

# Gradient Descent. Non-smooth case. Linear Least squares with $l_1$ -regularization.

Daniil Merkulov

Optimization methods. MIPT

## Previously

- Gradient Descent. Convergence for strongly convex quadratic function. Optimal hyperparameters.

$$\alpha = \frac{2}{\mu + L} \quad \kappa = \frac{L}{\mu} \geq 1 \quad \rho = \frac{\kappa - 1}{\kappa + 1}$$

$$\|x_k - x^*\| \leq \rho^k \|x_0 - x^*\|$$

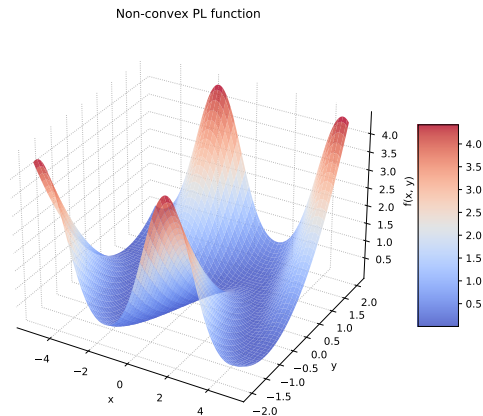


Figure 1: PL function

## Previously

- Gradient Descent. Convergence for strongly convex quadratic function. Optimal hyperparameters.

$$\alpha = \frac{2}{\mu + L} \quad \kappa = \frac{L}{\mu} \geq 1 \quad \rho = \frac{\kappa - 1}{\kappa + 1}$$

$$\|x_k - x^*\| \leq \rho^k \|x_0 - x^*\|$$

- Gradient Descent. Smooth convex case convergence.

$$f(x_k) - f^* \leq \frac{L \|x_0 - x^*\|^2}{2k}.$$

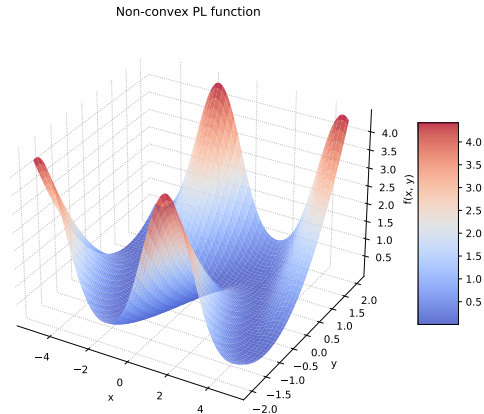


Figure 1: PL function

## Previously

- Gradient Descent. Convergence for strongly convex quadratic function. Optimal hyperparameters.

$$\alpha = \frac{2}{\mu + L} \quad \kappa = \frac{L}{\mu} \geq 1 \quad \rho = \frac{\kappa - 1}{\kappa + 1}$$

$$\|x_k - x^*\| \leq \rho^k \|x_0 - x^*\|$$

- Gradient Descent. Smooth convex case convergence.

$$f(x_k) - f^* \leq \frac{L \|x_0 - x^*\|^2}{2k}.$$

- Gradient Descent. Smooth PL case convergence.

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*).$$

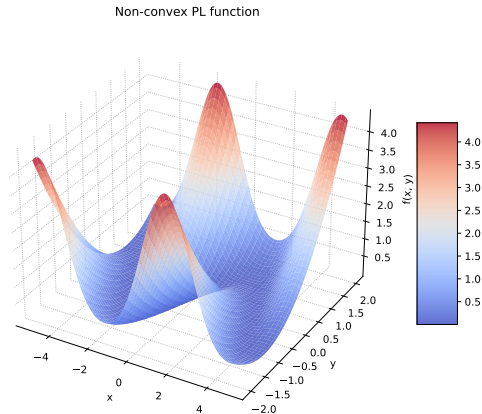


Figure 1: PL function

# Any $\mu$ -strongly convex differentiable function is a PL-function

## Theorem

If a function  $f(x)$  is differentiable and  $\mu$ -strongly convex, then it is a PL-function.

## Proof

By first order strong convexity criterion:

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

Putting  $y = x^*$ :

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

# Any $\mu$ -strongly convex differentiable function is a PL-function

## Theorem

If a function  $f(x)$  is differentiable and  $\mu$ -strongly convex, then it is a PL-function.

