CSIE5410 Optimization algorithms

Lecture 6: Composite convex optimization

Yen-Huan Li (yenhuan.li@csie.ntu.edu.tw)

18.10.2018

Department of Computer Science and Information Engineering
National Taiwan University

Abstract

Consider the optimization problem

$$F^\star = \min_x \left\{ f(x) + g(x) \mid x \in \mathbb{R}^p \right\},$$

for some smooth convex function f and proper closed convex function g. How does one solve the problem efficiently?

This lecture introduces the notion of a *proximal operator*, and the *proximal point method* and *proximal gradient method* to solve the problem.

Recommended reading

- J. Eckstein. 1989. *Splitting Methods for Monotone Operators with Applications to Parallel Optimization*. (Chapter 3).
- P. L. Combettes and J.-C. Pesquet. 2011. Proximal splitting methods in signal processing.
- Yu. Nesterov. 2005. Smooth minimization of non-smooth functions.
- *H. H. Bauschke and P. L. Combettes. 2011. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*.

Table of contents

Examples of composite convex optimization problems

Problem formulation

Consider the problem

$$F^\star = \min_x \{ f(x) + g(x) \mid x \in \mathbb{R}^p \},$$

where f is a convex L-smooth function, and g is a proper closed convex function.

Example. Let g be the indicator function of a closed convex set. Then the problem is equivalent to minimizing f subject to the constraint that $x \in \mathcal{X}$.

Example: High-dimensional linear regression (1/3)

Consider the linear regression model

$$y_i \coloneqq \langle x_i, \beta^\natural \rangle + w_i, \quad i \in \mathbb{N},$$

where $x_i, \beta \in \mathbb{R}^p$ and w_i are i.i.d. Gaussian r.v.'s of mean zero and unit variance. The goal is to estimate β given the data $(x_1, y_1), \ldots, (x_n, y_n)$.

Problem. Is it possible to estimate $\beta^\natural$ accurately, even when $n \ll p$ (i.e., when the data size is much smaller than the ambient dimension)?

T. Hastie *et al.* 2015. *Statistical Learning with Sparsity: The Lasso and generalizations*.

Example: High-dimensional linear regression (2/3)

Assumption. The vector $\beta^\natural$ is s-sparse with $s \ll p$.

Lasso. The lasso (aka the ℓ_1-penalized least squares estimator) is given by

$$\hat{\beta}_n \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \langle x_i, \beta \rangle\right)^2 + \lambda_n \|\beta\|_1 \;\middle|\; \beta \in \mathbb{R}^p \right\},$$

for some properly chosen *penalization parameter* $\lambda_n > 0$.

Proposition. The vector $\hat{\beta}_n$ is typically sparse. (Why?)

R. Tibshirani. 1996. Regression shrinkage and selection via the lasso.

Example: High-dimensional linear regression (3/3)

Theorem. Suppose that x_i's are i.i.d. Gaussian random vectors of zero mean and identity covariance matrix. Then with $\lambda_n = \sqrt{c(\log p)/n}$ for some $c > \sqrt{2}$, it holds that with probability at least $1 - O(p^{1-c^2/2})$,

$$\|\hat{\beta}_n - \beta^\natural\|_2 = O\left(\sqrt{\frac{s \log p}{n}}\right).$$

Remark. Therefore, a data size of $O(s \log p)$ suffices for good statistical accuracy.

P. Bickel. 2009. Simultaneous analysis of Lasso and Dantzig selector.

Example: High-dimensional statistics

The idea can be easily applied to other cases.

Example. (Sparse logistic regression)

$$\hat{w} \in \arg\min_{w} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \mathrm{e}^{-y_i \langle x_i, w \rangle}\right) + \lambda_n \|w\|_1 \;\middle|\; w \in \mathbb{R}^p \right\}.$$

Example. (Sparse positron emission tomography)

$$\hat{x} \in \arg\min_{x} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[y_i \langle a_i, x \rangle - \log \langle a_i, x \rangle \right] + \lambda_n \|x\|_1 \;\middle|\; x \in \Delta \subset \mathbb{R}^p \right\}$$

T. Hastie *et al.* 2015. *Statistical Learning with Sparsity: The Lasso and generalizations*.

M. Raginsky *et al.* 2010. Compressed sensing performance bounds under Poisson noise.

Example: Low-rank matrix estimation (1/3)

Let $X^\natural :=\in \mathbb{R}^{p_1 \times p_2}$. Suppose that we only observe

$$y_i := \mathrm{tr}(A_i^{\mathrm{T}} X^\natural) + w_i, \quad i = 1, \ldots, n,$$

for some matrices $A_1, \ldots, A_n$, where w_i denotes the noise.

Question. How do we recover $X^\natural$ given the observations?

Remark. Application include recommender systems, quantum state tomography, phase retrieval, etc.

E. Candès and B. Recht. 2009. Exact matrix completion via convex optimization.

D. Gross *et al.* 2010. Quantum state tomography via compressed sensing.

E. J. Candès *et al.* 2013. Phase retrieval via matrix completion.

Example: Low-rank matrix estimation (2/3)

Matrix lasso. With regard to the linear regression case, a natural estimator is the following:

$$\hat{X}_n \in \arg\min_X \left\{ \sum_{i=1}^{n} \big(y_i - \mathrm{tr}(A_i^{\mathrm{T}} X)\big)^2 + \lambda_n \|X\|_{S^1} \;\middle|\; X \in \mathbb{R}^{p_1 \times p_2} \right\},$$

for some properly chosen penalization parameter $\lambda_n > 0$.

