SDSC6015 Stochastic Optimization for Machine Learning

Lu Yu

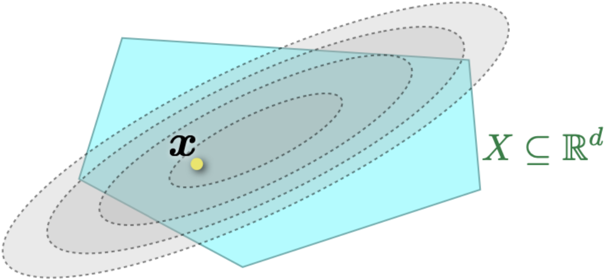
Department of Data Science, City University of Hong Kong

September 25, 2025

Projected Gradient Descent

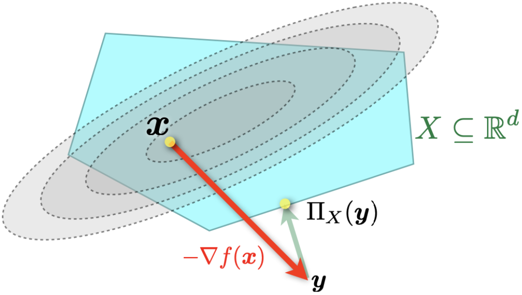
Constrained Optimization Problem

$$\begin{aligned} \text{minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{x} \in X \end{aligned}$$



Projected Gradient Descent

Idea: project onto X after every step: $\Pi_X(\boldsymbol{y}) := \arg\min_{\boldsymbol{x} \in X} \|\boldsymbol{x} - \boldsymbol{y}\|$



Projected gradient descent: $\boldsymbol{x}_{t+1} = \Pi_X[\boldsymbol{x}_t - \eta \nabla f(\boldsymbol{x}_t)]$

Projected Gradient Descent

Projected gradient descent:

$$
\begin{aligned}
\boldsymbol{y}_{t+1} &:= \boldsymbol{x}_t - \eta_t \nabla f(\boldsymbol{x}_t) \\
\boldsymbol{x}_{t+1} &:= \Pi_X(\boldsymbol{y}_{t+1}) = \operatorname*{arg\,min}_{\boldsymbol{x} \in X} \|\boldsymbol{x} - \boldsymbol{y}\|
\end{aligned}
$$

for stepsize $\eta_t > 0$ and timesteps $t = 0, 1, \ldots$

Convergence Rate of Projected Gradient Descent

The same number of steps as a gradient over $\mathbb{R}^d$!

- Lipschitz convex functions over X : $\mathcal{O}(1/\varepsilon^2)$ steps
- Smooth convex functions over X : $\mathcal{O}(1/\varepsilon)$ steps
- Smooth and strongly convex functions over X : $\mathcal{O}(\log(1/\varepsilon))$ steps

We will adapt the previous proofs for gradient descent.
BUT:

- Each step involves a projection onto X
- may or may not be efficient...

Smooth functions over X

Recall:

f is called smooth (with parameter L) over X if

$$f(\boldsymbol{y}) \leqslant f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in X.$$

Lemma 1

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and smooth with parameter L over X. Choosing stepsize

$$\eta = \frac{1}{L}\,,$$

projected gradient descent with arbitrary $\boldsymbol{x}_0 \in X$ satisfies

$$f(\boldsymbol{x}_{t+1}) \leqslant f(\boldsymbol{x}_t) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_t)\|^2 + \frac{L}{2}\|\boldsymbol{y}_{t+1} - \boldsymbol{x}_{t+1}\|^2, \qquad t \geqslant 0\,.$$

Projected Gradient Descent on Smooth Functions

Theorem 1

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. Let $X \subseteq \mathbb{R}^d$ be a closed convex set, and assume that there is a minimizer $\boldsymbol{x}^*$ of f over X; furthermore, suppose that f is smooth over X with parameter L. Choosing stepsize

$$\eta = \frac{1}{L}\,,$$

projected gradient descent yields

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^*) \leqslant \frac{L}{2T}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2, \qquad T > 0\,.$$

Projected Gradient Descent on Smooth Functions

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^*) \leqslant \frac{L}{2T}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2, \qquad T > 0\,.$$

Proof.

As before, use sufficient decrease to bound sum of squared gradients in vanilla analysis:

$$\frac{1}{2L}\|\nabla f(\boldsymbol{x}_t)\|^2 \leqslant f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{t+1}) + \frac{L}{2}\|\boldsymbol{y}_{t+1} - \boldsymbol{x}_{t+1}\|^2\,.$$

But now: extra term $\frac{L}{2}\|\boldsymbol{y}_{t+1} - \boldsymbol{x}_{t+1}\|^2$.