## Proof

By first order strong convexity criterion:

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

Putting  $y = x^*$ :

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

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

# Any $\mu$ -strongly convex differentiable function is a PL-function

## Theorem

If a function  $f(x)$  is differentiable and  $\mu$ -strongly convex, then it is a PL-function.

## Proof

By first order strong convexity criterion:

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

Putting  $y = x^*$ :

$$\begin{aligned} f(x^*) &\geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\mu}{2} \|x^* - x\|_2^2 \\ f(x) - f(x^*) &\leq \nabla f(x)^T (x - x^*) - \frac{\mu}{2} \|x^* - x\|_2^2 = \\ &= \left( \nabla f(x)^T - \frac{\mu}{2} (x^* - x) \right)^T (x - x^*) = \end{aligned}$$

# Any $\mu$ -strongly convex differentiable function is a PL-function

## Theorem

If a function  $f(x)$  is differentiable and  $\mu$ -strongly convex, then it is a PL-function.

## Proof

By first order strong convexity criterion:

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

Putting  $y = x^*$ :

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

$$\begin{aligned} f(x) - f(x^*) &\leq \nabla f(x)^T (x - x^*) - \frac{\mu}{2} \|x^* - x\|_2^2 = \\ &= \left( \nabla f(x)^T - \frac{\mu}{2} (x^* - x) \right)^T (x - x^*) = \\ &= \frac{1}{2} \left( \frac{2}{\sqrt{\mu}} \nabla f(x)^T - \sqrt{\mu} (x^* - x) \right)^T \sqrt{\mu} (x - x^*) = \end{aligned}$$



# Any $\mu$ -strongly convex differentiable function is a PL-function

## Theorem

If a function  $f(x)$  is differentiable and  $\mu$ -strongly convex, then it is a PL-function.

## Proof

By first order strong convexity criterion:

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

Putting  $y = x^*$ :

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

$$\begin{aligned} f(x) - f(x^*) &\leq \nabla f(x)^T (x - x^*) - \frac{\mu}{2} \|x^* - x\|_2^2 = \\ &= \left( \nabla f(x)^T - \frac{\mu}{2} (x^* - x) \right)^T (x - x^*) = \\ &= \frac{1}{2} \left( \frac{2}{\sqrt{\mu}} \nabla f(x)^T - \sqrt{\mu} (x^* - x) \right)^T \sqrt{\mu} (x - x^*) = \end{aligned}$$

# Any $\mu$ -strongly convex differentiable function is a PL-function

## Theorem

If a function  $f(x)$  is differentiable and  $\mu$ -strongly convex, then it is a PL-function.

## Proof

By first order strong convexity criterion:

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

Putting  $y = x^*$ :

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

$$\begin{aligned} f(x) - f(x^*) &\leq \nabla f(x)^T (x - x^*) - \frac{\mu}{2} \|x^* - x\|_2^2 = \\ &= \left( \nabla f(x)^T - \frac{\mu}{2} (x^* - x) \right)^T (x - x^*) = \\ &= \frac{1}{2} \left( \frac{2}{\sqrt{\mu}} \nabla f(x)^T - \sqrt{\mu} (x^* - x) \right)^T \sqrt{\mu} (x - x^*) = \end{aligned}$$

$$\begin{aligned} \text{Let } a &= \frac{1}{\sqrt{\mu}} \nabla f(x) \text{ and} \\ b &= \sqrt{\mu} (x - x^*) - \frac{1}{\sqrt{\mu}} \nabla f(x) \end{aligned}$$

# Any $\mu$ -strongly convex differentiable function is a PL-function

## Theorem

If a function  $f(x)$  is differentiable and  $\mu$ -strongly convex, then it is a PL-function.

