All ERMs Can Fail in Stochastic Convex Optimization

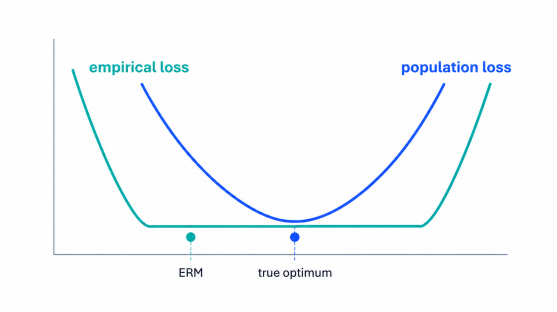
Lower Bounds in Linear Dimension

Tal Burla Roi Livni
Tel Aviv University • ICML 2026

All ERMs Can Fail in Stochastic Convex Optimization

Lower Bounds in Linear Dimension

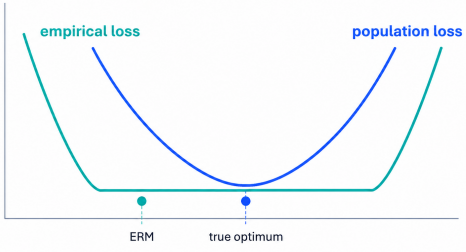
Tal Burla Roi Livni
Tel Aviv University • ICML 2026



All ERMs Can Fail in Stochastic Convex Optimization

Lower Bounds in Linear Dimension

Tal Burla Roi Livni
Tel Aviv University • ICML 2026



A simple but sufficient model: Stochastic Convex Optimization

SCO Setting

Instance space $\mathcal{Z}$

SCO Setting

Instance space $\mathcal{Z}$

Parameter space $\mathcal{W}_d = \{w \in \mathbb{R}^d : \|w\|_2 \leq 1\}$

SCO Setting

Instance space	$\mathcal{Z}$
Parameter space	$\mathcal{W}_d = \{w \in \mathbb{R}^d : \|w\|_2 \leq 1\}$
Loss	$f : \mathcal{W}_d \times \mathcal{Z} \to \mathbb{R}$

SCO Setting

Instance space $\mathcal{Z}$

Parameter space $\mathcal{W}_d = \{w \in \mathbb{R}^d : \|w\|_2 \leq 1\}$

Loss $f : \mathcal{W}_d \times \mathcal{Z} \to \mathbb{R}$

Convex $$f(\theta w_1 + (1-\theta)w_2, z) \leq \theta f(w_1, z) + (1-\theta) f(w_2, z)$$
$$\theta \in [0, 1]$$

SCO Setting

Instance space $\mathcal{Z}$

Parameter space $\mathcal{W}_d = \{w \in \mathbb{R}^d : \|w\|_2 \leq 1\}$

Loss $f : \mathcal{W}_d \times \mathcal{Z} \to \mathbb{R}$

Convex $f(\theta w_1 + (1-\theta)w_2, z) \leq \theta f(w_1, z) + (1-\theta) f(w_2, z)$
$\theta \in [0, 1]$

L-Lipschitz $|f(w, z) - f(w', z)| \leq L \|w - w'\|_2$

Learning

Data

D over $\mathcal{Z}$, $\quad S = (z_1, \ldots, z_m) \sim D^m$

Learning

Data

$$D \text{ over } \mathcal{Z}, \qquad S = (z_1, \ldots, z_m) \sim D^m$$

Algorithm

$$\mathcal{A}(S) = w_S \in \mathcal{W}_d$$

Learning

Data

$$D \text{ over } \mathcal{Z}, \qquad S = (z_1, \ldots, z_m) \sim D^m$$

Algorithm

$$\mathcal{A}(S) = w_S \in \mathcal{W}_d$$

Goal

$$F(w) = \mathbb{E}_{z \sim D}[f(w, z)], \qquad \min_{w \in \mathcal{W}_d} F(w)$$