Projected Gradient Descent on Smooth Functions

- Replace $\boldsymbol{x}_{t+1}$ in the vanilla analysis with $\boldsymbol{y}_{t+1}$ (the unprojected gradient step):

$$\boldsymbol{g}_t^\top(\boldsymbol{x}_t - \boldsymbol{x}^*) = \frac{1}{2\eta}\left(\eta^2\|\boldsymbol{g}_t\|^2 + \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - \|\boldsymbol{y}_{t+1} - \boldsymbol{x}^*\|^2\right).$$

- Use Fact (ii): $\|\boldsymbol{x} - \Pi_X(\boldsymbol{y})\|^2 + \|\boldsymbol{y} - \Pi_X(\boldsymbol{y})\|^2 \leqslant \|\boldsymbol{x} - \boldsymbol{y}\|^2$.
- With $\boldsymbol{x} = \boldsymbol{x}^*, \boldsymbol{y} = \boldsymbol{y}_{t+1}$, we have $\Pi_X(\boldsymbol{y}) = \boldsymbol{x}_{t+1}$, and hence

$$\|\boldsymbol{x}^* - \boldsymbol{x}_{t+1}\|^2 + \underline{\|\boldsymbol{y}_{t+1} - \boldsymbol{x}_{t+1}\|^2} \leqslant \|\boldsymbol{x}^* - \boldsymbol{y}_{t+1}\|^2$$

- We get back to the vanilla analysis...but with a saving!

$$\boldsymbol{g}_t^\top(\boldsymbol{x}_t - \boldsymbol{x}^*) \leqslant \frac{1}{2\eta}\left(\eta^2\|\boldsymbol{g}_t^2\| + \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 - \underline{\|\boldsymbol{y}_{t+1} - \boldsymbol{x}_{t+1}\|^2}\right)$$

Projected Gradient Descent on Smooth Functions

- Using $f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leqslant \boldsymbol{g}_t^\top(\boldsymbol{x}_t - \boldsymbol{x}^*)$ (convexity), vanilla analysis with saving, $\eta = 1/L$:

$$\sum_{t=0}^{T-1}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)) \leqslant \sum_{t=0}^{T-1} \boldsymbol{g}_t^\top(\boldsymbol{x}_t - \boldsymbol{x}^*)$$

$$\leqslant \frac{1}{2L}\sum_{t=0}^{T-1}\|\boldsymbol{g}_t\|^2 + \frac{L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 - \underline{\frac{L}{2}\sum_{t=0}^{T-1}\|\boldsymbol{y}_{t+1} - \boldsymbol{x}_{t+1}\|^2}.$$

- Using sufficient decrease to bound $\frac{1}{2L}\sum_{t=0}^{T-1}\|\boldsymbol{g}_t\|^2$ by

$$\sum_{t=0}^{T-1}\left(f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{t+1}) + \frac{L}{2}\|\boldsymbol{y}_{t+1} - \boldsymbol{x}_{t+1}\|^2\right)$$

$$= f(\boldsymbol{x}_0) - f(\boldsymbol{x}_T) + \frac{L}{2}\sum_{t=0}^{T-1}\|\boldsymbol{y}_{t+1} - \boldsymbol{x}_{t+1}\|^2$$

Projected Gradient Descent on Smooth Functions

- Putting it together: extra terms cancel, and as in unconstrained case, we get

$$\sum_{t=1}^{T}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)) \leqslant \frac{L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 .$$

- By the definitions of $\boldsymbol{x}_{t+1}$ and $\boldsymbol{y}_{t+1}$,

$$\|\boldsymbol{y}_{t+1} - \boldsymbol{x}_{t+1}\| \leqslant \|\boldsymbol{y}_{t+1} - \boldsymbol{x}_t\| = \eta\|\nabla f(\boldsymbol{x}_t)\| .$$

  Combining this with "Succifiect Decrease "

$$f(\boldsymbol{x}_{t+1}) \leqslant f(\boldsymbol{x}_t) .$$

  Again, we make progress at every step!

- Hence,

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^*) \leqslant \frac{1}{T}\sum_{t=1}^{T}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)) \leqslant \frac{L}{2T}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 .$$

Strongly Convex Functions over X

Recall:

f is strongly convex (with parameter μ) over X if

$$f(\boldsymbol{y}) \geqslant f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in X.$$

Lemma 2

A strongly convex function has a unique minimizer $\boldsymbol{x}^*$ of f over X.

We prove that projected gradient descent converges to $\boldsymbol{x}^*$.

Projected GD on Strongly Convex and Smooth Functions

Theorem 2

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. Let $X \subset \mathbb{R}^d$ be a nonempty closed and convex set and suppose that f is smooth over X with parameter L and strongly convex over X with parameter $\mu > 0$. Choosing $\eta = 1/L$, projected gradient descent with arbitrary $\boldsymbol{x}_0$ satisfies the following two properties.

- Squared distances to $\boldsymbol{x}^*$ are geometrically decreasing:

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \leqslant \left(1 - \frac{\mu}{L}\right)\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2, \qquad t \geqslant 0\,.$$

- The absolute error after T iterations is exponentially small in T:

$$\begin{aligned} f(\boldsymbol{x}_T) - f(\boldsymbol{x}^*) &\leqslant \|\nabla f(\boldsymbol{x}^*)\|\left(1 - \frac{\mu}{L}\right)^{T/2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| \leftarrow \text{ usually, } \nabla f(\boldsymbol{x}^*) \neq \mathbf{0}! \\ &+ \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2, \quad T > 0. \leftarrow \text{ as unconstrained case} \end{aligned}$$

Projected GD on Strongly Convex and Smooth Functions

Proof.