## Proof

By first order strong convexity criterion:

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

Putting  $y = x^*$ :

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

$$\begin{aligned} f(x) - f(x^*) &\leq \nabla f(x)^T (x - x^*) - \frac{\mu}{2} \|x^* - x\|_2^2 = \\ &= \left( \nabla f(x)^T - \frac{\mu}{2} (x^* - x) \right)^T (x - x^*) = \\ &= \frac{1}{2} \left( \frac{2}{\sqrt{\mu}} \nabla f(x)^T - \sqrt{\mu} (x^* - x) \right)^T \sqrt{\mu} (x - x^*) = \end{aligned}$$

$$\begin{aligned} \text{Let } a &= \frac{1}{\sqrt{\mu}} \nabla f(x) \text{ and} \\ b &= \sqrt{\mu} (x - x^*) - \frac{1}{\sqrt{\mu}} \nabla f(x) \\ \text{Then } a + b &= \sqrt{\mu} (x - x^*) \text{ and} \\ a - b &= \frac{2}{\sqrt{\mu}} \nabla f(x) - \sqrt{\mu} (x - x^*) \end{aligned}$$

## Any $\mu$ -strongly convex differentiable function is a PL-function

$$f(x) - f(x^*) \leq \frac{1}{2} \left( \frac{1}{\mu} \|\nabla f(x)\|_2^2 - \left\| \sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}} \nabla f(x) \right\|_2^2 \right)$$

which is exactly PL-condition. It means, that we already have linear convergence proof for any strongly convex function.

## Any $\mu$ -strongly convex differentiable function is a PL-function

$$f(x) - f(x^*) \leq \frac{1}{2} \left( \frac{1}{\mu} \|\nabla f(x)\|_2^2 - \left\| \sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}} \nabla f(x) \right\|_2^2 \right)$$
$$f(x) - f(x^*) \leq \frac{1}{2\mu} \|\nabla f(x)\|_2^2,$$

which is exactly PL-condition. It means, that we already have linear convergence proof for any strongly convex function.

# Non-smooth optimization

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

A classical convex optimization problem is considered. We assume that  $f(x)$  is a convex function, but now we do not require smoothness.



Figure 2: Norm cones for different  $p$  - norms are non-smooth

# Non-smooth optimization

Wolfe's example



Figure 3: Wolfe's example. [Open in Colab](#)

# Algorithm

A vector  $g$  is called the **subgradient** of the function  $f(x) : S \rightarrow \mathbb{R}$  at the point  $x_0$  if  $\forall x \in S$ :

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle$$



# Algorithm

A vector  $g$  is called the **subgradient** of the function  $f(x) : S \rightarrow \mathbb{R}$  at the point  $x_0$  if  $\forall x \in S$ :

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle$$

The idea is very simple: let's replace the gradient  $\nabla f(x_k)$  in the gradient descent algorithm with a subgradient  $g_k$  at point  $x_k$ :

$$x_{k+1} = x_k - \alpha_k g_k, \tag{SD}$$

where  $g_k$  is an arbitrary subgradient of the function  $f(x)$  at the point  $x_k$ ,  $g_k \in \partial f(x_k)$

# Convergence bound

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^* - \alpha_k g_k\|^2 =$$

## Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle\end{aligned}$$

## Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ 2\alpha_k \langle g_k, x_k - x^* \rangle &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - \|x_{k+1} - x^*\|^2\end{aligned}$$

## Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ 2\alpha_k \langle g_k, x_k - x^* \rangle &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - \|x_{k+1} - x^*\|^2\end{aligned}$$

## Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ 2\alpha_k \langle g_k, x_k - x^* \rangle &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - \|x_{k+1} - x^*\|^2\end{aligned}$$

Let us sum the obtained equality for  $k = 0, \dots, T - 1$ :

## Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ 2\alpha_k \langle g_k, x_k - x^* \rangle &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - \|x_{k+1} - x^*\|^2\end{aligned}$$

Let us sum the obtained equality for  $k = 0, \dots, T-1$ :

$$\sum_{k=0}^{T-1} 2\alpha_k \langle g_k, x_k - x^* \rangle = \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2$$

## Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ 2\alpha_k \langle g_k, x_k - x^* \rangle &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - \|x_{k+1} - x^*\|^2\end{aligned}$$

Let us sum the obtained equality for  $k = 0, \dots, T-1$ :

$$\begin{aligned}\sum_{k=0}^{T-1} 2\alpha_k \langle g_k, x_k - x^* \rangle &= \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq \|x_0 - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2\end{aligned}$$



## Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ 2\alpha_k \langle g_k, x_k - x^* \rangle &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - \|x_{k+1} - x^*\|^2\end{aligned}$$

Let us sum the obtained equality for  $k = 0, \dots, T-1$ :

$$\begin{aligned}\sum_{k=0}^{T-1} 2\alpha_k \langle g_k, x_k - x^* \rangle &= \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq \|x_0 - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq R^2 + G^2 \sum_{k=0}^{T-1} \alpha_k^2\end{aligned}$$

## Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ 2\alpha_k \langle g_k, x_k - x^* \rangle &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - \|x_{k+1} - x^*\|^2\end{aligned}$$

Let us sum the obtained equality for  $k = 0, \dots, T-1$ :

$$\begin{aligned}\sum_{k=0}^{T-1} 2\alpha_k \langle g_k, x_k - x^* \rangle &= \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq \|x_0 - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq R^2 + G^2 \sum_{k=0}^{T-1} \alpha_k^2\end{aligned}$$

- Let's write down how close we came to the optimum  $x^* = \arg \min_{x \in \mathbb{R}^n} f(x) = \arg f^*$  on the last iteration:

## Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ 2\alpha_k \langle g_k, x_k - x^* \rangle &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - \|x_{k+1} - x^*\|^2\end{aligned}$$

Let us sum the obtained equality for  $k = 0, \dots, T-1$ :

$$\begin{aligned}\sum_{k=0}^{T-1} 2\alpha_k \langle g_k, x_k - x^* \rangle &= \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq \|x_0 - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq R^2 + G^2 \sum_{k=0}^{T-1} \alpha_k^2\end{aligned}$$

- Let's write down how close we came to the optimum  $x^* = \arg \min_{x \in \mathbb{R}^n} f(x) = \arg f^*$  on the last iteration:
- For a subgradient:  $\langle g_k, x_k - x^* \rangle \leq f(x_k) - f(x^*) = f(x_k) - f^*$ .

# Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ 2\alpha_k \langle g_k, x_k - x^* \rangle &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - \|x_{k+1} - x^*\|^2\end{aligned}$$

Let us sum the obtained equality for  $k = 0, \dots, T-1$ :

$$\begin{aligned}\sum_{k=0}^{T-1} 2\alpha_k \langle g_k, x_k - x^* \rangle &= \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq \|x_0 - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq R^2 + G^2 \sum_{k=0}^{T-1} \alpha_k^2\end{aligned}$$

- Let's write down how close we came to the optimum  $x^* = \arg \min_{x \in \mathbb{R}^n} f(x) = \arg f^*$  on the last iteration:
- For a subgradient:  $\langle g_k, x_k - x^* \rangle \leq f(x_k) - f(x^*) = f(x_k) - f^*$ .
- We additionally assume, that  $\|g_k\|^2 \leq G^2$

# Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ 2\alpha_k \langle g_k, x_k - x^* \rangle &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - \|x_{k+1} - x^*\|^2\end{aligned}$$

Let us sum the obtained equality for  $k = 0, \dots, T-1$ :

$$\begin{aligned}\sum_{k=0}^{T-1} 2\alpha_k \langle g_k, x_k - x^* \rangle &= \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq \|x_0 - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq R^2 + G^2 \sum_{k=0}^{T-1} \alpha_k^2\end{aligned}$$

- Let's write down how close we came to the optimum  $x^* = \arg \min_{x \in \mathbb{R}^n} f(x) = \arg f^*$  on the last iteration:
- For a subgradient:  $\langle g_k, x_k - x^* \rangle \leq f(x_k) - f(x^*) = f(x_k) - f^*$ .
- We additionally assume, that  $\|g_k\|^2 \leq G^2$
- We use the notation  $R = \|x_0 - x^*\|_2$

## Convergence bound

Assuming  $\alpha_k = \alpha$  (constant stepsize), we have:

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq \frac{R^2}{2\alpha} + \frac{\alpha}{2} G^2 T$$

## Convergence bound

Assuming  $\alpha_k = \alpha$  (constant stepsize), we have:

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq \frac{R^2}{2\alpha} + \frac{\alpha}{2} G^2 T$$

Minimizing the right-hand side by  $\alpha$  gives  $\alpha^* = \frac{R}{G} \sqrt{\frac{1}{T}}$  and

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}.$$

## Convergence bound

Assuming  $\alpha_k = \alpha$  (constant stepsize), we have:

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq \frac{R^2}{2\alpha} + \frac{\alpha}{2} G^2 T$$

Minimizing the right-hand side by  $\alpha$  gives  $\alpha^* = \frac{R}{G} \sqrt{\frac{1}{T}}$  and

