Y_i (w-u)^\top X_i| \\
&\leq \gamma_\varphi c_\infty^{\mathcal{X}} \|w-u\|_1
\end{aligned}
$$

- To get a dim.-free Lipschitz constant, we need Lipschitz w.r.t. $\|\cdot\|_1$ norm...

Local-to-Global Properties w.r.t. a Generic Norm

Previous properties can be defined for any norm $\|\cdot\|$ in $\mathbb{R}^d$

- **Convex:** $\boxed{f(y) \geq f(x) + \nabla f(x)^T (y-x) \quad \forall x, y \in \mathbb{R}^d}$

- **α-Strongly Convex:**

$$\boxed{\exists \alpha > 0 \text{ such that } f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\alpha}{2}\|y-x\|^2 \quad \forall x, y \in \mathcal{C}}$$

- **β-Smooth:**

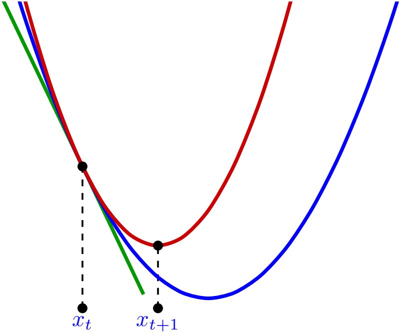
$$\boxed{\exists \beta > 0 \text{ such that } f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{\beta}{2}\|y-x\|^2 \quad \forall x, y \in \mathcal{C}}$$

- **γ-Lipschitz:**

$$\boxed{\exists \gamma > 0 \text{ such that } f(x) - \gamma\|y-x\| \leq f(y) \leq f(x) + \gamma\|y-x\| \ \forall x, y \in \mathcal{C}}$$

Q. What about designing gradient descent that works in *any* geometry?

Gradient descent

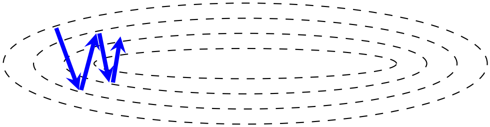


Gradient descent:

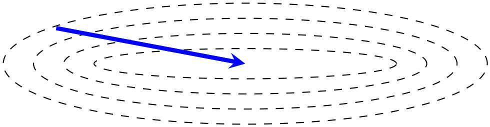
$$x_{t+1} = x_t - \eta_t \nabla f(x_t)$$

$$x_{t+1} = \arg\min_{y \in \mathbb{R}^d} \left\{ \underbrace{f(x_t) + \langle \nabla f(x_t), y - x_t \rangle}_{\text{linear approximation}} + \underbrace{\frac{1}{2\eta_t} \|y - x_t\|_2^2}_{\text{proximal term}} \right\}$$

Newton's method



Gradient descent



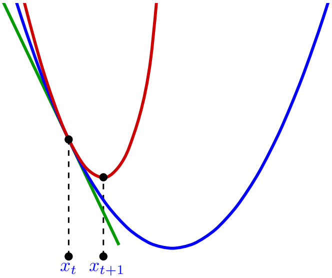
Newton's method

Newton's method:

$$x_{t+1} = x_t - \eta_t (\nabla^2 f(x_t))^{-1} \nabla f(x_t)$$

$$x_{t+1} = \arg\min_{y \in \mathbb{R}^d} \Big\{ \underbrace{f(x_t) + \langle \nabla f(x_t), y - x_t \rangle}_{\text{linear approximation}} + \underbrace{\frac{1}{2\eta_t} (y - x_t)^\top \nabla^2 f(x_t) (y - x_t)}_{\text{proximal term}} \Big\}$$

