Stochastic optimization

Darina Dvinskikh

HSE University

November 8, 2023

Course organization

- Lectures
- Exercises
- Homework

Grading: the average grade obtained by Exercise Sessions and Homework.

Problem Statement

Problem

$$\min_{w \in W \subseteq \mathbb{R}^d} F(w) \triangleq \mathbb{E}_Z f(w, Z),$$

where function f is convex in w, W is a convex set, and $\mathbb{E}_Z f(w, Z)$ is the expectation of f with respect to Z.

Goal: find a 'cheap' algorithm to approximate minimizer

$$w^* \triangleq \arg\min_{w \in W} \mathbb{E}_Z f(w, Z).$$

Complexity of an algorithm is determined by

- number of iterations N
- number of oracle calls, e.g., number of gradient calculations (1-st order oracle)

Preliminaries

Definition (ε-solution)

$\hat{w}$ is an ε-solution of stochastic problem

$$\min_{w \in W} F(w) \triangleq \mathbb{E}_Z f(w, Z)$$

if the non-optimality gap in terms of the objective function is upper bounded by ε:

$$\mathbb{E}F(\hat{w}) - \min_{w \in W} F(w) \leq \varepsilon$$

Two approaches

There are two competing approaches to solve the stochastic problem

$$\min_{w \in W} F(w) \triangleq \mathbb{E}_Z f(w, Z).$$

- Stochastic Approximation (SA):
  stochastic problem is solved e.g., by stochastic gradient descent (SGD)
- Sample Average Approximation (SAA), or Monte Carlo approach:
  stochastic problem is replaced by its empirical counterpart

$$\min_{w \in W} \hat{F}(w) \triangleq \frac{1}{m} \sum_{i=1}^{m} f(w, z_i),$$

where $z_1, z_2, ..., z_m$ are the realizations of a random variable Z.

Outline

Projected GD

(let's start with deterministic method)

Projected GD repeats for $k = 1, \dots, N$:

$$w^{k+1} = \arg\min_{w \in W} \left\| w^k - \alpha_k \nabla F(w^k) - w \right\|_2^2$$

$$= \arg\min_{w \in W} \left\{ F(w^k) + \langle \nabla F(w^k), w - w^k \rangle + \frac{1}{2\alpha_k} \|w - w^k\|_2^2 \right\}.$$

If $W \equiv \mathbb{R}^d$, then

$$w^{k+1} = w^k - \alpha_k \nabla F(w^k).$$

Function class

Let the minimal assumption for the objective $f(w, Z)$ holds:

Definition (M-Lipschitz continuity)

A function $f : W \times \mathcal{Z} \to \mathbb{R}$ is M-Lipschitz continuous with respect to $w \in W$ in the norm $\|\cdot\|$ if for all $w, u \in W$

$$|F(w) - F(u)| \leq M\|w - u\|.$$

From this definition it follows that for all $w \in W$

$$\|\nabla F(w)\|_* \leq M,$$

where $\|\cdot\|_*$ is the dual norm:

$$\|s\|_* = \max_{z \in \mathcal{Z}} \{\langle s, w \rangle : \|z\| = 1\}.$$

Convergence of Projected GD

Theorem

Let F be convex and M-Lipschitz. Let w^1 be a starting point. Let $R = \|w^* - w^1\|$ be the radius of a solution and stepsize $\alpha = \frac{R}{M\sqrt{N}}$. Then for $\hat{w}^N = \frac{1}{N}\sum_{k=1}^{N} w^k$, the following holds

$$F(\hat{w}^N) - \min_{w \in W} F(w) \leq \frac{MR}{\sqrt{N}}.$$

Corollary

Projected GD after

$$N = \frac{M^2 R^2}{\varepsilon^2}$$

number of iterations outputs $\hat{w}^N = \frac{1}{N}\sum_{k=1}^{N} w^k$, for which we can guarantee $F(\hat{w}^N) - \min_{w \in W} F(w) \leq \varepsilon$.