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}.$$

$$f(\bar{x}) - f^* = f\left(\frac{1}{T} \sum_{k=0}^{T-1} x_k\right) - f^* \leq \frac{1}{T} \left( \sum_{k=0}^{T-1} (f(x_k) - f^*) \right)$$



## Convergence bound

Assuming  $\alpha_k = \alpha$  (constant stepsize), we have:

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq \frac{R^2}{2\alpha} + \frac{\alpha}{2} G^2 T$$

Minimizing the right-hand side by  $\alpha$  gives  $\alpha^* = \frac{R}{G} \sqrt{\frac{1}{T}}$  and

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}.$$

$$\begin{aligned} f(\bar{x}) - f^* &= f\left(\frac{1}{T} \sum_{k=0}^{T-1} x_k\right) - f^* \leq \frac{1}{T} \left( \sum_{k=0}^{T-1} (f(x_k) - f^*) \right) \\ &\leq \frac{1}{T} \left( \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \right) \end{aligned}$$

## Convergence bound

Assuming  $\alpha_k = \alpha$  (constant stepsize), we have:

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq \frac{R^2}{2\alpha} + \frac{\alpha}{2} G^2 T$$

Minimizing the right-hand side by  $\alpha$  gives  $\alpha^* = \frac{R}{G} \sqrt{\frac{1}{T}}$  and

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}.$$

$$\begin{aligned} f(\bar{x}) - f^* &= f\left(\frac{1}{T} \sum_{k=0}^{T-1} x_k\right) - f^* \leq \frac{1}{T} \left( \sum_{k=0}^{T-1} (f(x_k) - f^*) \right) \\ &\leq \frac{1}{T} \left( \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \right) \\ &\leq GR \frac{1}{\sqrt{T}} \end{aligned}$$

## Convergence bound

Assuming  $\alpha_k = \alpha$  (constant stepsize), we have:

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq \frac{R^2}{2\alpha} + \frac{\alpha}{2} G^2 T$$

Minimizing the right-hand side by  $\alpha$  gives  $\alpha^* = \frac{R}{G} \sqrt{\frac{1}{T}}$  and

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}.$$

$$\begin{aligned} f(\bar{x}) - f^* &= f\left(\frac{1}{T} \sum_{k=0}^{T-1} x_k\right) - f^* \leq \frac{1}{T} \left( \sum_{k=0}^{T-1} (f(x_k) - f^*) \right) \\ &\leq \frac{1}{T} \left( \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \right) \\ &\leq GR \frac{1}{\sqrt{T}} \end{aligned}$$

## Convergence bound

Assuming  $\alpha_k = \alpha$  (constant stepsize), we have:

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq \frac{R^2}{2\alpha} + \frac{\alpha}{2} G^2 T$$

Minimizing the right-hand side by  $\alpha$  gives  $\alpha^* = \frac{R}{G} \sqrt{\frac{1}{T}}$  and

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}.$$

$$\begin{aligned} f(\bar{x}) - f^* &= f\left(\frac{1}{T} \sum_{k=0}^{T-1} x_k\right) - f^* \leq \frac{1}{T} \left( \sum_{k=0}^{T-1} (f(x_k) - f^*) \right) \\ &\leq \frac{1}{T} \left( \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \right) \\ &\leq GR \frac{1}{\sqrt{T}} \end{aligned}$$

Important notes:

- Obtaining bounds not for  $x_T$  but for the arithmetic mean over iterations  $\bar{x}$  is a typical trick in obtaining estimates for methods where there is convexity but no monotonic decreasing at each iteration. There is no guarantee of success at each iteration, but there is a guarantee of success on average

## Convergence bound

Assuming  $\alpha_k = \alpha$  (constant stepsize), we have:

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq \frac{R^2}{2\alpha} + \frac{\alpha}{2} G^2 T$$

Minimizing the right-hand side by  $\alpha$  gives  $\alpha^* = \frac{R}{G} \sqrt{\frac{1}{T}}$  and

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}.$$

$$\begin{aligned} f(\bar{x}) - f^* &= f\left(\frac{1}{T} \sum_{k=0}^{T-1} x_k\right) - f^* \leq \frac{1}{T} \left( \sum_{k=0}^{T-1} (f(x_k) - f^*) \right) \\ &\leq \frac{1}{T} \left( \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \right) \\ &\leq GR \frac{1}{\sqrt{T}} \end{aligned}$$

Important notes:

- Obtaining bounds not for  $x_T$  but for the arithmetic mean over iterations  $\bar{x}$  is a typical trick in obtaining estimates for methods where there is convexity but no monotonic decreasing at each iteration. There is no guarantee of success at each iteration, but there is a guarantee of success on average
- To choose the optimal step, we need to know (assume) the number of iterations in advance. Possible solution: initialize  $T$  with a small value, after reaching this number of iterations double  $T$  and restart the algorithm. A more intelligent way: adaptive selection of stepsize.