Mirror descent



Bregman divergence: given $\Phi : \mathbb{R}^d \to \mathbb{R}$ strictly convex and differentiable

$$D^{\Phi}(x, y) = \Phi(x) - \Phi(y) - \langle \nabla\Phi(y), x - y \rangle$$

Mirror descent: [Nemirovsky and Yudin, 1983]

$$x_{t+1} = (\nabla\Phi)^{-1}\big(\nabla\Phi(x_t) - \eta_t \nabla f(x_t)\big)$$

$$x_{t+1} = \arg\min_{y \in \mathbb{R}^d} \Big\{ \underbrace{f(x_t) + \langle \nabla f(x_t), y - x_t \rangle}_{\text{linear approximation}} + \underbrace{\frac{1}{\eta_t} D^{\Phi}(y, x_t)}_{\text{proximal term}} \Big\}$$

Bregman divergences

Function name	$\varphi(x)$	$\operatorname{dom}\varphi$	$D_\varphi(x;y)$
Squared norm	$\frac{1}{2}x^2$	$(-\infty,+\infty)$	$\frac{1}{2}(x-y)^2$
Shannon entropy	$x\log x - x$	$[0,+\infty)$	$x\log\frac{x}{y} - x + y$
Bit entropy	$x\log x + (1-x)\log(1-x)$	$[0,1]$	$x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$
Burg entropy	$-\log x$	$(0,+\infty)$	$\frac{x}{y} - \log\frac{x}{y} - 1$
Hellinger	$-\sqrt{1-x^2}$	$[-1,1]$	$(1-xy)(1-y^2)^{-1/2} - (1-x^2)^{1/2}$
ℓ_p quasi-norm	$-x^p \qquad (0<p<1)$	$[0,+\infty)$	$-x^p + p\,xy^{p-1} - (p-1)\,y^p$
ℓ_p norm	$\lvert x\rvert^p \qquad (1<p<\infty)$	$(-\infty,+\infty)$	$\lvert x\rvert^p - p\,x\operatorname{sgn}y\,\lvert y\rvert^{p-1} + (p-1)\,\lvert y\rvert^p$
Exponential	$\exp x$	$(-\infty,+\infty)$	$\exp x - (x-y+1)\exp y$
Inverse	$1/x$	$(0,+\infty)$	$1/x + x/y^2 - 2/y$

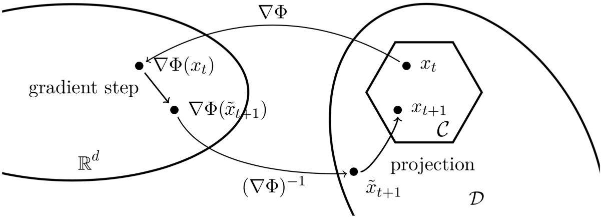
Figure: Table from [Dhillon and Tropp, 2008]

Projected Mirror Descent

Goal: $\boxed{\min_{x \in \mathcal{C}} f(x)}$ with f convex, $\mathcal{C} \subset \overline{\mathcal{D}}$ convex ($\overline{\mathcal{D}}$ is closure of $\mathcal{D}$), $\mathcal{C} \cap \mathcal{D} \neq \emptyset$

Projected Mirror Descent

$$\boxed{\begin{aligned} \nabla\Phi(\tilde{x}_{t+1}) &= \nabla\Phi(x_t) - \eta_t g_t, \text{where } g_t \in \partial f(x_t) \\ x_{t+1} &= \Pi_{\mathcal{C}}^{\Phi}(\tilde{x}_{t+1}) \end{aligned}}$$



Mirror Maps, Bregman Divergence, Bregman Projection

Mirror map (Definition 10.5)

$\Phi\colon \mathcal{D} \subseteq \mathbb{R}^d \to \mathbb{R}$ is a *mirror map* if:

i) Φ is strictly convex and differentiable

ii) The gradient $\nabla\Phi\colon \mathcal{D} \to \mathbb{R}^d$ is a surjective map

iii) The gradient diverges on the boundary of $\mathcal{D}$: $\lim_{x\to\partial\mathcal{D}} \|\nabla\Phi(x)\| = \infty$

Bregman divergence (Definition 10.6)

The Bregman divergence associated with a differentiable $\Phi : \mathbb{R}^d \to \mathbb{R}$ is

$$\boxed{D^{\Phi}(x,y) = \Phi(x) - \Phi(y) - \nabla\Phi(y)^{\top}(x-y)}$$

Bregman projection (Definition 10.7)