(i) Geometric decrease plus noise: $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \leqslant \cdots$

- unconstrained case:

$$2\eta(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)) + \eta^2\|\nabla f(\boldsymbol{x}_t)\|^2 \qquad\qquad + \underline{(1-\mu\eta)\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2}\,.$$

- constrained case (vanilla analysis with a saving):

$$2\eta(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)) + \eta^2\|\nabla f(\boldsymbol{x}_t)\|^2 - \|\boldsymbol{y}_{t+1} - \boldsymbol{x}_{t+1}\|^2 + \underline{(1-\mu\eta)\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2}\,.$$

Projected GD on Strongly Convex and Smooth Functions

To bound the noise, we use sufficient decrease.

- unconstrained case:

$$f(\boldsymbol{x}_{t+1}) \leqslant f(\boldsymbol{x}_t) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_t)\|^2, \qquad t \geqslant 0,$$

- constrained case:

$$f(\boldsymbol{x}_{t+1}) \leqslant f(\boldsymbol{x}_t) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_t)\|^2 + \frac{L}{2}\|\boldsymbol{y}_{t+1} - \boldsymbol{x}_{t+1}\|^2, \qquad t \geqslant 0\,.$$

Putting it together, the terms $\|\boldsymbol{y}_{t+1} - \boldsymbol{x}_{t+1}\|^2$ cancel, and we get

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \leqslant (1-\mu\eta)\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 = \big(1 - \frac{\mu}{L}\big)\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2$$

in both cases.

Projected GD on Strongly Convex and Smooth Functions

(ii) Error bound from smoothness:

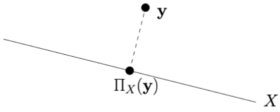
$$
\begin{aligned}
f(\boldsymbol{x}_T) - f(\boldsymbol{x}^*) &\leqslant \nabla f(\boldsymbol{x}^*)^\top (\boldsymbol{x}_T - \boldsymbol{x}^*) + \frac{L}{2}\|\boldsymbol{x}^* - \boldsymbol{x}_T\|^2 \\
&\leqslant \|\nabla f(\boldsymbol{x}^*)\|\|\boldsymbol{x}_T - \boldsymbol{x}^*\| + \frac{L}{2}\|\boldsymbol{x}^* - \boldsymbol{x}_T\|^2 \text{ (Cauchy-Schwarz)} \\
&\leqslant \|\nabla f(\boldsymbol{x}^*)\|\left(1 - \frac{\mu}{L}\right)^{T/2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| + \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^{T}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 \quad (i)
\end{aligned}
$$

constrained error bound $\approx \sqrt{\text{unconstrained error bound}}$
required number of steps roughly doubles.

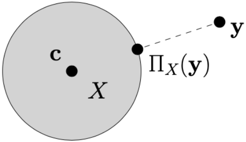
The Projection Step

Computing $\Pi_X(\boldsymbol{y}) := \arg\min_{\boldsymbol{x} \in X} \|\boldsymbol{x} - \boldsymbol{y}\|$ is an optimization problem itself. It can efficiently be solved in relevant cases:

- Projecting onto an affine subspace (leads to system of linear equations, similar to least squares)



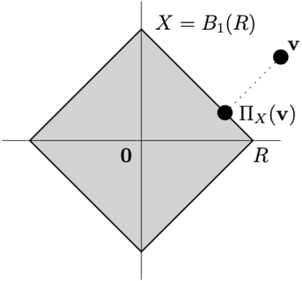
- Projecting onto a Euclidean ball with center $\boldsymbol{c}$ (simply scale the vector $\boldsymbol{y} - \boldsymbol{c}$)



Projecting onto ℓ_1-balls (needed in Lasso)

W.l.o.g. restrict to center at $\mathbf{0}$

$$B_1(R) := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_1 = \sum_{i=1}^{d} |x_i| \leqslant R\}.$$



- $B_1(R)$ is the cross polytope ($2d$ vertices, 2^d facets)

Projecting onto ℓ_1-balls

- This problem can be reduced to a projection onto a simplex set

$$\Pi_X(\boldsymbol{v}) = \arg\min_{\boldsymbol{x} \in \triangle_d} \|\boldsymbol{x} - \boldsymbol{v}\|^2 \,,$$

  where $\triangle_d := \{\boldsymbol{x} \in \mathbb{R}^d : \sum_{i=1}^d x_i = 1, x_i \geqslant 0,\ \forall i\}$ is called the standard simplex.

- Projection onto a simplex can be computed in $\mathcal{O}(d \log d)$ time (can be improved to $\mathcal{O}(d)$) [DSSSC08]

Questions?

Proximal Gradient Descent

Composite Optimization Problems

Consider objective functions composed as

$$f(\boldsymbol{x}) := g(\boldsymbol{x}) + h(\boldsymbol{x})$$

- g is a "nice" function
- h is a "simple" additional term, which however doesn't satisfy the assumptions of niceness which we used in the convergence analysis so far.
- Proximal Gradient Descent is useful for solving nonsmooth, constrained, or structured optimization problems.