# Steepest subgradient descent convergence bound

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^* - \alpha_k g_k\|^2 =$$

# Steepest subgradient descent convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \doteq\end{aligned}$$

## Steepest subgradient descent convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \doteq \\ \alpha_k &= \frac{\langle g_k, x_k - x^* \rangle}{\|g_k\|^2} \text{ (from minimizing right hand side over stepsize)}\end{aligned}$$



## Steepest subgradient descent convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\&= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \doteq \\ \alpha_k &= \frac{\langle g_k, x_k - x^* \rangle}{\|g_k\|^2} \text{ (from minimizing right hand side over stepsize)} \\ &\doteq \|x_k - x^*\|^2 - \frac{\langle g_k, x_k - x^* \rangle^2}{\|g_k\|^2}\end{aligned}$$

## Steepest subgradient descent convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\&= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \stackrel{\circ}{=} \\&\alpha_k = \frac{\langle g_k, x_k - x^* \rangle}{\|g_k\|^2} \text{ (from minimizing right hand side over stepsize)} \\&\stackrel{\circ}{=} \|x_k - x^*\|^2 - \frac{\langle g_k, x_k - x^* \rangle^2}{\|g_k\|^2} \\ \langle g_k, x_k - x^* \rangle^2 &= \left( \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right) \|g_k\|^2 \leq \left( \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right) G^2\end{aligned}$$

## Steepest subgradient descent convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\&= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \doteq \\ \alpha_k &= \frac{\langle g_k, x_k - x^* \rangle}{\|g_k\|^2} \text{ (from minimizing right hand side over stepsize)} \\ &\doteq \|x_k - x^*\|^2 - \frac{\langle g_k, x_k - x^* \rangle^2}{\|g_k\|^2}\end{aligned}$$

$$\begin{aligned}\langle g_k, x_k - x^* \rangle^2 &= \left( \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right) \|g_k\|^2 \leq \left( \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right) G^2 \\ \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle^2 &\leq \sum_{k=0}^{T-1} \left( \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right) G^2 \leq \left( \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 \right) G^2\end{aligned}$$

## Steepest subgradient descent convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\&= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \doteq \\ \alpha_k &= \frac{\langle g_k, x_k - x^* \rangle}{\|g_k\|^2} \text{ (from minimizing right hand side over stepsize)} \\&\doteq \|x_k - x^*\|^2 - \frac{\langle g_k, x_k - x^* \rangle^2}{\|g_k\|^2}\end{aligned}$$

$$\langle g_k, x_k - x^* \rangle^2 = (\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2) \|g_k\|^2 \leq (\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2) G^2$$

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle^2 \leq \sum_{k=0}^{T-1} (\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2) G^2 \leq (\|x_0 - x^*\|^2 - \|x_T - x^*\|^2) G^2$$

$$\frac{1}{T} \left( \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \right)^2 \leq \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle^2 \leq R^2 G^2 \quad \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}$$

## Steepest subgradient descent convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\&= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \doteq \\ \alpha_k &= \frac{\langle g_k, x_k - x^* \rangle}{\|g_k\|^2} \text{ (from minimizing right hand side over stepsize)} \\ &\doteq \|x_k - x^*\|^2 - \frac{\langle g_k, x_k - x^* \rangle^2}{\|g_k\|^2}\end{aligned}$$

$$\langle g_k, x_k - x^* \rangle^2 = (\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2) \|g_k\|^2 \leq (\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2) G^2$$

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle^2 \leq \sum_{k=0}^{T-1} (\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2) G^2 \leq (\|x_0 - x^*\|^2 - \|x_T - x^*\|^2) G^2$$

$$\frac{1}{T} \left( \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \right)^2 \leq \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle^2 \leq R^2 G^2 \quad \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}$$