Learning

Data

$$D \text{ over } \mathcal{Z}, \qquad S = (z_1, \ldots, z_m) \sim D^m$$

Algorithm

$$\mathcal{A}(S) = w_S \in \mathcal{W}_d$$

Goal

$$F(w) = \mathbb{E}_{z \sim D}[f(w, z)], \qquad \min_{w \in \mathcal{W}_d} F(w)$$

Empirical loss

$$F_S(w) = \frac{1}{m} \sum_{i=1}^{m} f(w, z_i)$$

Learning

Data

$$D \text{ over } \mathcal{Z}, \qquad S = (z_1, \ldots, z_m) \sim D^m$$

Algorithm

$$\mathcal{A}(S) = w_S \in \mathcal{W}_d$$

Goal

$$F(w) = \mathbb{E}_{z \sim D}[f(w, z)], \qquad \min_{w \in \mathcal{W}_d} F(w)$$

Empirical loss

$$F_S(w) = \frac{1}{m} \sum_{i=1}^{m} f(w, z_i)$$

ε-ERM

$$F_S(w_S) - \min_{w \in \mathcal{W}_d} F_S(w) \leq \varepsilon \qquad (\varepsilon = 0 : \text{ ERM})$$

ERM Can Fail

There exists

$$f : \mathcal{W}_{6m} \times \mathcal{Z} \to \mathbb{R} \quad \text{convex, } 1\text{-Lipschitz}$$

ERM Can Fail

There exists

$$f : \mathcal{W}_{6m} \times \mathcal{Z} \to \mathbb{R} \quad \text{convex, } 1\text{-Lipschitz}$$

Exact ERM

$$F(w_S) - \min_{w \in \mathcal{W}_d} F(w) \geq \Omega(1) \quad \textbf{New!}$$

ERM Can Fail

There exists

$$f : \mathcal{W}_{6m} \times \mathcal{Z} \to \mathbb{R} \quad \text{convex, } 1\text{-Lipschitz}$$

Exact ERM

$$F(w_S) - \min_{w \in \mathcal{W}_d} F(w) \geq \Omega(1) \quad \textbf{New!}$$

Every ε-ERM fails too

$$\varepsilon = \Theta(m^{-3/2})$$

But the problem is still learnable.

Strong Convexity

λ-strong convexity

$$g \in \partial f(w, z), \qquad f(w', z) \geq f(w, z) + \langle g, w' - w \rangle + \frac{\lambda}{2} \|w' - w\|_2^2$$

Strong Convexity

λ-strong convexity

$$g \in \partial f(w, z), \qquad f(w', z) \geq f(w, z) + \langle g, w' - w \rangle + \frac{\lambda}{2} \|w' - w\|_2^2$$

Known upper bound for ERM

$$F(w_S) - \min_{w \in \mathcal{W}_d} F(w) \leq O\left(\frac{1}{\lambda m}\right)$$

Strong Convexity

λ-strong convexity

$$g \in \partial f(w,z), \qquad f(w',z) \geq f(w,z) + \langle g, w' - w \rangle + \frac{\lambda}{2} \|w' - w\|_2^2$$

Known upper bound for ERM

$$F(w_S) - \min_{w \in \mathcal{W}_d} F(w) \leq O\left(\frac{1}{\lambda m}\right)$$

Our refined lower bound

$$m^{-3/2} \leq \lambda \leq m^{-1/2}, \qquad \varepsilon = \Theta\left(\frac{1}{\lambda m^3}\right)$$

$$F(w_S) - \min_{w \in \mathcal{W}_d} F(w) \geq \Omega\left(\frac{1}{\lambda m^{3/2}}\right) \quad \textbf{New!}$$

Projected GD

Update and output

$$w_{t+1} = \Pi_{\mathcal{W}_d}(w_t - \eta\, g_t)\,, \qquad g_t \in \partial F_S(w_t)$$

