Optimization for Machine Learning

Lecture 11: Proximal Gradient Methods

Sebastian Stich

CISPA – `https://cms.cispa.saarland/optml24/`

July 2, 2024

Lecture Outline

Composite Optimization Problems

Projected Gradient Descent

Proximal Gradient Descent

Stochastic Proximal Gradient Descent

Composite Optimization Problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + \psi(\mathbf{x})$$

- f: $\mathbb{R}^d \to \mathbb{R}$, L-smooth
- ψ: $\mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ proper, closed and convex regularizer

Example: Constrained Minimization

Let $X \subseteq \mathbf{dom}(f)$ be a convex set.

$$\min_{\mathbf{x}\in X} f(\mathbf{x}) = \min_{\mathbf{x}\in\mathbb{R}^d} \underbrace{f(\mathbf{x}) + \psi(\mathbf{x})}_{} =$$

where $\psi(\mathbf{x}) := \mathbf{1}_X(\mathbf{x})$

$$\min_{x} \begin{cases} \text{if } x \in X = f(x) \\ \text{if } x \notin X = \infty \end{cases} = \min_{x\in X} f(x)$$

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

$$\mathbf{1}_X : \mathbb{R}^d \to \mathbb{R} \cup +\infty$$

$$\mathbf{x} \mapsto \mathbf{1}_X(\mathbf{x}) := \begin{cases} 0 & \text{if } \mathbf{x} \in X, \\ +\infty & \text{otherwise.} \end{cases}$$

Example: Regularization

Lasso: Sparsity inducing regularization

$f(x) = \|Ax-b\|^2$

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + \lambda \|\mathbf{x}\|_1$$

with $\|\mathbf{x}\|_1 := \sum_{i=1}^d |\mathbf{x}_i|$.

Ridge regression:

$f(x) = \|Ax-b\|^2$

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + \frac{\lambda}{2} \|\mathbf{x}\|_2^2$$

with $\|\mathbf{x}\|_2^2 := \sum_{i=1}^d |\mathbf{x}_i|^2$.

Example: Consensus Formulation

Distributed optimization:

$$\min_{\mathbf{x}\in\mathbb{R}^d}\left[f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})\right] = \min_{\mathbf{x}_1,\dots,\mathbf{x}_n\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}_i) + \psi(\mathbf{x}_1,\dots,\mathbf{x}_n)\,,$$

where $\psi(\mathbf{x}_1,\dots,\mathbf{x}_n) := \begin{cases} 0, & \text{if } \mathbf{x}_1 = \cdots = \mathbf{x}_n \\ +\infty, & \text{otherwise} \end{cases}.$

Lecture Outline

Composite Optimization Problems

Projected Gradient Descent

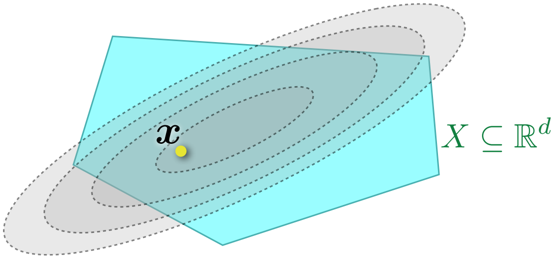
Proximal Gradient Descent

Stochastic Proximal Gradient Descent

$$\min_{x \in \mathbb{R}^d} f(x) + \varphi(x)$$

Constrained Optimization

$$\begin{array}{rl} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in X \end{array}$$



Constrained Minimization

Definition 11.1

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be convex and let $X \subseteq \mathbf{dom}(f)$ be a convex set. A point $\mathbf{x} \in X$ is a minimizer of f over X if

$$f(\mathbf{x}) \leq f(\mathbf{y}) \quad \forall \mathbf{y} \in X.$$

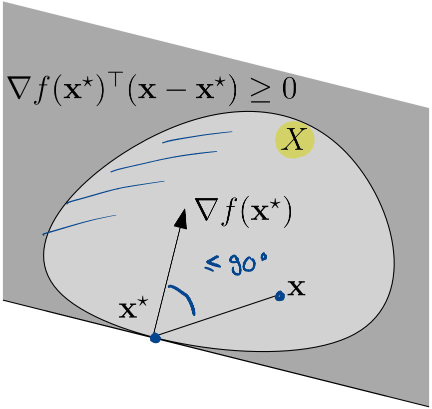
Lemma 11.2

Suppose that $f : \mathbf{dom}(f) \to \mathbb{R}$ is convex and differentiable over an open domain $\mathbf{dom}(f) \subseteq \mathbb{R}^d$, and let $X \subseteq \mathbf{dom}(f)$ be a convex set. Point $\mathbf{x}^\star \in X$ is a minimizer of f over X if and only if

$$\nabla f(\mathbf{x}^\star)^\top (\mathbf{x} - \mathbf{x}^\star) \geq 0 \quad \forall \mathbf{x} \in X.$$