The *Bregman projection* associated to a mirror map Φ is given by

$$\boxed{\Pi_{\mathcal{C}}^{\Phi}(y) = \operatorname*{argmin}_{x\in\mathcal{C}\cap\mathcal{D}} D^{\Phi}(x,y)}$$

Euclidean Balls $\Rightarrow$ Gradient Descent

- $\mathcal{C} \qquad \mathcal{D} = \mathbb{R}^d$
- Mirror map: $\Phi(x) = \frac{1}{2}\|x\|_2^2$
- $\nabla\Phi(x) = x$
- Bregman divergence:

$$\begin{aligned} D^{\Phi}(x,y) &= \Phi(x) - \Phi(y) - \nabla\Phi(y)^{\top}(x-y) \\ &= \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|y\|_2^2 - y^{\top}x + y^{\top}y \\ &= \frac{1}{2}\|x-y\|_2^2 \end{aligned}$$

- Projection:

$$\Pi_{\mathcal{C}}^{\Phi}(y) = \operatorname*{argmin}_{x\in\mathcal{C}\cap\mathcal{D}} D^{\Phi}(y,x) = \operatorname*{argmin}_{x\in\mathcal{C}} \|x-y\|_2^2 \equiv \Pi_{\mathcal{C}}(y)$$

We recover the projected subgradient descent algorithm

Negative Entropy $\Rightarrow$ Exponential Gradient Descent

- $\mathcal{C} = \Delta_d \qquad \mathcal{D} = \{x \in \mathbb{R}^d \colon x_i > 0, i = 1, \ldots, d\}$
- Mirror map: $\Phi(x) = \sum_{i=1}^d x_i \log x_i$ (negative entropy)
- $\nabla\Phi(x) = 1 + \log(x)$
- Bregman divergence: $D^\Phi(x, y) = \sum_{i=1}^d x_i \log\left(\frac{x_i}{y_i}\right)$
- Projection: $\Pi_{\mathcal{C}}^{\Phi}(y) = \operatorname{argmin}_{x \in \mathcal{C} \cap \mathcal{D}} D^\Phi(y, x) = \frac{y}{\|y\|_1}$

$$\boxed{\begin{aligned} &\log(\tilde{x}_{t+1}) = \log(x_t) - \eta g_t \qquad \Longleftrightarrow \qquad \tilde{x}_{t+1} = x_t e^{-\eta g_t} \\ &x_{t+1} = \frac{\tilde{x}_{t+1}}{\|\tilde{x}_{t+1}\|_1} \end{aligned}}$$

Decomposition via Bregman Divergences

Property (Proposition 10.9)

For any differentiable function $\Phi : \mathbb{R}^d \to \mathbb{R}$ we have

$$\boxed{(\nabla \Phi(x) - \nabla \Phi(y))^\top (x - z) = D^\Phi(x, y) + D^\Phi(z, x) - D^\Phi(z, y)}$$

Analogous to the Euclidean decomposition

$2a^\top b = \|a\|_2^2 + \|b\|_2^2 - \|a - b\|_2^2$

Non-Expansivity of Projections

Non-expansivity (Proposition 10.10)

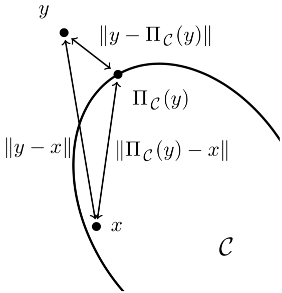
Let $x \in \mathcal{C} \cap \mathcal{D}$ and $y \in \mathcal{D}$. Then,

$$\left(\nabla\Phi\left(\Pi_{\mathcal{C}}^{\Phi}(y)\right) - \nabla\Phi(y)\right)^{\top}\left(\Pi_{\mathcal{C}}^{\Phi}(y) - x\right) \leq 0$$

which implies $D^{\Phi}(x, \Pi_{\mathcal{C}}^{\Phi}(y)) + D^{\Phi}(\Pi_{\mathcal{C}}^{\Phi}(y), y) \leq D^{\Phi}(x, y)$ and, in particular,

$$\boxed{D^{\Phi}(x, \Pi_{\mathcal{C}}^{\Phi}(y)) \leq D^{\Phi}(x, y)}$$

Analogous property as for Euclidean projections:



Results for Lipschitz Functions

f is γ-**Lipschitz on** $\mathcal{C}$ w.r.t. norm $\|\cdot\|$ if $\exists\gamma > 0$ such that (equivalent)

- For every $x, y \in \mathcal{C}$, $|f(y) - f(x)| \leq \gamma\|x - y\|$
- For every $x \in \mathcal{C}$, any subgradient $g \in \partial f(x)$ satisfies $\|g\|_* \leq \gamma$, where

$$\boxed{\|g\|_* := \sup\left\{|g^\top x|; x \in \mathbb{R}^d, \|x\| = 1\right\}} \qquad \textbf{(dual norm)}$$

Projected Mirror Descent—Lipschitz (Theorem 10.11)

- Function f is γ-Lipschitz w.r.t. the norm $\|\cdot\|$
- Mirror map Φ is α-strongly convex on $\mathcal{C} \cap \mathcal{D}$ w.r.t. the norm $\|\cdot\|$
- Initial condition is $x_1 \in \operatorname{argmin}_{x \in \mathcal{C}\cap\mathcal{D}} \Phi(x)$
- Assume $c^2 = \sup_{x \in \mathcal{C}\cap\mathcal{D}} \Phi(x) - \Phi(x_1)$

Then, projected mirror descent with $\eta_s \equiv \eta = \frac{c}{\gamma}\sqrt{\frac{2\alpha}{t}}$ satisfies

$$\boxed{f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^\star) \leq c\gamma\sqrt{\frac{2}{\alpha t}}}$$

Proof of Theorem 10.11 (Part I)

- Using that $g_s = \frac{1}{\eta}(\nabla\Phi(x_s) - \nabla\Phi(\tilde{x}_{s+1}))$.:

$$\begin{aligned}
f(x_s) - f(x) \leq g_s^\top (x_s - x) &= \frac{1}{\eta}(\nabla\Phi(x_s) - \nabla\Phi(\tilde{x}_{s+1}))^\top (x_s - x) \\
&= \frac{1}{\eta}(D^\Phi(x_s, \tilde{x}_{s+1}) + D^\Phi(x, x_s) - D^\Phi(x, \tilde{x}_{s+1})) \\
&\leq \frac{1}{\eta}(D^\Phi(x_s, \tilde{x}_{s+1}) + D^\Phi(x, x_s) - D^\Phi(x, x_{s+1}))
\end{aligned}$$

- Strong conv. of Φ: $\Phi(\tilde{x}_{s+1}) \geq \Phi(x_s) + \nabla\Phi(x_s)^\top(\tilde{x}_{s+1} - x_s) + \frac{\alpha}{2}\|\tilde{x}_{s+1} - x_s\|^2$
- Lipschitz continuity of f: $\|g_s\|_* \leq \gamma$
- Using these two inequalities, along with Hölder's inequality, we obtain

$$\begin{aligned}
D^\Phi(x_s, \tilde{x}_{s+1}) &= \Phi(x_s) - \Phi(\tilde{x}_{s+1}) - \nabla\Phi(\tilde{x}_{s+1})^\top(x_s - \tilde{x}_{s+1}) \\
&\leq (\nabla\Phi(x_s) - \nabla\Phi(\tilde{x}_{s+1}))^\top(x_s - \tilde{x}_{s+1}) - \frac{\alpha}{2}\|\tilde{x}_{s+1} - x_s\|^2 \\
&= \eta g_s^\top(x_s - \tilde{x}_{s+1}) - \frac{\alpha}{2}\|\tilde{x}_{s+1} - x_s\|^2 \\
&\leq \eta\|g_s\|_*\|x_s - \tilde{x}_{s+1}\| - \frac{\alpha}{2}\|\tilde{x}_{s+1} - x_s\|^2 \\
&\leq \eta\gamma\|x_s - \tilde{x}_{s+1}\| - \frac{\alpha}{2}\|\tilde{x}_{s+1} - x_s\|^2 \leq \frac{\eta^2\gamma^2}{2\alpha}
\end{aligned}$$

where we used the inequality $az - bz^2 \leq \max_{z\in\mathbb{R}}(az - bz^2) = a^2/4b$ for all $z \in \mathbb{R}$.