Idea

The classical gradient step for minimizing a differentiable function $g : \mathbb{R}^d \to \mathbb{R}$ is typically written as:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \nabla g(\boldsymbol{x}_t)$$

An equivalent way to express this step is by minimizing a local quadratic approximation of $g(\boldsymbol{x})$, given by:

$$\boldsymbol{x}_{t+1} = \arg\min_{\boldsymbol{x}} \left\{ g(\boldsymbol{x}_t) + \nabla g(\boldsymbol{x}_t)(\boldsymbol{x} - \boldsymbol{x}_t) + \frac{1}{2\eta} \|\boldsymbol{x} - \boldsymbol{x}_t\|^2 \right\}$$

Idea

Now for $f = g + h$, keep the same for g, and add h unmodified.

$$\begin{aligned}\boldsymbol{x}_{t+1} &:= \arg\min_{\boldsymbol{y}} \left\{ g(\boldsymbol{x}_t) + \nabla g(\boldsymbol{x}_t)^\top (\boldsymbol{y} - \boldsymbol{x}_t) + \frac{1}{2\eta} \|\boldsymbol{y} - \boldsymbol{x}_t\|^2 + h(\boldsymbol{y}) \right\} \\ &= \arg\min_{\boldsymbol{y}} \left\{ \frac{1}{2\eta} \|\boldsymbol{y} - (\boldsymbol{x}_t - \eta \nabla g(\boldsymbol{x}_t))\|^2 + h(\boldsymbol{y}) \right\},\end{aligned}$$

the proximal gradient descent update.

Proximal Gradient Descent Algorithm

An iteration of proximal gradient descent is defined as

$$\boldsymbol{x}_{t+1} := \mathsf{prox}_{h,\eta}(\boldsymbol{x}_t - \eta \nabla g(\boldsymbol{x}_t)) \, .$$

where the proximal mapping for a given function h, and parameter $\eta > 0$ is defined as

$$\mathsf{prox}_{h,\eta}(\boldsymbol{z}) := \arg\min_{\boldsymbol{y}} \left\{ \frac{1}{2\eta} \|\boldsymbol{y} - \boldsymbol{z}\|^2 + h(\boldsymbol{y}) \right\} .$$

The update step can be equivalently written as

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta G_{h,\eta}(\boldsymbol{x}_t)$$

for $G_{h,\eta}(\boldsymbol{x}) = \frac{1}{\eta}\big(\boldsymbol{x} - \mathsf{prox}_{h,\eta}(\boldsymbol{x} - \eta \nabla g(\boldsymbol{x}))\big)$ being the so called generalized gradient of f.

A Generalization of Gradient Descent?

- $h = 0$: recover gradient descent
- $h = \mathbf{1}_X$: recover projected gradient descent

Given a closed convex set X, the indicator function of the set X is given as the convex function

$$\mathbf{1}_X : \mathbb{R}^d \to \mathbb{R} \cup +\infty$$
$$\boldsymbol{x} \mapsto \mathbf{1}_X(\boldsymbol{x}) := \begin{cases} 0 & \text{if } \boldsymbol{x} \in X \\ +\infty & \text{otherwise}\,. \end{cases}$$

Proximal mapping becomes

$$\mathrm{prox}_{h,\eta}(\boldsymbol{x}) = \arg\min_{\boldsymbol{y}} \left\{ \frac{1}{2\eta} \|\boldsymbol{y} - \boldsymbol{z}\|^2 + \mathbf{1}_X(\boldsymbol{y}) \right\} = \arg\min_{\boldsymbol{y} \in X} \|\boldsymbol{y} - \boldsymbol{z}\|^2$$

Proximal Gradient Descent

Aim to minimize

$$f(\boldsymbol{x}) := g(\boldsymbol{x}) + h(\boldsymbol{x})$$

Proximal Gradient Descent iteration:

$$\boldsymbol{x}_{t+1} := \mathsf{prox}_{h,\eta}(\boldsymbol{x}_t - \eta \nabla g(\boldsymbol{x}_t))\,.$$

where

$$\mathsf{prox}_{h,\eta}(\boldsymbol{z}) := \arg\min_{\boldsymbol{y}} \left\{ \frac{1}{2\eta} \|\boldsymbol{y} - \boldsymbol{z}\|^2 + h(\boldsymbol{y}) \right\}.$$

- Convergence of proximal gradient can be as fast as classic gradient descent
- In every iteration, we have to additionally compute the proximal mapping h.

Convergence of Proximal Gradient Descent

Theorem 3

Let $g : \mathbb{R}^d \to \mathbb{R}$ be convex and smooth with parameter L, h convex, and $\mathrm{prox}_{h,\eta}(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}}\{\|\boldsymbol{x} - \boldsymbol{z}\|^2/(2\eta) + h(\boldsymbol{z})\}$ can be computed. Choosing the fixed stepsize

$$\eta = \frac{1}{L}\,,$$

proximal gradient descent with arbitrary $\boldsymbol{x}_0$ satisfies

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^*) \leqslant \frac{L}{2T}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2, \qquad T > 0\,.$$

- Intuitively, this means that proximal method only "sees" the nice smooth part g of the objective, and is not impacted by the additional h, which it treats separately in each step.

Convergence of Proximal Gradient Descent

Proof.