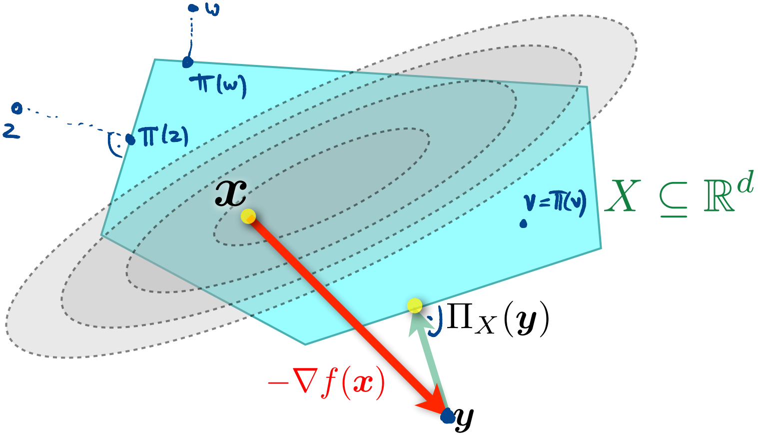
case $X = \mathbb{R}^d \Rightarrow \quad \nabla f(x^\star)^\top (y) \geq 0 \quad y \in \mathbb{R}^d \Rightarrow \nabla f(x^\star) = 0$!

Constrained Minimization



Projected Gradient Descent

Idea: project onto X after every step: $\Pi_X(\mathbf{y}) := \operatorname{argmin}_{\mathbf{x}\in X} \|\mathbf{x}-\mathbf{y}\|$



Projected gradient descent: $\mathbf{x}_{t+1} := \Pi_X\left[\mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t)\right]$

The Algorithm

Projected gradient descent:

$$
\begin{aligned}
\mathbf{y}_{t+1} &:= \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t), \\
\mathbf{x}_{t+1} &:= \Pi_X(\mathbf{y}_{t+1}) := \operatorname*{argmin}_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{y}_{t+1}\|^2.
\end{aligned}
$$

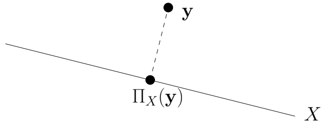
for **timesteps** $t = 0, 1, \ldots,$ and **stepsize** $\gamma \geq 0$.

The Projection Step: $\Pi_X(\mathbf{y}) := \operatorname{argmin}_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{y}\|$

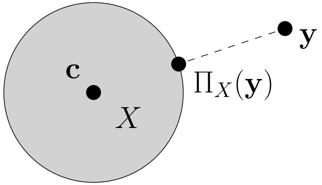
Computing $\Pi_X(\mathbf{y})$ is an optimization problem itself.

It can efficiently be solved in relevant cases:

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

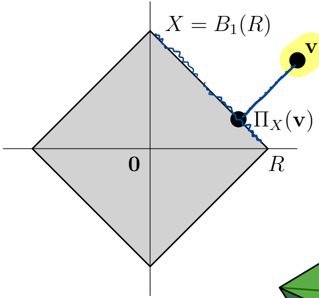


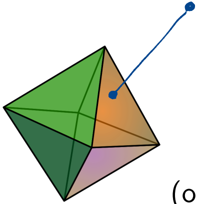
- Projecting onto a Euclidean ball with center $\mathbf{c}$ (simply scale the vector $\mathbf{y} - \mathbf{c}$)



Projecting onto ℓ_1-balls (needed in Lasso)

W.l.o.g. restrict to center at $\mathbf{0}$: $B_1(R) = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_1 = \sum_{i=1}^d |x_i| \le R\}$.





$B_1(R)$ is the cross polytope ($2d$ vertices, 2^d facets). (octahedron, $d = 3$)

Section 4.5: projection can be computed in $\mathcal{O}(d \log d)$ time

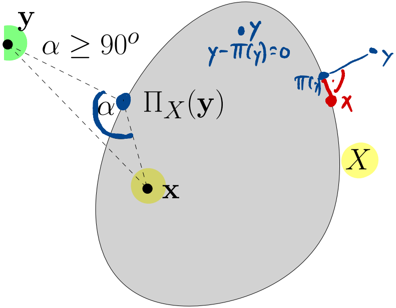
Properties of Projection

Fact 11.3

Let $X \subseteq \mathbb{R}^d$ be closed and convex, $\mathbf{x} \in X, \mathbf{y} \in \mathbb{R}^d$. Then

(i) $(\mathbf{x} - \Pi_X(\mathbf{y}))^\top(\mathbf{y} - \Pi_X(\mathbf{y})) \leq 0$.