Projected SGD

Question: What if $\nabla F(w)$ is unavalable, e.g., $F(w) = \mathbb{E}_Z f(w, Z)$.
Answer: Use stochastic gradient $\nabla f(w, Z)$ such that

- it is an unbiased estimate for all $w \in W$

$$\mathbb{E}_Z \nabla f(w, Z) = \nabla F(w)$$

- $\mathbb{E}_Z \left[\|\nabla f(x, Z)\|_*^2 \right] \leq M^2$.

Then iterations of Projected SGD:

$$w^{k+1} = \arg\min_{w \in W} \left\{ f(w^k, Z^k) + \langle \nabla_w f(w^k, Z^k), w - w^k \rangle + \frac{1}{2\alpha_k} \|w - w^k\|_2^2 \right\}.$$

Convergence of Projected SGD

Theorem

Let F be convex and M-Lipschitz. Let w^1 be a starting point. Let $R = \|w^* - w^1\|$ is the radius of a solution and stepsize $\alpha_k = \frac{R}{M\sqrt{k}}$ for each k. Then for $\hat{w}^N = \frac{1}{N}\sum_{k=1}^{N} w^k$, the following holds

$$\mathbb{E}F(\hat{w}^N) - \min_{w \in W} F(w) \leq \frac{3MR}{2\sqrt{N}}.$$

Fitting problem geometry

Question: How to fit the problem geometry?
Answer: Use the Bregman divergence instead of the Euclidezn projection

Definition

Let $\omega(w)$ be a 1-strongly convex function, then

$$V(w, y) = \omega(w) - \omega(y) - \langle \nabla\omega(y), w - y \rangle$$

is referred as the Bregman divergence. Function $w(w)$ is also known as the distance generating function (DGF), or prox-function.

Propertises:

- $V(w, y)$ is 1-strongly convex w.r.t. w
- $V(w, y) = 0$ iff $w = y$
- $V(w, y) \geq \frac{1}{2}\|w - y\|^2$ (from the definition of strong convexity of ω).

Bregman divergence. Examples

Euclidean setup

DGF $\omega(w) = \frac{1}{2}\|w\|_2^2$ is 1-strongly convex in the ℓ_2-norm. Then the Bregman divergence is

$$V(w, y) = \frac{1}{2}\|w - y\|^2.$$

Entropy setup

DGF $\omega(w) = \sum_{i=1}^d w_i \log w_i$ (negative entropy) is 1-strongly convex in the ℓ_1-norm. Then the Bregman divergence is KL-divergence

$$V(w, y) = \sum_{i=1}^d w_i \log \frac{w_i}{y_i}.$$

Stochastic Mirror Descent

Let $g^k \triangleq \nabla_w f(w^k, Z^k)$, then

Projected SGD repeats

$$w^{k+1} = \arg\min_{w \in W} \left\{ \langle g^k, w - w^k \rangle + \frac{1}{2\alpha_k} \|w - w^k\|_2^2 \right\}$$

Stochastic Mirror Descent repeats

$$w^{k+1} = \arg\min_{w \in W} \left\{ \langle g^k, w - w^k \rangle + \frac{1}{\alpha_k} V(w, w^k) \right\}$$

If $V(w, w^k) = \frac{1}{2}\|w - w^k\|_2^2$, then Stochastic Mirror Descent is the Projected SGD

Special case of Mirror Descent: Entropy setup

Let $V(w, y) = \sum_{i=1}^{d} w_i \log \frac{w_i}{y_i}$ (KL divergence). Then Mirror Descent has a closed-form: for all component $i = 1, \ldots, d$

$$w_i^{k+1} = \frac{w_i^k \exp(-\alpha_k g_i^k)}{\sum_{l=1}^{d} w_l^k \exp(-\alpha_k g_l^k)}$$

This is known as *entropic mirror descent*.

Convergence of Stochastic Mirror Descent

Theorem

Let F be convex and M-Lipschitz. Let w^1 be a starting point. Let stepsize $\alpha = \frac{R}{M\sqrt{k}}$ for each k with $R = V(w^*, w^1)$. Then for $\hat{w}^N = \frac{1}{N}\sum_{k=1}^{N} w^k$, the following holds

$$\mathbb{E}F(\hat{w}^N) - \min_{w \in W} F(w) \leq \frac{MR}{\sqrt{N}}.$$

Outline

SAA approach

Stochastic problem is replaced by its empirical counterpart

$$\min_{w \in W} \hat{F}(w) \triangleq \frac{1}{m} \sum_{i=1}^{m} f(w, z_i),$$

where $z_1, z_2, ..., z_m$ are the realizations of a random variable Z.