- Recall that the proximal step could be written as

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \arg\min_{\boldsymbol{y}\in\mathbb{R}^d}\{g(\boldsymbol{x}_t) + \nabla g(\boldsymbol{x}_t)^\top(\boldsymbol{y}-\boldsymbol{x}_t) + \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}_t\|^2 + h(\boldsymbol{y})\} \\ &= \arg\min_{\boldsymbol{y}\in\mathbb{R}^d}\{\psi(\boldsymbol{y})\}\,, \end{aligned}$$

  where the function

$$\psi(\boldsymbol{y}) = g(\boldsymbol{x}_t) + \nabla g(\boldsymbol{x}_t)^\top(\boldsymbol{y}-\boldsymbol{x}_t) + \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}_t\|^2 + h(\boldsymbol{y})$$

  is strongly convex with L.

Convergence of Proximal Gradient Descent

- By the definition of $\boldsymbol{x}_{t+1}$ and strong convexity of ψ

$$\psi(\boldsymbol{y}) \geqslant \psi(\boldsymbol{x}_{t+1}) + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}_{t+1}\|^2 \,,$$

which is equivalent to

$$\begin{aligned}
&\nabla g(\boldsymbol{x}_t)^\top(\boldsymbol{y} - \boldsymbol{x}_t) + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}_t\|^2 + h(\boldsymbol{y}) \\
&\geqslant \nabla g(\boldsymbol{x}_t)^\top(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t) + \frac{L}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2 + h(\boldsymbol{x}_{t+1}) + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}_{t+1}\|^2
\end{aligned}$$

- Rearranging terms and subtracting $h(\boldsymbol{x}_t)$ from both sides,

$$\begin{aligned}
&\nabla g(\boldsymbol{x}_t)^\top(\boldsymbol{y} - \boldsymbol{x}_t) + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}_t\|^2 - \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}_{t+1}\|^2 + h(\boldsymbol{y}) - h(\boldsymbol{x}_t) \\
&\geqslant \nabla g(\boldsymbol{x}_t)^\top(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t) + \frac{L}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2 + h(\boldsymbol{x}_{t+1}) - h(\boldsymbol{x}_t)
\end{aligned}$$

Convergence of Proximal Gradient Descent

- As the function g is L-smooth, we can estimate the right side as

$$\nabla g(\boldsymbol{x}_t)^\top(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t) + \frac{L}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2 \geqslant g(\boldsymbol{x}_{t+1}) - g(\boldsymbol{x}_t)\,.$$

- Since g is convex, on the left side we estimate

$$\nabla g(\boldsymbol{x}_t)^\top(\boldsymbol{y} - \boldsymbol{x}_t) \leqslant g(\boldsymbol{y}) - g(\boldsymbol{x}_t)\,.$$

- Putting this together

$$f(\boldsymbol{y}) - f(\boldsymbol{x}_t) + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}_t\|^2 - \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}_{t+1}\|^2 \geqslant f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t)\,.$$

This holds for any $\boldsymbol{y} \in \mathbb{R}^d$.

- Let $\boldsymbol{y} = \boldsymbol{x}^*$ and sum up the inequality over $t = 0$ to $t = T - 1$

$$\sum_{t=0}^{T-1} (f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)) + \frac{L}{2}\|\boldsymbol{x}^* - \boldsymbol{x}_0\|^2 - \frac{L}{2}\|\boldsymbol{x}^* - \boldsymbol{x}_T\|^2 \geqslant f(\boldsymbol{x}_T) - f(\boldsymbol{x}_0)\,.$$

- Note that $f(\boldsymbol{x}_{t+1}) \leqslant f(\boldsymbol{x}_t)$, as $\psi(\boldsymbol{x}_{t+1}) \leqslant \psi(\boldsymbol{x}_t)$ for each $0 \leqslant t \leqslant T$

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^*) \leqslant \frac{1}{T}\sum_{t=1}^{T} (f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)) \leqslant \frac{L}{2T}\|\boldsymbol{x}^* - \boldsymbol{x}_0\|^2\,.$$

Convergence of Proximal Gradient Descent

$R^2 = \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2.$

$$T \geqslant \frac{R^2 L}{2\varepsilon} \qquad \Rightarrow \qquad \text{error} \leqslant \frac{L}{2T} R^2 \leqslant \varepsilon\,.$$

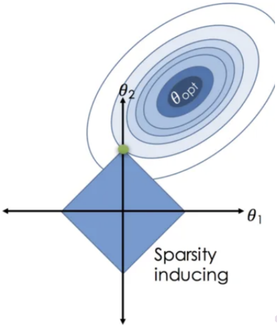
- The convergence rate $\mathcal{O}(\frac{1}{\varepsilon})$ is the same as gradient descent on convex smooth functions (Theorem 2 from lecture 2), but now for any h for which we can compute the proximal mapping.