E. J. Candès and Y. Plan. 2011. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements.

S. Negahban and M. J. Wainwright. 2011. Estimation of (near) low-rank matrices with noise and high-dimensional scaling.

Example: Low-rank matrix estimation (3/3)

Theorem. Suppose that the entries of A_i are i.i.d. Gaussian r.v.'s of zero mean and unit variance. Suppose that the matrix $X^\natural$ is of rank $r \in \mathbb{N}$. Then, with probability at least $1 - O(\mathrm{e}^{-c(p_1+p_2)})$ for some $c > 0$, it holds that

$$\|\hat{X}_n - X^\natural\|_{\mathsf{F}} = O\left(\sqrt{\frac{r(p_1+p_2)}{n}}\right).$$

Remark. Notice that in general, we need $n = \Omega(p_1 \times p_2)$ observations to have a small estimation error.

S. Negahban and M. J. Wainwright. 2011. Estimation of (near) low-rank matrices with noise and high-dimensional scaling.

Example: Total variation denoising (1/3)

Consider the problem of estimating an image $x^\natural \in \mathbb{R}^{d\times d}$ given

$$y := x^\natural + \varepsilon,$$

where ε is a matrix of i.i.d. standard Gaussian r.v.'s.

The ML estimator of $x^\natural$ is given by

$$\hat{x}_{\mathsf{ML}} \in \arg\min_x \left\{ \|y - x\|_{\mathsf{F}}^2 \;\middle|\; x \in \mathbb{R}^{d\times d} \right\}.$$

Then we get $\hat{x}_{\mathsf{ML}} = y$.

Question. How do we get a non-trivial estimate?

Example: Total variation denoising (2/3)

Definition. Define

$$D_1 \coloneqq \begin{bmatrix} +1 & 0 & 0 & \cdots & 0 \\ -1 & +1 & \ddots & \ddots & \vdots \\ 0 & -1 & +1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 & +1 \end{bmatrix}, \quad D_2 \coloneqq \begin{bmatrix} I \otimes D_1 \\ D_1 \otimes I \end{bmatrix}.$$

The total variation of an image $x \in \mathbb{R}^{d \times d}$ is given by

$$V(x) \coloneqq \|D_2 \operatorname{vec}(x)\|_1.$$

L. I. Rudin *et al.* 1992. Nonlinear total variation based noise removal algorithms.

Example: Total variation denoising (3/3)

Fact. The total variation of an image is typically small.

Total variation denoising.

$$\hat{x} \in \arg\min_x \left\{ \frac{1}{d^2}\|y - x\|_2^2 + \lambda_d V(x) \;\middle|\; x \in \mathbb{R}^{d\times d} \right\},$$

for some penalization parameter $\lambda_d > 0$.

Theorem. For a properly chosen λ_d, it holds with probability at least 0.9 that

$$\frac{1}{d^2}\|\hat{x} - x^\natural\|_{\mathsf{F}}^2 = O\left(\frac{\min\left\{ \|D_2\mathrm{vec}(x^\natural)\|_0, \|D_2\mathrm{vec}(x^\natural)\|_1 \right\}}{d^2} \log^2 d \right).$$

J.-C. Hütter and P. Rigollet. 2016. Optimal rates for total variation denoising.

Proximal gradient algorithm

Fixed-point characterization of a minimizer (1/2)

Fermat's rule says that $x^\star \in \mathbb{R}^p$ is a minimizer of $f + g$ on $\mathbb{R}^p$, if and only if

$$0 \in \partial(f+g)(x^\star).$$

Suppose that $\mathrm{ri}(\mathrm{dom}\, f) \cap \mathrm{ri}(\mathrm{dom}\, g) \neq \emptyset$. We write

$$0 \in \partial f(x^\star) + \partial g(x^\star) = \nabla f(x^\star) + \partial g(x^\star).$$

Then, we have

$$-\eta \nabla f(x^\star) \in \eta \partial g(x^\star), \quad \forall \eta > 0.$$

Fixed-point characterization of a minimizer (2/2)

Define the identity mapping $I : x \mapsto x$. We write

$$(I - \eta\nabla f)(x^\star) \in (I + \eta\partial g)(x^\star).$$

That is,

$$x^\star \in (I + \eta\partial g)^{-1}(I - \eta\nabla f)(x^\star).$$

Suppose that $(I + \eta\partial g)^{-1}$ is single-valued. Then we obtain a fixed-point characterization of $x^\star$

$$x^\star = (I + \eta\partial g)^{-1}(I - \eta\nabla f)(x^\star).$$

Proximal gradient algorithm

Definition. Let g be a proper closed convex function on $\mathbb{R}^p$. The proximal mapping associated with g is given by

$$\operatorname{prox}_g(x) := (I + \partial g)^{-1}(x),$$

Proximal gradient algorithm.

$$x_t \leftarrow (I + \eta_{t-1}\partial g)^{-1}(I - \eta_{t-1}\nabla f)x_{t-1}, \text{ or}$$
$$x_t \leftarrow \operatorname{prox}_{\eta_{t-1}g}(x_{t-1} - \eta_{t-1}\nabla f(x_{t-1})).$$

J.-J. Moreau. 1962. Fonctions convexes duales et points proximaux dans un espace hilbertien.