Questions:

- Why can we use $\hat{w}^* \triangleq \arg\min_{w \in W} \frac{1}{m} \sum_{i=1}^{m} f(w, z_i)$ instead of $w^* \triangleq \arg\min_{w \in W} \mathbb{E}_Z f(w, Z)$?
- How many samples should we take to approximate w^* by $\hat{w}$ with a desired accuracy ε w.r.t the non-optimality function gap.

Justification of the SAA approach

Law of Large Numbers

$$\hat{F}(w) \xrightarrow{a.s.} F(w) \quad \text{as} \quad m \to \infty,$$

where $\xrightarrow{a.s.}$ is almost surely converge.

Central Limit Theorem type results

$$\sqrt{m}(\hat{F}(w) - F(w)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(w)),$$

where $\xrightarrow{d}$ means the convergence in distribution and $\sigma^2(w) = \mathrm{Var}(f(w, Z))$.

Supervised Learning: Motivation

In machine learning: $f(w, z_i) \triangleq$ loss.

Definition (Expected risk)

Given a prediction function $h : X \to Y$, a loss function $l : Y \times Y \to \mathbb{R}$, and a probability distribution p on $\mathcal{X} \times \mathcal{Y}$ the expected risk of h is

$$\mathcal{R}(h) \triangleq \mathbb{E}l(y, h(x)) = \int_{\mathcal{Z}\times\mathcal{Y}} l(y, h(x))dp(x, y)$$

Remark: h depends on the random training data and not on the testing data

Definition (Empirical risk)

Given a prediction function $h : X \to Y$, a loss function $l : Y \times Y \to \mathbb{R}$, and data $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1, ..., m$, the empirical risk of h is defined as:

$$\hat{\mathcal{R}}(h) = \frac{1}{m}\sum_{i=1}^{m} l(y_i, h(x_i))$$

Special case: Linear regression

Regression and classification are the two main tasks in supervised learning.

Julie needs to decide whether she should go to the restaurant "Bamboo Garden" for lunch or not (give a rating) based on her and her friends historical ratings for the restaurants

Restaurant	Judy's rating	Jim's rating	Julie's ratings?
Goodfellas	1	5	2.5
Hakkasan	4.5	4	5
...	...	...	...
Bamboo Garden	3	3	?

Special case: Linear regression

Setup

- Predictor function $h(x_i) = \langle w, x_i \rangle$
- Quadratic loss $l(y, h(x_i)) = (y_i - h(x_i))^2$

Empirical risk minimization

$$\hat{w} \triangleq \arg\min_{w \in W} \sum_{i=1}^{m} (\langle w, x_i \rangle - y)^2 = \min_{w \in W} \|X^\top w - Y\|_2^2$$

where $X = [x_1 \dots x_m]$ is the design matrix and $Y = [y_1 \dots y_m]^\top$.

Analytical solution if $W = \mathbb{R}^d$

$$\hat{w} = (XX^\top)^{-1} XY.$$

Special case: Logistic regression (binary classification)

Suppose that Julie only cares about whether she will like the restaurant "Bamboo Garden" or not, rather her own ratings. Moreover, she only recorded some historical data indicating whether she likes or dislikes some restaurants.

Restaurant	Judy's rating	Jim's rating	Julie likes?
Goodfellas	1	5	No
Hakkasan	4.5	4	Yes
...	...	...	...
Bamboo Garden	3	3	?

Special case: Logistic regression (binary classification)

- Labels $y_i = \{-1, 1\}$
- Predictor function $h : W \times X \to [0, 1]$ is the probability that x is 1:

$$h(x) = \frac{1}{1 + \exp(-w^\top x)}$$

  is the sigmoid function
- Logistic loss

$$l(y, h(x)) = \log(1 + \exp(-y\langle w, x\rangle))$$

Empirical risk minimization (ERM)

$$\hat{w} \triangleq \min_{w \in \mathcal{W}} \frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp(y_i \langle w, x_i \rangle)).$$

The ERM problem can be efficiently solved by using numerical methods.