(ii) $\|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2$.



Properties of Projection II

Fact 11.4

Let $X \subseteq \mathbb{R}^d$ be closed and convex, $\mathbf{x} \in X, \mathbf{y} \in \mathbb{R}^d$. Then

(i) $(\mathbf{x} - \Pi_X(\mathbf{y}))^\top(\mathbf{y} - \Pi_X(\mathbf{y})) \leq 0$.

(ii) $\|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2$.

Proof.

(i) $\Pi_X(\mathbf{y})$ is minimizer of (differentiable) convex function $d_\mathbf{y}(\mathbf{x}) = \|\mathbf{x} - \mathbf{y}\|^2$ over X.
By first-order characterization of optimality (**Lemma 2.28**),

$$
\begin{aligned}
0 &\leq \nabla d_\mathbf{y}(\Pi_X(\mathbf{y}))^\top(\mathbf{x} - \Pi_X(\mathbf{y})) \\
&= 2(\Pi_X(\mathbf{y}) - \mathbf{y})^\top(\mathbf{x} - \Pi_X(\mathbf{y})) \quad \checkmark \\
\Leftrightarrow \quad 0 &\geq 2(\mathbf{y} - \Pi_X(\mathbf{y}))^\top(\mathbf{x} - \Pi_X(\mathbf{y})) \\
\Leftrightarrow \quad 0 &\geq (\mathbf{x} - \Pi_X(\mathbf{y}))^\top(\mathbf{y} - \Pi_X(\mathbf{y})) \quad \checkmark
\end{aligned}
$$

□

Properties of Projection III

Fact 11.5

Let $X \subseteq \mathbb{R}^d$ *be closed and convex,* $\mathbf{x} \in X, \mathbf{y} \in \mathbb{R}^d$. *Then*

(i) $(\mathbf{x} - \Pi_X(\mathbf{y}))^\top (\mathbf{y} - \Pi_X(\mathbf{y})) \leq 0$.

(ii) $\|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2$.

Proof.

(ii)

$$\mathbf{v} := (\mathbf{x} - \Pi_X(\mathbf{y})), \quad \mathbf{w} := (\mathbf{y} - \Pi_X(\mathbf{y})).$$

$\|v - w\|^2 = \|v\|^2 + \|w\|^2 - 2 v^T w$

By (i),

$$\begin{aligned} 0 \geq 2\mathbf{v}^\top \mathbf{w} &= \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2 \\ &= \|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 - \|\mathbf{x} - \mathbf{y}\|^2. \quad \checkmark \end{aligned}$$

$\square$

Results for projected gradient descent over closed and convex X

The same number of steps as 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 (one) of the previous proofs for gradient descent.

BUT:

- Each step involves a projection onto X
- may or may not be efficient (in relevant cases, it is)...

Smooth convex functions over X: $\mathcal{O}(1/\varepsilon)$ steps

Theorem 11.6

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 $\mathbf{x}^\star$ of f over X; furthermore, suppose that f is smooth over X with parameter L. Choosing stepsize

$$\gamma := \frac{1}{L},$$

projected gradient descent yields

$$\frac{1}{T}\sum_{t=1}^{T} f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

(**Exercise 29** in the lecture notes ask you to prove $f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$).

Step I: Sufficient decrease for projected gradient descent

Lemma 11.7

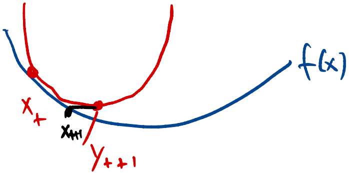
Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and smooth with parameter L over X. Choosing stepsize

$$\gamma := \frac{1}{L},$$

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

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2, \quad t \geq 0.$$

$\Pi(x_{t+1})$



Sufficient decrease II

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

Proof.