Projected GD

Update and output

$$w_{t+1} = \Pi_{\mathcal{W}_d}(w_t - \eta\, g_t)\,, \qquad g_t \in \partial F_S(w_t)$$

$$w_S^{\mathrm{GD}} = \frac{1}{T}\sum_{t=1}^{T} w_t$$

Projected GD

Update and output

$$w_{t+1} = \Pi_{\mathcal{W}_d}(w_t - \eta\, g_t)\,, \qquad g_t \in \partial F_S(w_t)$$

$$w_S^{\mathrm{GD}} = \frac{1}{T}\sum_{t=1}^{T} w_t$$

GD as an ε-ERM

$$F_S(w_S^{\mathrm{GD}}) - \min_{w\in\mathcal{W}_d} F_S(w) = O\left(\eta + \frac{1}{\eta T}\right)$$

$$\eta T \gtrsim m^{3/2} \quad \Longrightarrow \quad F(w_S^{\mathrm{GD}}) - \min_{w\in\mathcal{W}_d} F(w) \geq \Omega(1) \quad \textbf{New!}$$

Projected GD

Update and output

$$w_{t+1} = \Pi_{\mathcal{W}_d}(w_t - \eta\, g_t)\,, \qquad g_t \in \partial F_S(w_t)$$

$$w_S^{\mathrm{GD}} = \frac{1}{T}\sum_{t=1}^{T} w_t$$

GD as an ε-ERM

$$F_S(w_S^{\mathrm{GD}}) - \min_{w\in\mathcal{W}_d} F_S(w) = O\left(\eta + \frac{1}{\eta T}\right)$$

$$\eta T \gtrsim m^{3/2} \quad \Longrightarrow \quad F(w_S^{\mathrm{GD}}) - \min_{w\in\mathcal{W}_d} F(w) \geq \Omega(1) \quad \textbf{New!}$$

Known upper bound and learning

$$F(w_S^{\mathrm{GD}}) - \min_{w\in\mathcal{W}_d} F(w) \leq O\left(\eta\sqrt{T} + \frac{\eta T}{m}\right)$$

$$T = m^2, \quad \eta = m^{-3/2} \quad \Longrightarrow \quad O(m^{-1/2})$$

GD Lower Bound

There exists a convex 1-Lipschitz loss

$$F(w_S^{\mathrm{GD}}) - \min_{w \in \mathcal{W}_d} F(w) \geq \Omega\left(\min\left\{\sqrt{\frac{\eta T}{m^{3/2}}}, 1\right\}\right)$$

New!

GD Lower Bound

There exists a convex 1-Lipschitz loss

$$F(w_S^{\mathrm{GD}}) - \min_{w \in \mathcal{W}_d} F(w) \geq \Omega\left(\min\left\{\sqrt{\frac{\eta T}{m^{3/2}}}, 1\right\}\right)$$ **New!**

Together with the known $\Omega(\eta\sqrt{T})$ lower bound

$$F(w_S^{\mathrm{GD}}) - \min_{w \in \mathcal{W}_d} F(w) \geq \Omega\left(\eta\sqrt{T} + \sqrt{\frac{\eta T}{m^{3/2}}}\right)$$

Open Directions

Gradient descent

$$\Omega\left(\sqrt{\frac{\eta T}{m^{3/2}}}\right) \quad \text{vs.} \quad O\left(\frac{\eta T}{m}\right)$$

Open Directions

Gradient descent

$$\Omega\left(\sqrt{\frac{\eta T}{m^{3/2}}}\right) \quad \text{vs.} \quad O\left(\frac{\eta T}{m}\right)$$

Strongly convex objectives

$$\Omega\left(\frac{1}{\lambda m^{3/2}}\right) \quad \text{vs.} \quad O\left(\frac{1}{\lambda m}\right)$$

Open Directions

Gradient descent

$$\Omega\left(\sqrt{\frac{\eta T}{m^{3/2}}}\right) \quad \text{vs.} \quad O\left(\frac{\eta T}{m}\right)$$

Strongly convex objectives

$$\Omega\left(\frac{1}{\lambda m^{3/2}}\right) \quad \text{vs.} \quad O\left(\frac{1}{\lambda m}\right)$$

Smooth objectives

Can smooth losses exhibit the same phenomenon?

Thank you

All ERMs Can Fail in Stochastic Convex Optimization

Lower Bounds in Linear Dimension

Tal Burla Roi Livni