Statistics: Motivation

In statistics: $f(w, Z) = -\log p(w, Z)$ (negative log-density).

Definition (Maximum likelihood estimator)

Let $Z_1, ..., Z_m$ be i.i.d., then maximum likelihood estimator is

$$\hat{w} \triangleq \arg\max_{w \in W} \frac{1}{m} \sum_{i=1}^{m} \log p(w, Z_i).$$

Whereas the true w is $w^* \triangleq \arg\max_{w \in W} \mathbb{E} \log p(w, Z)$.

Properties of $\hat{w}$ (under some conditions):

- consistent: $\hat{w} \xrightarrow{\mathbb{P}} w$
- asymptotic normal $\sqrt{m}(\hat{w} - w) \xrightarrow{d} \mathcal{N}(0, (i(w))^{-1})$, where $i(w) = \mathbb{E}\left[\left(\frac{\partial}{\partial w} \log p(w, Z_1)\right)^2\right]$ is the Fisher information
- asymptotic optimal and efficient

How to choose m?

The number of realizations m is adjusted by the desired precision.

- Stochastic Approximation (SA): the number of iterations of SGD is

$$N = m = \mathcal{O}\left(\frac{M^2R^2}{\varepsilon^2}\right) \tag{1}$$

  to find an ε-solution $\hat{w}$. This is equivalent to the sample size of $z_1, ..., z_m$.

- Sample Average Approximation (SAA): the sample size is[1]

$$m = \widetilde{\mathcal{O}}\left(\frac{dM^2R^2}{\varepsilon^2}\right).$$

  This is d times more than in the SA approach (d is the problem's dimension).

[1] Alexander Shapiro and Arkadi Nemirovski. "On complexity of stochastic programming problems". In: *Continuous optimization*. Springer, 2005, pp. 111–146.

SAA: linear dependence on d

Moreover, the estimate on the sample size was obtained under the assumptions that empirical problem

$$\min_{w \in W} \hat{F}(w) \triangleq \frac{1}{m} \sum_{i=1}^{m} f(w, z_i)$$

is solved exactly.

! Linear dependence on d is the main drawback of the SAA approach.

Problem class: Lipschitz and strongly convex functions

Let $f(w, z)$ is M-Lipschitz continuous and γ-strongly convex.

Definition (γ-strong convexity)

A function $f : W \times z \to \mathbb{R}$ is λ-strongly convex with respect to w in norm $\|\cdot\|$ if it is continuously differential and it satisfies for all $w, y \in W,\ \forall Z \in \mathcal{Z}$

$$f(w, z) \geq f(y, z) + \langle \nabla f(y, z), w - y \rangle + \frac{\lambda}{2}\|w - y\|^2.$$

Sample size in the strongly convex case

If the objective $f(w, z)$ is λ-strongly convex in w, the sample sizes are equal up to logarithmic factors:

- Stochastic Approximation (SA): the number of iterations of SGD is

$$m = \mathcal{O}\left(\frac{M^2}{\lambda\varepsilon}\right).$$

- Sample Average Approximation (SAA): the sample size is

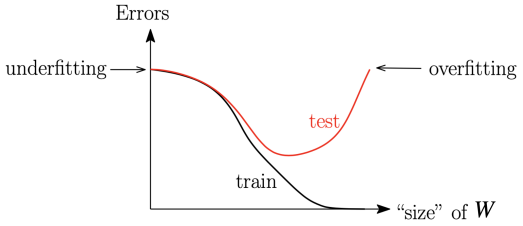
$$m = \widetilde{\mathcal{O}}\left(\frac{M^2}{\lambda\varepsilon}\right).$$

  Moreover, it suffices to solve empirical problem with accuracy $\varepsilon' = \mathcal{O}\left(\frac{\varepsilon^2\lambda}{M^2}\right)$.

Conclusion: to eliminate the linear dependence on d in the SAA approach for a non-strongly convex objective, regularization should be used.