Use smoothness, $\mathbf{y}_{t+1} - \mathbf{x}_t = -\nabla f(\mathbf{x}_t)/L$, $2\mathbf{v}^\top \mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2$:

$$\begin{aligned}
f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\
&= f(\mathbf{x}_t) - L(\mathbf{y}_{t+1} - \mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\
&= f(\mathbf{x}_t) - \frac{L}{2}\left(\|\mathbf{y}_{t+1} - \mathbf{x}_t\|^2 + \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 - \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2\right) + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\
&= f(\mathbf{x}_t) - \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_t\|^2 + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2 \\
&= f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2.
\end{aligned}$$

Proof I

- By convexity:

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \le \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^\star)$$

- With $\mathbf{y}_{t+1} = \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t)$ we have

$\|v-w\|^2 = \|v\|^2 + \|w\|^2 - 2v^\top w$

$$\nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^\star) = \frac{1}{2\gamma}\left(\gamma^2 \|\nabla f(\mathbf{x}_t)\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \boxed{\|\mathbf{y}_{t+1} - \mathbf{x}^\star\|^2)}\right.$$

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

$$\|\mathbf{x}^\star - \mathbf{x}_{t+1}\|^2 + \underline{\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2} \le \|\mathbf{x}^\star - \mathbf{y}_{t+1}\|^2$$

- This saving term is crucial to make telescoping work again!

$$\nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^\star) \le \frac{1}{2\gamma}\left(\gamma^2 \|\nabla f(\mathbf{x}_t)\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 - \underline{\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2}\right)$$

- Set $\gamma = \frac{1}{L}$ and use the sufficient decrease lemma to bound $\|\nabla f(\mathbf{x}_t)\|^2$:

$$\begin{aligned}\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}^\star) &\leq \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 - \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2 \\ &\leq f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\end{aligned}$$

- This "trick" makes telescoping work again!

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

Hence

$$\frac{1}{T}\sum_{t=1}^{T} f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$$

Lecture Outline

Composite Optimization Problems

Projected Gradient Descent

Proximal Gradient Descent

Stochastic Proximal Gradient Descent

Composite optimization problems

Consider objective functions composed as

$$F(\mathbf{x}) := f(\mathbf{x}) + \psi(\mathbf{x})$$

where f is a "nice" function, where as ψ is a "simple" additional term, which however doesn't satisfy the assumptions of niceness which we used in the convergence analysis so far.

In particular, an important case is when ψ is not differentiable.

Idea

The classical gradient step for minimizing f:

$$\mathbf{x}_{t+1} = \underset{\mathbf{y}}{\operatorname{argmin}}\ f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{y} - \mathbf{x}_t) + \frac{1}{2\gamma}\|\mathbf{y} - \mathbf{x}_t\|^2 \ .$$

For the stepsize $\gamma := \frac{1}{L}$ it exactly minimizes the local quadratic model of g at our current iterate $\mathbf{x}_t$, formed by the smoothness property with parameter L.

Now for $F = f + \psi$, keep the same for f, and add ψ unmodified.

$$\begin{aligned}
\mathbf{x}_{t+1} &:= \underset{\mathbf{y}}{\operatorname{argmin}}\ f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{y} - \mathbf{x}_t) + \frac{1}{2\gamma}\|\mathbf{y} - \mathbf{x}_t\|^2 + \psi(\mathbf{y}) \\
&= \underset{\mathbf{y}}{\operatorname{argmin}}\ \frac{1}{2\gamma}\|\mathbf{y} - \underbrace{(\mathbf{x}_t - \gamma\nabla f(\mathbf{x}_t))}_{= y_{t+1}}\|^2 + \psi(\mathbf{y}) \ ,
\end{aligned}$$

the proximal gradient descent update.

The proximal gradient descent algorithm

An iteration of proximal gradient descent is defined as

$$\mathbf{x}_{t+1} := \mathrm{prox}_{\psi,\gamma}(\mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t)) \ .$$

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

$$\mathrm{prox}_{\psi,\gamma}(\mathbf{z}) := \operatorname*{argmin}_{\mathbf{y}} \left\{ \frac{1}{2\gamma} \|\mathbf{y} - \mathbf{z}\|^2 + \psi(\mathbf{y}) \right\} \ .$$

"simple" $\hat{=}$ this proximal problem can be solved efficiently!

A generalization of gradient descent?

- $\psi \equiv 0$: recover gradient descent
- $\psi \equiv \mathbf{1}_X$: recover projected gradient descent!
  Proximal mapping becomes

$$\mathrm{prox}_{h,\gamma}(\mathbf{z}) := \underset{\mathbf{y}}{\mathrm{argmin}} \left\{ \frac{1}{2\gamma} \|\mathbf{y} - \mathbf{z}\|^2 + \mathbf{1}_X(\mathbf{y}) \right\} = \underset{\mathbf{y} \in X}{\mathrm{argmin}} \ \|\mathbf{y} - \mathbf{z}\|^2$$

  which is the projection onto X.