P. L. Lions and B. Mercier. 1979. Splitting algorithms for the sum of two nonlinear operators.

Proximal gradient algorithm.

$$x_t \leftarrow (I + \eta_{t-1}\partial g)^{-1}(I - \eta_{t-1}\nabla f)x_{t-1}, \text{ or}$$
$$x_t \leftarrow \operatorname{prox}_{\eta_{t-1}g}(x_{t-1} - \eta_{t-1}\nabla f(x_{t-1})).$$

Question. Is the algorithm well-defined? In particular, is $(I + \eta\partial g)^{-1}x$ well-defined for every $\eta > 0$ and $x \in \mathbb{R}^p$?

Question. How does one compute the proximal mapping?

Question. Does the algorithm converge? What is the convergence rate?

Properties of the proximal mapping

Theorem. Let f be a proper closed convex function on $\mathbb{R}^p$. Then the proximal mapping prox_f is well-defined everywhere on $\mathbb{R}^p$, and has the equivalent formulation

$$\operatorname{prox}_f(x) = \arg\min_y \left\{ f(y) + \frac{1}{2}\|y - x\|_2^2 \;\middle|\; y \in \mathbb{R}^p \right\}, \quad \forall x \in \mathbb{R}^p.$$

Sanity check. Let $z \in (I + \partial f)^{-1} x$. Then $x \in z + \partial f(z)$; equivalently, we write $0 \in \partial f(z) + (z - x)$, which is the optimality condition for the expression above.

H. H. Bauschke and P. L. Combettes. 2011. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*.

Sketch of the formal proof (1/3): Monotone operators

Definition. We say that a set-valued operator $A : \mathbb{R}^p \to 2^{\mathbb{R}^p}$ is monotone, if and only if

$$\langle x_A - y_A, x - y \rangle \geq 0, \quad \forall x, y \in \mathbb{R}^p, x_A \in Ax, y_A \in Ay.$$

Definition. We say that a set-valued operator A is maximally monotone, if and only if it is monotone and

$$u \in Ax \Leftrightarrow \forall y \in \operatorname{dom} A \text{ and } v \in Ay : \langle u - v, x - y \rangle \geq 0$$

H. H. Bauschke and P. L. Combettes. 2011. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*.

Sketch of the formal proof (2/3): Maximal monotonicity implies well-defined proximal mapping

Theorem. (Minty's theorem) A set-valued operator $A : \mathbb{R}^p \to 2^{\mathbb{R}^p}$ is maximally monotone, if and only if $(I + A)^{-1}$ is well-defined everywhere on $\mathbb{R}^p$.

Theorem. Let f be a proper closed convex function on $\mathbb{R}^p$. Then ∂f is maximally monotone.

Remark. Minimizing a proper closed convex function f is then equivalent to solving the monotone inclusion problem $0 \in \partial f(x^\star)$.

Corollary. Therefore, $(I + \partial f)^{-1}$ is well-defined everywhere in $\mathbb{R}^p$.

H. H. Bauschke and P. L. Combettes. 2011. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*.

Sketch of the formal proof (3/3): Uniqueness

Proposition. Let f be a proper closed convex function on $\mathbb{R}^p$. The associated proximal mapping is single-valued and can be computed as

$$(I + \partial f)^{-1} x = \arg\min_{y} \left\{ f(y) + \frac{1}{2}\|y - x\|_2^2 \;\middle|\; y \in \mathbb{R}^p \right\}.$$

Proof. The formulation is verified via the optimality condition (recall our sanity check). Notice that the function $\varphi(y) \coloneqq f(y) + (1/2)\|y - x\|_2^2$ is 1-strongly convex. Define $y^\star := (I + \partial f)^{-1} x$. We write

$$\varphi(y) \geq \varphi(y^\star) + \langle \nabla \varphi(y^\star), y - y^\star \rangle + \frac{1}{2}\|y - y^\star\|_2^2, \quad \forall y \in \mathbb{R}^p.$$

As $0 \in \partial \varphi(y^\star)$, the proposition follows.

Examples of the proximal mapping

Two simple cases

Example. Let $f(x) \coloneqq (1/2)\|x\|_2^2$. Then $\mathrm{prox}_f(x) = (1/2)x$.

Proof. Simple exercise.

Example. Let $f(x)$ be the indicator function of a closed convex set $\mathcal{X} \subseteq \mathbb{R}^p$. Then $\mathrm{prox}_f(x) = \mathrm{proj}_{\mathcal{X}}(x)$.

Proof. We write

$$\begin{aligned}\mathrm{prox}_f(x) &= \arg\min_y \left\{ f(y) + \frac{1}{2}\|y-x\|_2^2 \;\middle|\; y \in \mathbb{R}^p \right\} \\ &= \arg\min_y \left\{ \frac{1}{2}\|y-x\|_2^2 \;\middle|\; y \in \mathcal{X} \right\}.\end{aligned}$$

Absolute value

Example. Let $f(x) \coloneqq |x|$ on $\mathbb{R}$. Then $\operatorname{prox}_f(x) = \operatorname{soft}_1(x)$, where the *soft thresholding operator* is given by

$$\operatorname{soft}_\lambda(x) \coloneqq \begin{cases} 0, & x \in [-\lambda, \lambda], \\ x - \lambda, & x > \lambda, \\ x + \lambda, & \text{otherwise.} \end{cases}$$

Proof. Recall that

$$y^\star \coloneqq \operatorname{prox}_f(x) = \arg\min_y \left\{ |y| + \frac{1}{2}(y - x)^2 \;\middle|\; y \in \mathbb{R} \right\}.$$

The optimality condition says that

$$0 \in \partial(|\cdot|)(y^\star) + (y^\star - x),$$

which is satisfied by the expression of prox_f above.