Proof of Theorem 10.11 (Part II)

- By convexity, we finally obtain

$$
\begin{aligned}
f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^\star) &\leq \frac{1}{t}\sum_{s=1}^{t}(f(x_s) - f(x^\star)) \\
&\leq \frac{1}{\eta t}\sum_{s=1}^{t}(D^\Phi(x^\star, x_s) - D^\Phi(x^\star, x_{s+1})) + \frac{\eta\gamma^2}{2\alpha} \\
&= \frac{1}{\eta t}(D^\Phi(x^\star, x_1) - D^\Phi(x^\star, x_{t+1})) + \frac{\eta\gamma^2}{2\alpha} \\
&\leq \frac{D^\Phi(x^\star, x_1)}{\eta t} + \frac{\eta\gamma^2}{2\alpha},
\end{aligned}
$$

  as $D^\Phi(x, x_{s+1}) \geq 0$.

- The proof follows by optimizing the bound over η, and using that

$$
D^\Phi(x^\star, x_1) = \Phi(x^\star) - \Phi(x_1) - \nabla\Phi(x_1)^\top(x^\star - x_1) \leq \sup_{x \in \mathcal{C}\cap\mathcal{D}} \Phi(x) - \Phi(x_1),
$$

  where we used the optimality condition in Proposition 8.10 to claim that

$$
\nabla\Phi(x_1)^\top(x^\star - x_1) \geq 0
$$

  as $x_1 \in \operatorname{argmin}_{x \in \mathcal{C}\cap\mathcal{D}} \Phi(x)$ by assumption.

Back to Learning: Boosting

- $\mathcal{C} = \Delta_d \qquad \mathcal{D} = \{x \in \mathbb{R}^d \colon x_i > 0, i = 1, \ldots, d\}$
- Mirror map: $\Phi(w) = \sum_{i=1}^d w_i \log w_i$ (negative entropy)
- Starting point $w_1 \in \operatorname{argmin}_{w \in \mathcal{C} \cap \mathcal{D}} \Phi(w) = \frac{1}{d}\mathbb{1}$
- As $\Phi(w) \leq 0$, we have $c^2 = \sup_{w \in \mathcal{C} \cap \mathcal{D}} \Phi(w) - \Phi(w_1) = \log d$
- Φ is α-strongly convex with respect to the $\| \cdot \|_1$ norm, with $\alpha = 1$ (consequence of Pinsker's inequality)
- R is γ-Lipschitz with respect to the $\| \cdot \|_1$ norm, with $\gamma = \gamma_\varphi c_\infty^{\mathcal{X}}$ (Hölder's)

$$|R(w) - R(u)| \leq \gamma_\varphi c_\infty^{\mathcal{X}} \|w - u\|_1$$

If $\eta = \frac{c}{\gamma}\sqrt{\frac{2\alpha}{t}} = \frac{1}{\gamma_\varphi c_\infty^{\mathcal{X}}}\sqrt{\frac{2 \log d}{t}}$, we have (recall $c_1^{\mathcal{W}} = 1$)

$$\boxed{\texttt{Optimization}_\Delta := R(\overline{W}_t) - R(W_\Delta^\star) \leq c\gamma\sqrt{\frac{2}{\alpha t}} = c_\infty^{\mathcal{X}} c_1^{\mathcal{W}} \gamma_\varphi \sqrt{\frac{2 \log d}{t}}}$$

Principled approach: Enough to run algorithm for $t \sim n$ time steps (ONLY BASED ON UPPER BOUNDS!)