Example: Iterative Soft-Thresholding Algorithm (ISTA)

Lasso regression:

$$\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{y} - A\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

- $A \in \mathbb{R}^{n \times d}$ is the feature matrix, $\boldsymbol{\beta} \in \mathbb{R}^d$ is the coefficient vector, $\boldsymbol{y} \in \mathbb{R}^n$ is the response variable
- $\lambda > 0$ is a tuning parameter that controls sparsity (the ℓ_1-norm forces coefficients β_j to be exactly zero, performing feature selection)



Example: Iterative Soft-Thresholding Algorithm (ISTA)

$$f(\boldsymbol{\beta}) = \underbrace{\frac{1}{2}\|\boldsymbol{y} - A\boldsymbol{\beta}\|_2^2}_{g(\boldsymbol{\beta})} + \underbrace{\lambda\|\boldsymbol{\beta}\|_1}_{h(\boldsymbol{\beta})}$$

Proximal mapping is now

$$\mathrm{prox}_{h,\eta}(\boldsymbol{z}) = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2\eta}\|\boldsymbol{\beta} - \boldsymbol{z}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\} =: S_{\lambda,\eta}(\boldsymbol{z})\,.$$

Here, $S_{\lambda,\eta}(\boldsymbol{z})$ is the soft-thresholding operator

$$[S_{\lambda,\eta}(\boldsymbol{z})]_i = \begin{cases} z_i - \eta\lambda & \text{if } z_i > \eta\lambda \\ 0 & \text{if } |z_i| < \eta\lambda \\ z_i + \eta\lambda & \text{if } z_i < -\eta\lambda \end{cases}, \quad i = 1, \ldots, d$$

Example: Iterative Soft-Thresholding Algorithm (ISTA)

Recall $\nabla g(\boldsymbol{\beta}) = -A^\top(\boldsymbol{y} - A\boldsymbol{\beta})$, hence proximal gradient update is:

$$\boldsymbol{x}_{t+1} = S_{\lambda,\eta}\big(\boldsymbol{x}_t + \eta A^\top(\boldsymbol{y} - A\boldsymbol{x}_t)\big) .$$

Often called the iterative soft-thresholding algorithm (ISTA)

Questions?

Mirror Descent

Mirror Descent: Motivation

Consider the simplex-constrained optimization problem

$$\min_{\boldsymbol{x} \in \triangle_d} f(\boldsymbol{x}) \,,$$

where the simplex $\triangle_d := \{\boldsymbol{x} \in \mathbb{R}^d : \sum_{i=1}^d x_i = 1, x_i \geqslant 0, \ \forall i\}$

Now, we assume $\|\nabla f(\boldsymbol{x})\|_\infty = \max\limits_{i=1,\ldots,d} |[\nabla f(\boldsymbol{x})]_i| \leqslant 1, \forall \boldsymbol{x} \in \triangle_d$.

- The largest element of any gradient is bounded by 1.
- All the elements of any gradient are bounded by 1.
- The extreme cases here are the following two vectors taken as the gradient.

$$\text{(the minimal vector) } \mathbf{0}_d = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \,, \quad \text{(the maximal vector) } \mathbf{1}_d = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

Mirror Descent: Motivation

- For the vector $\mathbf{1}_d$, it has ℓ_2-norm $\|\mathbf{1}_d\|_2 = \sqrt{d}$
- In other words, $\|\nabla f(\boldsymbol{x})\|_\infty \leqslant 1$ gives $\|\nabla f(\boldsymbol{x})\|_2 \leqslant L = \sqrt{d}$
- Convergence of GD (on convex and L-Lipschitz functions):

$$f(\boldsymbol{x}_{\text{best}}^{(T)}) - f(\boldsymbol{x}^*) \leqslant R\sqrt{\frac{d}{T}}$$

- It turns out the rate $\mathcal{O}\big(\sqrt{\frac{d}{T}}\big)$ is not optimal, mirror descent can do better as $\mathcal{O}\big(\sqrt{\frac{\log d}{T}}\big)$

Mirror Descent: Preliminary

- Fix an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$, and a compact convex set $X \subseteq \mathbb{R}^d$. The dual norm $\|\cdot\|_*$ is defined as

$$\|\boldsymbol{g}\|_* = \sup_{\|\boldsymbol{x}\| \leqslant 1} \boldsymbol{g}^\top \boldsymbol{x}.$$

- We say that a convex function $f : X \to \mathbb{R}^d$ is
  - L-Lipschitz w.r.t. $\|\cdot\|$ if $\forall x \in X, \boldsymbol{g} \in \partial f(\boldsymbol{x}), \|\boldsymbol{g}\|_* \leqslant L$
  - β-smooth w.r.t. $\|\cdot\|$ if $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_* \leqslant \beta \|\boldsymbol{x} - \boldsymbol{y}\|, \forall \boldsymbol{x}, \boldsymbol{y} \in X$
  - μ-strongly convex w.r.t. $\|\cdot\|$ if $f(\boldsymbol{x}) - f(\boldsymbol{y}) \leqslant \boldsymbol{g}^\top (\boldsymbol{x} - \boldsymbol{y}) - \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2, \forall \boldsymbol{x}, \boldsymbol{y} \in X, \boldsymbol{g} \in \partial f(\boldsymbol{x})$

Consider the mirror descent [Nemirovski and Yudin (1983)] iteration

$$\boldsymbol{y}_{t+1} = (\nabla\Phi)^{-1}(\nabla\Phi(\boldsymbol{x}_t) - \eta_t \boldsymbol{g}_t) \quad \text{and} \quad \boldsymbol{x}_{t+1} = \operatorname*{arg\,min}_{\boldsymbol{x}\in X} D_\Phi(\boldsymbol{x}, \boldsymbol{y}_{t+1}),$$

with $\boldsymbol{g}_t \in \partial f(\boldsymbol{x}_t)$.