ℓ_1- and Schatten 1-norms

Example. Let $f(x) := \|x\|_1$ on $\mathbb{R}^p$. Then $\operatorname{prox}_f(x) = \operatorname{soft}_1(x)$, where $\operatorname{soft}_\lambda$ denotes the elementwise soft thresholding operator.

Proof. Notice that

$$\|y\|_1 + \frac{1}{2}\|y - x\|_2^2 = \sum_{i=1}^{p} \left[|y^{(i)}| + \frac{1}{2}\left(y^{(i)} - x^{(i)}\right)^2 \right].$$

Example. Let $f(X) := \|X\|_{S^1}$ on $\mathbb{R}^{m \times n}$. Let $X = U \operatorname{diag}(\sigma) V^{\mathrm{T}}$ be the singular value decomposition (SVD) of X. Then

$$\operatorname{prox}_f(X) = U \operatorname{diag}\left(\operatorname{soft}_1(\sigma)\right) V^{\mathrm{T}},$$

where $\operatorname{soft}_\lambda$ denotes the elementwise soft thresholding operator.

Derivation of the proximal mapping of the Schatten 1-norm (1/3)

Definition. The Frobenius norm of a matrix $X \in \mathbb{R}^{m \times n}$ is given by

$$\|X\|_{\mathsf{F}} := \sqrt{\operatorname{tr}(X^{\mathrm{T}} X)}.$$

Proposition. For any matrix $X \in \mathbb{R}^{m \times n}$, we have

$$\|X\|_{\mathsf{F}} := \|X\|_{S^2}$$

.

Corollary. Let $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ be unitary matrices. Then

$$\|X\|_{\mathsf{F}} = \|UX\|_{\mathsf{F}} = \|XV\|_{\mathsf{F}}.$$

Derivation of the proximal mapping of the Schatten 1-norm (2/3)

Definition. For any matrices $X \in \mathbb{R}^{m\times n}$, its singular value decomposition (SVD) always exists, and is uniquely defined as

$$X = U\,\mathrm{diag}(\sigma)V^{\mathrm{T}},$$

for some unitary matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$, and $\sigma \in \mathbb{R}^{\min\{\,m,n\,\}}$ of non-negative entries (called singular values).

Theorem. Let $X, Y \in \mathbb{R}^{m\times n}$ and σ_X and σ_Y be their vectors of singular values, respectively. Then

$$\mathrm{tr}(X^{\mathrm{T}}Y) \le \langle \sigma_X, \sigma_Y \rangle\,;$$

the equality holds if and only if $X = U\,\mathrm{diag}(\sigma_X)V^{\mathrm{T}}$ and $Y = U\,\mathrm{diag}(\sigma_Y)V^{\mathrm{T}}$ for some unitary matrices U and V.

Derivation of the proximal mapping of the Schatten 1-norm (3/3)

Proof. For any $Y \in \mathbb{R}^{m\times n}$, let $Y = A\,\mathrm{diag}(\rho)B^{\mathrm{T}}$ be its SVD. We write

$$Z := \mathrm{prox}_f(X) = \arg\min_Y \left\{ \|Y\|_{S^1} + \frac{1}{2}\|Y - X\|_{\mathsf{F}}^2 \;\middle|\; Y \in \mathbb{R}^{m\times n} \right\}.$$

Notice that $\|Y\|_{S^1} = \|\rho\|_1$. Also notice that

$$\begin{aligned} \|Y - X\|_{\mathsf{F}}^2 &= \|Y\|_{\mathsf{F}}^2 - 2\,\mathrm{tr}(Y^{\mathrm{T}}X) + \|X\|_{\mathsf{F}}^2 \\ &\geq \|\rho\|_2^2 - 2\,\langle \rho, \sigma\rangle + \|X\|_{\mathsf{F}}^2, \end{aligned}$$

and the equality holds if and only if $A = U$ and $B = V$. Then we get the desired formula of $\mathrm{prox}_f(X)$.

Total variation

Recall that the total variation function is given by

$$V(X) := \|D_2 \operatorname{vec}(X)\|_1 = (\|\cdot\|_1 \circ D_2)(\operatorname{vec}(X)).$$

Proposition. Let $f := g \circ A$ on $\mathbb{R}^p$ for some proper closed convex function g on $\mathbb{R}^m$ and $A \in \mathbb{R}^{m \times p}$. If $AA^{\mathrm{T}} = \nu I$ for some $\nu > 0$, then $\operatorname{prox}_f(x) = x + \nu^{-1} A^{\mathrm{T}} \big(\operatorname{prox}_{\nu g}(Ax) - Ax\big)$.

Remark. If $AA^{\mathrm{T}} \neq \nu I$ for all $\nu > 0$ (e.g., when $A = D_2$), then one needs an optimization algorithm to compute prox_f.