Regularization: Machine Learning Motivation

In supervised learning: regularization is used to prevent overfitting



Capacity control. To avoid overfitting, we need to make sure that the set of allowed functions h is not too large by typically reducing the number of parameters or by restricting the norm of predictors.

Ridge regression

Capacity control can be done by regularization, the main example is *ridge regression*.

Linear regression

$$Y = \Psi^\top w + \epsilon,$$

where $\mathbb{E}\epsilon = 0$, and $\mathrm{Var}(\epsilon) = \sigma^2 \mathrm{I}$.

$$\hat{w} \triangleq \arg\min_{w \in W} \frac{1}{m} \sum_{i=1}^{m} (y_i - w^\top \psi(x_i))^2 + \lambda \|w\|_2^2,$$

where $\lambda > 0$.

In vector form

$$\hat{w} \triangleq \arg\min_{w \in W} \frac{1}{m} \|Y - \Psi^\top w\|_2^2 + \lambda \|w\|_2^2 = \left(\Psi\Psi^\top + \frac{\lambda}{m} I\right)^{-1} \Psi Y.$$

Regularization: Statistical Motivation

Bayesian statistics

Regularization is a prior belief about the problem.

Gaussian linear regression

$$Y = \Psi^\top w + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. Let Gaussian prior $w \sim \mathcal{N}(0, \sigma_\pi^2 I)$.

Then the Bayesian estimator under quadratic loss is

$$\hat{w} = \left(\Psi\Psi^\top + \frac{1}{m}\frac{\sigma^2}{\sigma_\pi^2} I \right)^{-1} \Psi Y.$$

Thus, this corresponds to the ridge regression with $\lambda = \sigma^2/\sigma_\pi^2$.

Regularization technique. How to find λ?

Let us suppose we need to solve

$$\min_{w \in W} f(w). \tag{2}$$

Let us regularize it by $\frac{1}{2}\|w - w^1\|_2^2$ (it is 1-strongly convex in the Euclidean norm), where w^1 is a starting point.

$$\min_{w \in W} f_\lambda(w) \triangleq f(w) + \frac{\lambda}{2}\|w - w^1\|_2^2. \tag{3}$$

Thus, f_λ is λ-strongly convex in the Euclidean norm.

Theorem

Let $w_{\varepsilon/2}$ is the $\varepsilon/2$-solution of problem (7) with

$$\lambda = \frac{\varepsilon}{\|w^* - w^1\|_2^2},$$

where $w^* = \arg\min_{w \in W} f(w)$. Then $w_{\varepsilon/2}$ will be an ε-solution of problem (6).

Regularization technique. Proof

Proof: We will use $f_\lambda(w) \triangleq f(w) + \frac{\lambda}{2}\|w - w^1\|_2^2$. Let us consider

$$f(w_{\varepsilon/2}) - f(w^*) \leq f_\lambda(w_{\varepsilon/2}) - f(w^*). \tag{4}$$

Then we use

$$\min_{w \in W} f_\lambda(w) = \min_{w \in W}\{f(w) + \frac{\lambda}{2}\|w - w^1\|_2^2\} \leq f(w^*) + \frac{\lambda}{2}\|w^*, w^1\|_2^2, \tag{5}$$

where we used that the minimal value is not bigger than a value of f in an arbitrary point, e.g., $w = w^*$. Then we choose $\lambda = \frac{\varepsilon}{\|w^* - w^1\|_2^2}$ in (5):

$$\min_{w \in W} f_\lambda(w) = \min_{w \in W}\{f(w) + \frac{\lambda}{2}\|w - w^1\|_2^2\} \leq f(w^*) + \frac{\lambda}{2}\|w^* - w^1\|^2 \leq f(w^*) + \varepsilon/2.$$

Thus from this $-f(w) \leq \min_{w \in W} f_\lambda(w) + \varepsilon/2$. Using this for (4), we obtain

$$f(w_{\varepsilon/2}) - f(w^*) \leq f_\lambda(w_{\varepsilon/2}) - f(w^*) \leq f_\lambda(w_{\varepsilon/2}) - \min_{w \in W} f_\lambda(w) + \varepsilon/2.$$

Therefore, since $w_{\varepsilon/2}$ is the $\varepsilon/2$- minimizer of f_λ, $w_{\varepsilon/2}$ is an ε- minimizer of f. $\square$

Strong convexity in the general norm

Question: What if we need strong convexity of f_λ not in the Euclidean norm but in the general norm?
Answer: Use the Bregman divergence

Regularization technique in the general norm.