Convergence in $\mathcal{O}(1/\varepsilon)$ steps

For many classes of function f, it can be shown that proximal gradient descent on $f(\mathbf{x}) + \psi(\mathbf{x})$ converges in the same number of steps, as gradient descent on $f(\mathbf{x})$.

The the additional complexity is "hidden" in the proximal step, as it is assumed that the proximal update can be computed efficiently.

Lecture Outline

Composite Optimization Problems

Projected Gradient Descent

Proximal Gradient Descent

Stochastic Proximal Gradient Descent

Stochastic Proximal Gradient Method

$$\mathbf{x}_{t+1} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^d} \mathbf{g}_t{}^\top \mathbf{x} + \psi(\mathbf{x}) + \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{x}_t\|^2 \;,$$

where $\mathbb{E}\mathbf{g}_t = \nabla f(\mathbf{x}_t)$ with bounded variance:

$$\mathbb{E}\,\|\mathbf{g}_t - \nabla f(\mathbf{x}_t)\|^2 \leq \sigma^2 .$$

Be careful with stochastic prox!

- Again, we would expect that the Stochastic Proximal Gradient Method works similarly as the Stochastic Gradient Method.
- However, the proximal step with a stochastic gradients could amplify the stochastic variance.

$$\mathbf{x}_{t+1} = \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^d} \mathbf{g}_t^\top \mathbf{x} + \psi(\mathbf{x}) + \frac{1}{2\gamma} \|\mathbf{x} - \mathbf{x}_t\|^2$$

- In practice, this is often addressed with large batches. In theory, the batch size sometimes needs to be taken as large as $\frac{1}{\epsilon}$!

SPG with momentum

Large batches can be avoided with momentum.

SPG with momentum:
For an initialization $\mathbf{m}_{-1} \in \mathbb{R}^d$, and a momentum parameter η:

$$\mathbf{m}_t = (1-\eta)\mathbf{m}_{t-1} + \eta \mathbf{g}_t$$

$$\mathbf{x}_{t+1} = \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^d} \mathbf{m}_t^\top \mathbf{x} + \psi(\mathbf{x}) + \frac{1}{2\gamma} \|\mathbf{x} - \mathbf{x}_t\|^2 ,$$

where again $\mathbb{E}\mathbf{g}_t = \nabla f(\mathbf{x}_t)$ denotes a stochastic gradient.

SPG with momentum [GRS24]

Theorem 11.8

If $\mathbf{m}_0$ *is initialized such that* $\mathbb{E}\,\|\mathbf{m}_0 - \nabla f(\mathbf{x}_0)\|^2 = \mathcal{O}(LF_0)$ *with* $F_0 = f(\mathbf{x}_0) - f^\star$ *,* $\mathbb{E}\,\|\mathbf{g}_t - \nabla f(\mathbf{x}_t)\|^2 \le \sigma^2$*,* f *is* L*-smooth, and the momentum parameter* $\eta = \frac{3L\gamma}{1-L\gamma}$*, and* $\gamma = \min\left\{\frac{1}{4L}, \frac{C}{\sqrt{T}}\right\}$ *(for a constant* C*), then*

$$\frac{1}{T}\sum_{t=0}^{T} \mathbb{E}\,\|\nabla f(\mathbf{x}_t)\|^2 \le \mathcal{O}\left(\frac{LF_0}{T} + \frac{\sigma\sqrt{LF_0}}{\sqrt{T}}\right).$$

($\hat{=}$ SGD on unconstrained problems)

The initialization condition can for instance be reached for $\mathbf{m}_0 = \frac{1}{|B_0|}\sum_{i\in B_0} \mathbf{g}(\mathbf{x}_0)$ with a mini-batch of size $\max\left\{\frac{\sigma^2}{LF_0}, 1\right\}$. This batch size does not depend on ϵ.

✓ $|B_0|$ indep. oracle calls

Recommended reading: [GRS24]

Discussion

- composite problems $f(\mathbf{x}) + \psi(\mathbf{x})$
- under the assumption that $\psi(\mathbf{x})$ is simple, composite problems can usually be solved with proximal methods in the same number of iterations as it takes to minimize $f(\mathbf{x})$ alone

Bibliography I

Yuan Gao, Anton Rodomanov, and Sebastian U Stich.
Non-convex stochastic composite optimization with polyak momentum.
arXiv preprint arXiv:2403.02967, 2024.