P. L. Combettes and J.-C. Pesquet. 2011. Proximal splitting methods in signal processing.

A. Beck and M. Teboulle. 2009. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems.

Proximal point method & smoothing

Proximal point method

Consider the problem of minimizing a proper closed convex function f on $\mathbb{R}^p$.

Viewing the problem as minimizing $f + g$ with $g(x) := 0$ for all $x \in \mathbb{R}^p$, we obtain the proximal point method.

Algorithm Proximal point method

1: Set $x_0 \in \mathbb{R}^p$.
2: **for** $t = 1, \ldots, T$ **do**
3: $\quad x_t \leftarrow (I + \eta_{t-1}\partial f)^{-1} x_{t-1} = \mathrm{prox}_{\eta_{t-1} f}(x_{t-1})$
4: **end for**

B. Martinet. 1970. Regularisation d'inequations variationnelles par approximations successives.

Smoothing interpretation

The major difficulty in minimizing the function f is that it may not be smooth.

Question. How do we find a smooth approximation of the function f?

Definition. Let f be a proper closed convex function on $\mathbb{R}^p$. Its Moreau envelope is given by

$$f_\eta(x) \coloneqq \min_y \left\{ f(y) + \frac{1}{2\eta}\|x - y\|_2^2 \;\middle|\; y \in \mathbb{R}^p \right\}.$$

J.-J. Moreau. 1965. Proximité et dualité dans un espace hilbertien.

Smoothing interpretation

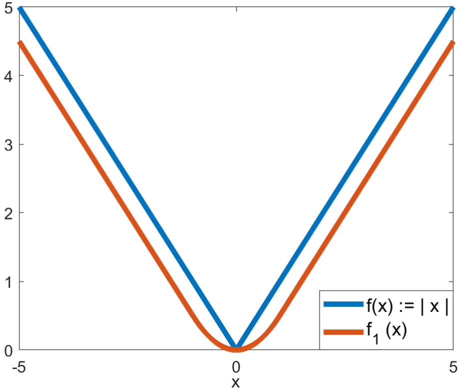
Theorem. (Moreau's theorem) Let f be proper closed convex on $\mathbb{R}^p$. The Moreau envelope f_η is convex, differentiable, and $(1/\eta)$-smooth on $\mathbb{R}^p$. Moreover,

$$\nabla f_\eta(x) = \frac{1}{\eta}(x - \mathrm{prox}_{\eta f}(x)), \quad \forall x \in \mathbb{R}^p.$$

Remark. One may also write $\nabla f_\eta(x) = \eta^{-1}(I - (I + \eta\partial f)^{-1})x$. The operator $\eta^{-1}(I - (I + \eta\partial f)^{-1})$ is called the *Yosida approximation* of ∂f.

J.-J. Moreau. 1965. Proximité et dualité dans un espace hilbertien.
K. Yosida. 1964. *Functional Analysis*.

Illustration: f vs. f_μ



Remark. The Moreau envelope of $|\cdot|$ is called the Huber loss.

P. J. Huber. 1964. Robust estimation of a location parameter.

Interpretation of the proximal point method

Algorithm Proximal point method

Set $x_0 \in \mathbb{R}^p$.
for $t = 1, \ldots, T$ **do**
    $x_t \leftarrow (I + \eta_{t-1}\partial f)^{-1} x_{t-1} = \mathrm{prox}_{\eta_{t-1} f}(x_{t-1})$
end for

The proximal point method iteratively does the following:

1. Construct a smooth approximation $f_{\eta_{t-1}}$ of the objective function f, which is $(1/\eta_{t-1})$-smooth.
2. Do a standard gradient descent step on $f_{\eta_{t-1}}$, i.e.,

$$\begin{aligned} x_t &= x_{t-1} - \eta_{t-1} \nabla f_{\eta_{t-1}}(x_{t-1}) \\ &= x_{t-1} - \eta_{t-1} \left[\frac{1}{\eta_{t-1}} \left(x_{t-1} - \mathrm{prox}_{\eta_{t-1} f}\, x_{t-1} \right) \right]. \end{aligned}$$

Checking smoothness of the Moreau envelope (1/3)

Strategy. Derive ∇f_η first, and then check for the convexity and smoothness of f_η.

Lemma. Let $p_x := \mathrm{prox}_f(x)$. Then

$$f(y) \geq f(p_x) + \langle x - p_x, y - p_x \rangle\,, \quad \forall y \in \mathbb{R}^p.$$

Proof. By definition, we have that $p_x = (I + \partial f)^{-1}(x)$, or

$$x - p_x \in \partial f(p_x).$$

The lemma follows from the definition of a subdifferential.

Checking smoothness of the Moreau envelope (2/3)

Theorem. The proximal mapping is firmly non-expansive, i.e.,

$$\|p_y - p_x\|_2^2 \leq \langle p_y - p_x, y - x \rangle, \quad \forall x, y \in \mathbb{R}^p,$$

where $p_y := \operatorname{prox}_f(y)$ and $p_x := \operatorname{prox}_f(x)$.

Proof. We write

$$f(p_x) \geq f(p_y) + \langle y - p_y, p_x - p_y \rangle,$$
$$f(p_y) \geq f(p_x) + \langle x - p_x, p_y - p_x \rangle.$$

Summing up the two inequalities, the theorem follows.