Let us suppose we need to solve

$$\min_{w \in W} f(w). \tag{6}$$

Let us regularize it by $V(w, w^1)$ (it is 1-strongly convex in the ℓ_p-norm w.r.t. w)

$$\min_{w \in W} f_\lambda(w) \triangleq f(w) + \lambda V(w, w^1). \tag{7}$$

Thus, f_λ is λ-strongly convex in the Euclidean norm.

Theorem

Let $w_{\varepsilon/2}$ is the $\varepsilon/2$-solution of problem (7) with

$$\lambda = \frac{\varepsilon}{2V(w^*, w^1)},$$

where $w^* = \arg\min_{w \in W} f(w)$. Then $w_{\varepsilon/2}$ will be an ε-solution of problem (6).

Regularization

Let us suppose that $f(w, z)$ is merely convex in w and M-Lipschitz continuous.
Then, we regularize the stochastic problem

$$\min_{w \in W} \mathbb{E} f(w, z) + \lambda V(w, w^1),$$

where $\lambda = \frac{\varepsilon}{2V(w^*, w^1)}$. The empirical counterpart of this problem is

$$\min_{w \in W} \frac{1}{m} \sum_{i=1}^{m} f(w, z_i) + \lambda V(w, w^1)$$

Then the sample sizes in the SA and the SAA approaches will be equal up to logarithmic terms.

Randomization

Question: Can we apply stochastic methods, e.g. SGD, when the problem is deterministic?
Answer: Yes.
Let we are given the empirical problem

$$\min_{w \in W} \hat{F}(w) \triangleq \frac{1}{m} \sum_{i=1}^{m} f(w, z_i). \tag{8}$$

Let us introduce $f_i(w) \triangleq f(w, z_i)$. Then we can reformulate problem (8) as

$$\min_{w \in W} \hat{F}(w) \triangleq \frac{1}{m} \sum_{i=1}^{m} f_i(w).$$

SGD for Finite Sum

SGD repeats:

$$w^{k+1} = w^k - \alpha_k \nabla f_{i_k}(x^k)$$

where $i_k \in 1, ..., m$ is some chosen index at iteration k.

Randomized rule: choose $i \in 1, ..., m$ uniformly at random

$$\mathbb{P}(\xi = i) = \frac{1}{m}$$

then $\nabla f_{i_k}(x^k)$ is an unbiased estimate of the full gradient:

$$\mathbb{E}_\xi[\nabla f_\xi(w)] = \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(w).$$

Step Sizes

Standard in SGD is to use diminishing step sizes, e.g., $\alpha_k = 1/k$.

Why not fixed step sizes?

For instance, $W \equiv \mathbb{R}^n$:

$$\frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\hat{w}^*) = 0, \quad \text{whereas} \quad \nabla f_i(\hat{w}^*) \neq 0.$$

Convergence Rates

Recall: for convex F, GD with fixed stepsize satisfies

$$F(\hat{w}^N) - F^* = \mathcal{O}(1/\sqrt{N})$$

When F is differentiable with Lipschitz gradient, we get for fixed stepsize

$$F(\hat{w}^N) - F^* = \mathcal{O}(1/N)$$

What about SGD?

For convex functions, SGD with diminishing stepsize satisfies

$$\mathbb{E}[F(\hat{w}^N)] - F^* = \mathcal{O}(1/\sqrt{N})$$

Unfortunately this does not improve when we further assume function has Lipschitz gradient.

Convergence rates

When F is strongly convex and has a Lipschitz gradient, GD satisfies

$$F(w^N) - F^* = \mathcal{O}(\gamma^N)$$

where $0 < \gamma < 1$.
But under same conditions, SGD gives

$$\mathbb{E}[F(w^N)] - F^* = \mathcal{O}(1/N).$$

So stochastic methods do not enjoy the linear convergence rate of GD under strong convexity.