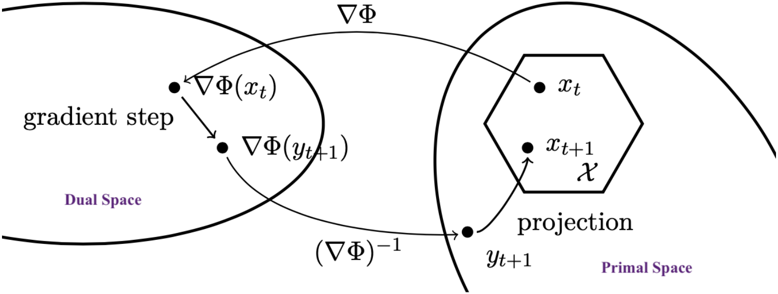
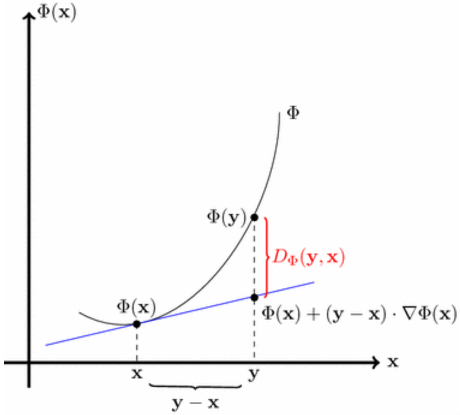


Figure: The "geometry" of mirror descent from [Bubeck 2015].

Mirror Descent: Key elements

- Mirror potential $\Phi : \mathbb{R}^d \to \mathbb{R}$ is strictly convex, continuously differentiable with $\lim_{\|\boldsymbol{x}\|_2 \to \infty} \|\nabla \Phi(x)\| = \infty$.
- Define the Bregman divergence associated to f as

$$D_f(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y}) \, .$$



- The projection via Bregman divergence associated to Φ

$$\Pi_X^{\Phi}(\boldsymbol{y}) = \arg\min_{\boldsymbol{x} \in X} D_{\Phi}(\boldsymbol{x}, \boldsymbol{y}), \quad \forall \boldsymbol{y} \in X\,.$$

- Properties of Φ ensures the existence and uniqueness of this projection Π_X^{Φ}.

Convergence of Mirror Descent

Let $\boldsymbol{x}_1 \in \arg\min_{\boldsymbol{x}\in X} \Phi(\boldsymbol{x})$. For $t \geqslant 1$, let $\boldsymbol{y}_{t+1} \in \mathbb{R}^d$ such that

$$\nabla\Phi(\boldsymbol{y}_{t+1}) = \nabla\Phi(\boldsymbol{x}_t) - \eta \boldsymbol{g}_t, \text{ where } \boldsymbol{g}_t \in \partial f(\boldsymbol{x}_t)\,,$$

and

$$\boldsymbol{x}_{t+1} \in \Pi_X^{\Phi}(\boldsymbol{y}_{t+1})\,.$$

Theorem 4

Let

- Φ be a mirror map ρ-strongly convex on X w.r.t $\|\cdot\|$.
- $R^2 = \sup_{\boldsymbol{x}\in X} \Phi(\boldsymbol{x}) - \Phi(\boldsymbol{x}_1)$.
- f be convex and L-Lipschitz w.r.t $\|\cdot\|$.

Mirror descent with $\eta = \frac{2R}{L}\sqrt{\frac{\rho}{T}}$ satisfies

$$f\Big(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\Big) - f(\boldsymbol{x}^*) \leqslant RL\sqrt{\frac{1}{\rho T}}\,.$$

To prove Theorem 4, we need some auxiliary arguments.

- Given the Bregman divergence associated to Φ, it holds that

$$(\nabla\Phi(\boldsymbol{x}) - \nabla\Phi(\boldsymbol{y}))^\top(\boldsymbol{x} - \boldsymbol{z}) = D_\Phi(\boldsymbol{x}, \boldsymbol{y}) + D_\Phi(\boldsymbol{z}, \boldsymbol{x}) - D_\Phi(\boldsymbol{z}, \boldsymbol{y})\,.$$

- Moreover,

$$D_\Phi(\boldsymbol{x}, \Pi_X^\Phi(\boldsymbol{y})) + D_\Phi(\Pi_X^\Phi(\boldsymbol{y}), \boldsymbol{y}) \leqslant D_\Phi(\boldsymbol{x}, \boldsymbol{y})\,.$$

Proof.