## Steepest subgradient descent convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\&= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \doteq \\ \alpha_k &= \frac{\langle g_k, x_k - x^* \rangle}{\|g_k\|^2} \text{ (from minimizing right hand side over stepsize)} \\&\doteq \|x_k - x^*\|^2 - \frac{\langle g_k, x_k - x^* \rangle^2}{\|g_k\|^2} \\ \langle g_k, x_k - x^* \rangle^2 &= (\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2) \|g_k\|^2 \leq (\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2) G^2 \\ \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle^2 &\leq \sum_{k=0}^{T-1} (\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2) G^2 \leq (\|x_0 - x^*\|^2 - \|x_T - x^*\|^2) G^2 \\ \frac{1}{T} \left( \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \right)^2 &\leq \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle^2 \leq R^2 G^2 \quad \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}\end{aligned}$$

Which leads to exactly the same bound of  $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$  on the primal gap. In fact, for this class of functions, you can't get a better result than  $\frac{1}{\sqrt{T}}$ .

# Linear Least Squares with $l_1$ -regularization

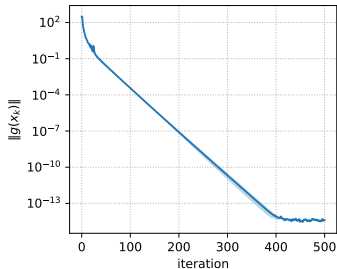
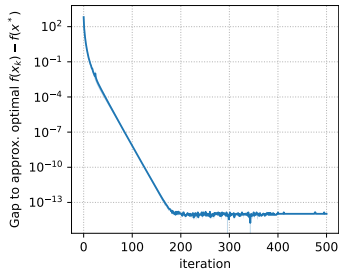
$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1$$

Algorithm will be written as:

$$x_{k+1} = x_k - \alpha_k \left( A^\top (Ax_k - b) + \lambda \text{sign}(x_k) \right)$$

where signum function is taken element-wise.

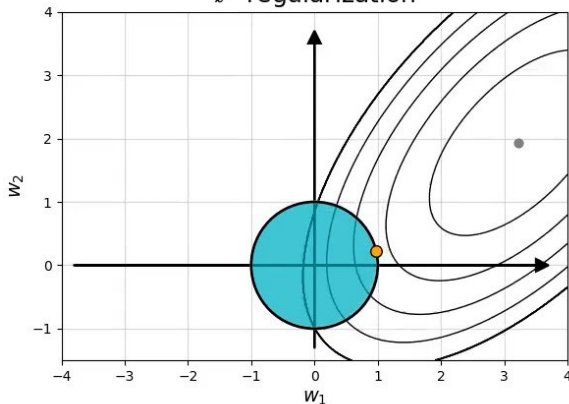
LLS with  $l_1$  regularization. 2 runs.  $\lambda = 1$



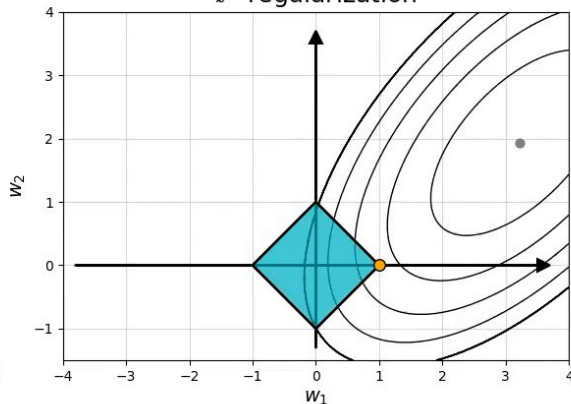
# Great illustration of $\ell_1$ -regularization

$\ell^1$  induces sparse solutions for least squares

$\ell^2$  regularization



$\ell^1$  regularization



by @itayevron



# Support Vector Machines

Let  $D = \{(x_i, y_i) \mid x_i \in \mathbb{R}^n, y_i \in \{\pm 1\}\}$

We need to find  $\omega \in \mathbb{R}^n$  and  $b \in \mathbb{R}$  such that

$$\min_{\omega \in \mathbb{R}^n, b \in \mathbb{R}} \frac{1}{2} \|\omega\|_2^2 + C \sum_{i=1}^m \max[0, 1 - y_i(\omega^\top x_i + b)]$$