Corollary. The proximal mapping is non-expansive. In particular, projection onto a closed convex set is non-expansive.

Checking smoothness of the Moreau envelope (3/3)

Proposition. The gradient of f_η is $(1/\eta)$-Lipschitz continuous on $\mathbb{R}^p$. Therefore, the Moreau envelope f_η is $(1/\eta)$-smooth on $\mathbb{R}^p$.

Proof. Define $p_y := \operatorname{prox}_{\eta f}(y)$ and $p_x := \operatorname{prox}_{\eta f}(x)$. We write

$$
\begin{aligned}
&\|\nabla f_\eta(y) - \nabla f_\eta(x)\|_2^2 \\
&\quad = \left\|\eta^{-1}\left(y - p_y\right) - \eta^{-1}\left(x - p_x\right)\right\|_2^2 \\
&\quad \le \eta^{-2}\left(\|y - x\|_2^2 + \|p_y - p_x\|_2^2 - 2\left\langle p_y - p_x, y - x\right\rangle\right) \\
&\quad \le \eta^{-2}\left(\|y - x\|_2^2 - \|p_y - p_x\|_2^2\right).
\end{aligned}
$$

Remark. See Proposition 12.29 in the reference for $\nabla f_\eta(x)$.

H. H. Bauschke and P. L. Combettes. 2011. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*.

Direct smoothing approach to convex optimization (1/3)

Consider the problem of minimizing a proper closed convex function f on a convex bounded closed set $\mathcal{X} \subset \mathbb{R}^p$.

Theorem. Suppose that $\partial f(x) \neq \emptyset$ on $\mathcal{X}$. Suppose that there exists some $L > 0$, such that for every $x \in \mathcal{X}$, there is some $\nabla f(x) \in \partial f(x)$ such that $\|\nabla f(x)\|_2 \leq L$. Then, it holds that

$$f(x) - \frac{\eta L^2}{2} \leq f_\eta(x) \leq f(x).$$

Remark. Therefore, we may run the accelerated gradient method (next lecture) on f_η to solve the original optimization problem.

Yu. Nesterov. 2005. Smooth minimization of nonsmooth functions.

A. Beck and M. Teboulle. 2012. Smoothing and first order methods: A unified framework.

Proof of the approximation error bound for f_η

Proof. We write

$$f_\eta(x) \le f(x) + \frac{1}{2\eta}\|x - x\|_2^2 = f(x).$$

Moreover, we write, for all $\nabla f(x) \in \partial f(x)$,

$$\begin{aligned} f_\eta(x) - f(x) &= \min_y \left\{ f(y) - f(x) + \frac{1}{2\eta}\|y - x\|_2^2 \;\middle|\; y \in \mathbb{R}^p \right\} \\ &\ge \min_y \left\{ \langle \nabla f(x), y - x \rangle + \frac{1}{2\eta}\|y - x\|_2^2 \;\middle|\; y \in \mathbb{R}^p \right\} \\ &= -\frac{\eta}{2}\|\nabla f(x)\|_2^2. \end{aligned}$$

Direct smoothing approach to convex optimization (2/3)

As f_η is $(1/\eta)$-smooth, the accelerated gradient descent achieves

$$f_\eta(x_t) - f_\eta^\star \le \frac{4R_{\mathcal{X}}^2}{\eta(t+2)^2},$$

where $f_\eta^\star$ and $x_\eta^\star$ denotes the minimum value and a minimizer, respectively, and $R_{\mathcal{X}} := \max_x \left\{ \|x - x_0\|_2^2 \;\middle|\; \in \mathcal{X} \right\}$.

Proposition. Let $\varepsilon > 0$. To find some iterate x_T such that $f(x_T) - f^\star \le \varepsilon$, it suffices to set

$$T = \frac{2\sqrt{2}RL}{\varepsilon} = O\left(\frac{RL}{\varepsilon}\right), \quad \eta = \frac{2\sqrt{2}R}{L(t+2)}.$$

Direct smoothing approach to convex optimization (3/3)

Proof of the proposition. We write

$$\begin{aligned} f(x_t) - f^\star &= (f(x_t) - f_\eta(x_t)) + \left(f_\eta(x_t) - f_\eta^\star\right) + \left(f_\eta^\star - f^\star\right) \\ &\le \frac{\eta L^2}{2} + \frac{4R^2}{\eta(t+2)^2}. \end{aligned}$$

By the inequality of arithmetic and geometric means, the optimal value of η is given by

$$\eta = \sqrt{\frac{8R^2}{L^2(t+2)^2}}.$$

It is then direct to find T to achieve ε numerical error.

Nesterov's smoothing

Reformulation

Our *directly smoothing* approach is to instead minimize

$$f_\eta(x) := \min_y \left\{ f(y) + \frac{1}{2\eta}\|y - x\|_2^2 \;\middle|\; y \in \mathcal{X} \right\}$$

on the constraint set $\mathcal{X}$, for some properly chosen $\eta > 0$.