$$
\begin{aligned}
& f(\boldsymbol{x}_t) - f(\boldsymbol{x}) \\
& \leqslant \boldsymbol{g}_t^\top (\boldsymbol{x}_t - \boldsymbol{x}) \\
& = \frac{1}{\eta}\big(\nabla\Phi(\boldsymbol{x}_t) - \nabla\Phi(\boldsymbol{y}_{t+1})\big)^\top (\boldsymbol{x}_t - \boldsymbol{x}) \\
& = \frac{1}{\eta}\Big(D_\Phi(\boldsymbol{x}, \boldsymbol{x}_t) + D_\Phi(\boldsymbol{x}_t, \boldsymbol{y}_{t+1}) - D_\Phi(\boldsymbol{x}, \boldsymbol{y}_{t+1})\Big) \\
& \leqslant \frac{1}{\eta}\Big(D_\Phi(\boldsymbol{x}, \boldsymbol{x}_t) + D_\Phi(\boldsymbol{x}_t, \boldsymbol{y}_{t+1}) - D_\Phi(\boldsymbol{x}, \boldsymbol{x}_{t+1}) - D_\Phi(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1})\Big) .
\end{aligned}
$$

The term $D_\Phi(\boldsymbol{x}, \boldsymbol{x}_t) - D_\Phi(\boldsymbol{x}, \boldsymbol{x}_{t+1})$ will lead to a telescopic sum when summing over $t = 1$ to $t = T$. It remains to bound the other term...

Convergence of Mirror Descent

Recall that Φ is ρ-strongly convex, and $az - bz^2 \leqslant \frac{a^2}{4b}, \forall z \in \mathbb{R}$.

$$
\begin{aligned}
& D_\Phi(\boldsymbol{x}_t, \boldsymbol{y}_{t+1}) - D_\Phi(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) \\
& = \Phi(\boldsymbol{x}_t) - \Phi(\boldsymbol{x}_{t+1}) - \nabla \Phi(\boldsymbol{y}_{t+1})^\top (\boldsymbol{x}_t - \boldsymbol{x}_{t+1}) \\
& \leqslant \big(\nabla \Phi(\boldsymbol{x}_t) - \nabla \Phi(\boldsymbol{y}_{t+1})\big)^\top (\boldsymbol{x}_t - \boldsymbol{x}_{t+1}) - \frac{\rho}{2} \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2 \\
& = \eta \boldsymbol{g}_t^\top (\boldsymbol{x}_t - \boldsymbol{x}_{t+1}) - \frac{\rho}{2} \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2 \\
& \leqslant \eta L \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\| - \frac{\rho}{2} \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2 \\
& \leqslant \frac{(\eta L)^2}{2\rho} \, .
\end{aligned}
$$

Thus,

$$
\sum_{t=1}^{T} \Big(f(\boldsymbol{x}_t) - f(\boldsymbol{x}) \Big) \leqslant \frac{D_\Phi(\boldsymbol{x}, \boldsymbol{x}_1)}{\eta} + \eta \frac{L^2 T}{2\rho} \, .
$$

Standard Setups for Mirror Descent

- **"Ball setup".** The mirror potential

$$\Phi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2, \quad \forall \boldsymbol{x} \in \mathbb{R}^d.$$

  - Associated Bregman divergence $D_\Phi(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$.
  - This is exactly equivalent to the projected subgradient descent.

Standard Setups for Mirror Descent

- **"Simplex setup".** The mirror potential

$$\Phi(\boldsymbol{x}) = \sum_{i=1}^{d} x_i \log(x_i), \quad \boldsymbol{x} \in \mathbb{R}^d_{++} = \{\boldsymbol{x} \in \mathbb{R}^d : x_i > 0, i = 1, \ldots, d\}.$$

  - The gradient update $\nabla\Phi(\boldsymbol{y}_{t+1}) = \nabla\Phi(\boldsymbol{x}_t) - \eta\nabla f(\boldsymbol{x}_t)$ can be written as
$$[\boldsymbol{y}_{t+1}]_i = [\boldsymbol{x}_t]_i \exp(-\eta[\nabla f(\boldsymbol{x}_t)]_i), \quad i = 1, \ldots, d.$$
  - The Bregman divergence of Φ is

$$D_\Phi(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{d} x_i \log(x_i / y_i) \quad \text{(Kullback-Leibler divergence)}$$

  - Projection of $\boldsymbol{y}$ onto the simplex $\triangle_d$ under the KL divergence leads to renormalization $\boldsymbol{y} \to \boldsymbol{y}/\|\boldsymbol{y}\|_1$ (see notes).
  - For $X = \triangle_d, \boldsymbol{x}_1 = (1/d, \ldots, 1/d)$ and $R^2 = \log d$ (see notes).

References

Sébastien Bubeck, *Convex optimization: Algorithms and complexity*, Foundations and Trends in Machine Learning **8** (2015), no. 3-4, 231–357.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra, *Efficient projections onto the l 1-ball for learning in high dimensions*, Proceedings of the 25th international conference on Machine learning, 2008, pp. 272–279.

Mor Harchol-Balter, *Introduction to probability for computing*, Cambridge University Press, 2023.

Laurent Lessard, Benjamin Recht, and Andrew Packard, *Analysis and design of optimization algorithms via integral quadratic constraints*, SIAM Journal on Optimization **26** (2016), no. 1, 57–95.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin, *Problem complexity and method efficiency in optimization*.

Stephen J Wright and Benjamin Recht, *Optimization for data analysis*, Cambridge University Press, 2022.