Proposition. We can equivalently write

$$f_\eta(x) := \max_y \left\{ \langle y, x\rangle - f^*(y) - \frac{\eta}{2}\|y\|_2^2 \;\middle|\; y \in \mathcal{X} \right\},$$

where f^* is the *conjugate* of f, defined as

$$f^*(y) := \max_z \{ \langle y, z\rangle - f(z) \mid z \in \mathbb{R}^p \}.$$

D. P. Bertsekas. 1996. *Constrained Optimization and Lagrange Multiplier Methods*.

Nesterov's smoothing (1/4)

Consider the optimization problem

$$f^\star = \min_x \{ f(x) \mid x \in \mathcal{X} \},$$

where

$$f(x) \coloneqq \max_y \{ \langle Ax, y \rangle - \varphi(y) \mid y \in \mathcal{Y} \},$$

for some matrix $A \in \mathbb{R}^{q \times p}$, bounded closed convex sets $\mathcal{X} \subset \mathbb{R}^p$ and $\mathcal{Y} \subset \mathbb{R}^q$, and convex function φ continuous on $\mathcal{Y}$.

Remark. This is not a black-box model.

Nesterov's smoothing (2/4)

Define for any $\eta \geq 0$,

$$f_\eta(x) \coloneqq \max_y \left\{ \langle Ax, y \rangle - \varphi(y) - \frac{\eta}{2}\|y\|_2^2 \;\middle|\; y \in \mathcal{Y} \right\}.$$

Proposition. The function f_η is well-defined and convex on $\mathbb{R}^p$. It holds that

$$f_\eta(x) \leq f(x) \leq f_\eta(x) + \max_y \left\{ \frac{\eta}{2}\|y\|_2^2 \;\middle|\; y \in \mathcal{Y} \right\}.$$

Yu. Nesterov. 2005. Smooth minimization of non-smooth functions.

Nesterov's smoothing (3/4)

Theorem. The function f_η is differentiable on $\mathbb{R}^p$. Let $u^\star_\eta(x)$ be the associated maximizer in the definition of f_η. Then, the gradient is given by

$$\nabla f_\eta(x) = A^\mathsf{T} u^\star_\eta(x).$$

Moreover, the gradient is L_η-Lipschitz continuous, with

$$L_\eta = \frac{1}{\eta}\|A\|^2_{2\to 2}.$$

Remark. Differentiability is a consequence of Danskin's theorem.

Yu. Nesterov. 2005. Smooth minimization of non-smooth functions.

Nesterov's smoothing (4/4)

Theorem. Let $\varepsilon > 0$. Run an accelerated gradient method on f_η for a properly chosen η. It takes $O(1/t)$ iterations to find an iterate x_t such that $f(x_t) \le f^\star + \varepsilon$.

Proof. Define $D_{\mathcal{Z}} \coloneqq \max_z \left\{ (1/2)\|z\|_2^2 \;\middle|\; z \in \mathcal{Z} \right\}$ for $\mathcal{Z} \in \{ \mathcal{X}, \mathcal{Y} \}$. We write

$$\begin{aligned} f(x_t) - f^\star &= f(x_t) - f_\eta(x_t) + f_\eta(x_t) - f_\eta^\star + f_\eta^\star - f^\star \\ &\le \frac{\eta}{2} D_{\mathcal{Y}} + O\left(\frac{\|A\|_{2\to 2}^2 D_{\mathcal{X}}}{\eta t^2} \right) \end{aligned}$$

Optimizing over η, we obtain

$$f(x_t) - f^\star = O\left(\frac{\|A\|_{2\to 2}\sqrt{D_{\mathcal{X}} D_{\mathcal{Y}}}}{t} \right).$$

Application: Minimax strategy (1/2)

Alice and Bob are playing a game.

- Alice can choose her action from the set $\mathcal{A} \coloneqq \{ a_1, \dots, a_p \}$.
- Bob can choose his action from the set $\mathcal{B} \coloneqq \{ b_1, \dots, b_q \}$.
- For every $(a, b) \in \mathcal{A} \times \mathcal{B}$, there is a number $\pi(a, b)$, which represents Alice's loss and Bob's pay-off.
- A *strategy* is a randomized action.

Question. How does Alice decide her strategy?

Application: Minimax strategy (2/2)

Alice's minimax strategy is given by

$$x^\star \in \arg\min_x \max_y \left\{ \sum_{i,j} \pi(a_i, b_j) x^{(i)} y^{(j)} \;\middle|\; x \in \Delta_p, y \in \Delta_q \right\},$$

where Δ_p and Δ_q denote the simplexes in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively.

Let $A \in \mathbb{R}^{q \times p}$ such that $A_{i,j} = \pi(a_i, b_j)$. We write, equivalently,

$$x^\star \in \arg\min_x \left\{ \max_y \left\{ \langle Ax, y \rangle \mid y \in \Delta_q \right\} \;\middle|\; x \in \Delta_p \right\}.$$

Application: Minimax strategy

Algorithm An $O(1/\varepsilon)$ algorithm

Choose T and the smoothing parameter $\mu > 0$ properly.
Set $y_1 = x_0 \in \Delta_p$ and $\eta_1 = 1$.
Set $L = \frac{\|A\|_{2\to 2}^2}{\mu}$
for $t = 1, \ldots, T$ **do**
    $u_t \leftarrow \arg\max_z \left\{ \langle Ay_t, z\rangle - \frac{\mu}{2}\|z\|_2^2 \;\middle|\; z \in \Delta_q \right\}$
    $g_t \leftarrow A^\mathsf{T} u_t$ ▷ Compute $\nabla f_\mu(y_t)$.
    $x_t \leftarrow \arg\min_x \left\{ \langle g_t, x - y_t\rangle + \frac{L}{2}\|x - y_t\|_2^2 \;\middle|\; x \in \Delta_p \right\}$
    $\eta_{t+1} \leftarrow \frac{1+\sqrt{1+4\eta_t^2}}{2}$
    $y_{t+1} \leftarrow x_t + \frac{\eta_t - 1}{\eta_{t+1}}(x_t - x_{t-1})$
end for

A. Beck and M. Teboulle. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems.

Convergence of the proximal point algorithm

Proximal point algorithm

Consider the optimization problem:

$$f^\star = \min_x \{ f(x) \mid x \in \mathbb{R}^p \},$$

for some proper closed convex function f.

Proximal point algorithm

$$x_t \leftarrow (I + \eta_t \partial f)^{-1} x_{t-1} = \mathrm{prox}_{\eta_t f}(x_{t-1}).$$

Theorem. Define $\sigma_t \coloneqq \sum_{\tau=1}^t \eta_\tau$. For any $x \in \mathbb{R}^p$, it holds that

$$f(x_t) - f(x) \le \frac{\|x - x_0\|_2^2}{2\sigma_t}.$$

O. Güler. 1991. On the convergence of the proximal point algorithm for convex minimization.

Proof of the convergence guarantee (1/3)

Lemma. Define $y_t := (x_{t-1} - x_t)/\eta_t$. The sequence $(\|y_t\|_2)_{t\in\mathbb{N}}$ is non-increasing.

Proof. Notice that (Why?)

$$y_t = (x_{t-1} - \mathrm{prox}_{\eta_t f}(x_{t-1}))/\eta_t \in \partial f(x_t).$$

Then we obtain (Why?)

$$\langle y_{t+1} - y_t, x_{t+1} - x_t \rangle = \eta_t \langle y_{t+1} - y_t, -y_{t+1} \rangle \geq 0.$$

Therefore, we write

$$\|y_{t+1}\|_2^2 \leq \langle y_t, y_{t+1} \rangle \leq \|y_t\|_2 \|y_{t+1}\|_2.$$

Proof of the convergence guarantee (2/3)

Proof of the theorem. Since $y_t \in \partial f(x_t)$, we write

$$f(x) - f(x_t) \geq \langle y_t, x - x_t \rangle = \eta_t^{-1} \langle x_{t-1} - x_t, x - x_t \rangle .$$

Setting $x = x_{t-1}$, we notice that

$$f(x_{t-1}) - f(x_t) \geq \eta_t^{-1} \|x_{t-1} - x_t\|_2^2 \geq 0,$$

meaning that $(f(x_t))_{t \in \mathbb{N}}$ is a non-increasing sequence.

By the *three-point equality*,

$$\begin{aligned} 2\eta_t(f(x) - f(x_t)) &\geq 2 \langle x_{t-1} - x_t, x - x_t \rangle \\ &= \|x_{t-1} - x_t\|_2^2 + \|x - x_t\|_2^2 - \|x - x_{t-1}\|_2^2 \\ &= \eta_t^2 \|y_t\|_2^2 + \|x - x_t\|_2^2 - \|x - x_{t-1}\|_2^2. \end{aligned}$$

Proof of the convergence guarantee (3/3)

Proof of the theorem continued. Summing over t, we obtain

$$2\left(\sigma_t f(x) - \sum_{\tau=1}^{t} \eta_\tau f(x_\tau)\right) \geq \sum_{\tau=1}^{t} \eta_\tau^2 \|y_\tau\|_2^2 + \|x - x_t\|_2^2 - \|x - x_0\|_2^2$$
$$\geq -\|x - x_0\|_2^2.$$

Recall that $(f(x_t))_{t \in \mathbb{N}}$ is non-increasing. Then we obtain

$$2\sigma_t(f(x) - f(x_t)) \geq -\|x - x_0\|_2^2.$$

Conclusions

Summary (1/2)

Let f, g be a proper closed convex function on $\mathbb{R}^p$.

- The proximal mapping is well-defined on $\mathbb{R}^p$ as

$$\begin{aligned}\operatorname{prox}_f(x) &:= (I + \partial f)^{-1} x \\ &= \arg\min_y \left\{ f(y) + \frac{1}{2}\|y - x\|_2^2 \;\middle|\; y \in \mathbb{R}^p \right\}.\end{aligned}$$

- The proximal gradient method:

$$x_{t+1} \leftarrow \operatorname{prox}_{\eta_t g}(x_t - \eta_t \nabla f(x_t)).$$

Summary (2/2)

Let f be a proper closed convex function on $\mathbb{R}^p$.

- The Moreau envelope is given by

$$f_\eta(x) = \min_y \left\{ f(y) + \frac{1}{2\eta}\|y - x\|_2^2 \;\middle|\; y \in \mathbb{R}^p \right\},$$

  which is convex, differentiable, and $(1/\eta)$-smooth on $\mathbb{R}^p$, with

$$\nabla f_\eta(x) = \frac{1}{\eta}(x - \mathrm{prox}_{\eta f}(x)).$$

- The proximal point method, which can be interpreted as a smoothing approach, is given by

$$x_t \leftarrow \mathrm{prox}_{\eta_{t-1} f}(x_{t-1}).$$

Next lecture

- Approach of the "Russian school".