# STA 250 – Homework 2, due: Sun, May 4

Instructor: Dogyoon Song

**Instructions:** Please feel free to collaborate with other students on your homework, but you must list the names of any collaborators at the top of your homework assignment. All final write-ups must be done individually, and submissions must be made via Gradescope in a single LaTeX-produced PDF file. You don't have to submit your solutions to problems that are marked "Optional," which are rather challenging and will not be graded, but could be possibly interesting for your own learning.

## Problem 1: Convexity

For each of the following functions, please address all of the following:

1. Determine whether or not the function is convex (specify its domain).
2. If it is convex, assess whether it is strongly convex and/or $L$-smooth (include explicit constants if known, or explain why they do not exist).
3. Find the convex conjugate $f^*$, defined by

$$f^*(y) \;=\; \sup_{x \in \operatorname{dom} f} \big( \langle y, x \rangle \;-\; f(x) \big),$$

and describe the effective domain of $f^*$.

**Note:** You may find relevant material e.g., in [BV04, Chapters 2 & 3].

**(a) Indicator Function of a Set.**
Let $S \subseteq \mathbb{R}^d$ (not necessarily convex) and define

$$I_S(x) = \begin{cases} 0, & x \in S, \\ +\infty, & x \notin S. \end{cases}$$

**(b) Norms.**

(i) $f(x) = \|x\|$, where $\|\cdot\|$ is some arbitrary norm on $\mathbb{R}^d$.

(ii) $f(x) = \frac{1}{2}\|x\|^2$, where $\|\cdot\|$ is again an arbitrary norm.

*Hint:* The dual norm $\|u\|_*$ of a norm $\|\cdot\|$ is defined by $\|u\|_* \;\coloneqq\; \sup\{\langle u, x\rangle : \|x\| \le 1\}$.

**(c) Negative Entropy (1D).**
Consider $f : (0, \infty) \to \mathbb{R}$ given by

$$f(x) \;=\; x \log x.$$

(If you wish, you may generalize to the vector case $f(x) = \sum_{i=1}^d x_i \log x_i$ for $x_i > 0$.)

**(d) Maximum of Linear Functions.**
Let $\{w_1, \ldots, w_m\} \subset \mathbb{R}^d$ be fixed. Define

$$g(x) \;=\; \max_{1 \le i \le m} \langle w_i,\; x \rangle.$$

**(e) Log-Sum-Exp.**
For $x \in \mathbb{R}^d$, define

$$f(x) \;=\; \log\left( \sum_{i=1}^{d} e^{x_i} \right).$$

**(f) A Huber-Like Loss.**
Let $\delta > 0$. For $x \in \mathbb{R}$ (one dimension), define

$$\ell_\delta(x) = \begin{cases} \frac{1}{2}\, x^2, & \text{if } |x| \le \delta, \\ \delta\, |x| \;-\; \frac{1}{2}\, \delta^2, & \text{otherwise.} \end{cases}$$

**(g) (Optional) Largest Eigenvalue Function.**
Consider the space $S^d$ of $d \times d$ real symmetric matrices. Define

$$h(X) \;=\; \lambda_{\max}(X).$$

Assume the inner product is $\langle X, Y \rangle = \operatorname{trace}(X^\top Y)$ and the norm is typically the Frobenius or operator norm (please specify which you use for smoothness).

**(h) (Optional) Negative Log-Det.**
For $X \in S^d$ such that $X \succ 0$, define

$$f(X) \;=\; -\log\bigl(\det X\bigr).$$

(You may assume the domain is the positive-definite cone in $S^d \coloneqq \{X \in \mathbb{R}^{d \times d} : X^T = X\}$.)

## Problem 2: Convergence Analysis

**Note:** You may find relevant material in, for example, [B+15, Sections 3 & 6] or [N+18, Chapter 2].

**(a) Gradient Descent with Diminishing Step Sizes.**
Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex and $L$-smooth function (but not necessarily strongly convex). Consider gradient descent where

$$x^{(k+1)} = x^{(k)} - \eta_k\, \nabla f\bigl(x^{(k)}\bigr), \quad \eta_k = \frac{1}{L}\,\frac{1}{k}.$$

Prove that

$$f\bigl(x^{(k)}\bigr) \;-\; f(x^*) \;=\; O\left(\frac{1}{k}\right),$$

where $x^*$ is a minimizer of $f$. Compare this with the convergence result for a fixed step size $\eta = \frac{1}{L}$; is there any difference in the rate or constant?

**(b) Stochastic Gradient Descent (SGD).**
Let $f(x) = \mathbb{E}_\xi\bigl[\ell(x;\, \xi)\bigr]$, where each $\ell(\cdot;\, \xi)$ is convex and $L$-smooth in $x$. Define the SGD iteration:

$$x^{(k+1)} = x^{(k)} - \alpha_k\, \nabla \ell\bigl(x^{(k)};\, \xi_k\bigr),$$

where $\{\xi_k\}$ are i.i.d. samples and $\alpha_k > 0$ is the step size.

(i) State a standard convergence theorem for SGD in the convex (but not strongly convex) setting, such as

$$\mathbb{E}\big[f(\bar{x}^{(k)}) - f(x^*)\big] \;\le\; O\left(\frac{1}{\sqrt{k}}\right),$$

where $\bar{x}^{(k)} = \frac{1}{k}\sum_{i=1}^{k} x^{(i)}$.

(ii) Outline the key steps in the proof with sufficient details.

(iii) (Optional) Complete full details of the proof outlined above.

(iv) If $f$ is *strongly convex*, explain how the rate can improve to $O(\frac{1}{k})$.

**(c) Nesterov's Accelerated Gradient (NAG).**

Finally, consider Nesterov's accelerated gradient method for minimizing a convex and $L$-smooth function $f : \mathbb{R}^n \to \mathbb{R}$. Recall the update:

$$y^{(k)} = x^{(k)} + \beta_k\big(x^{(k)} - x^{(k-1)}\big),$$
$$x^{(k+1)} = y^{(k)} - \eta\,\nabla f\big(y^{(k)}\big),$$

where $\beta_k$ and $\eta$ are chosen carefully (e.g. $\eta = \frac{1}{L}$ and $\beta_k = \frac{k-1}{k+2}$).

(i) Show that if $f$ is convex and $L$-smooth, then

$$f\big(x^{(k)}\big) - f\big(x^*\big) \;=\; O\left(\frac{1}{k^2}\right).$$

You may use a *potential function* argument (or partial version thereof) as in [N+18, Chapter 2] or [B+15, Section 3], or refer to other known proofs (e.g., Yudong Chen's lecture notes).

(ii) Compare this $O(\frac{1}{k^2})$ rate to the $O(\frac{1}{k})$ rate of standard gradient descent. Why is the *acceleration* technique so valuable in practice?

## Problem 3: Duality, KKT Conditions, and Kernel Methods

**(a) Quadratic Problem with Linear Constraints.**

Consider the following constrained optimization problem:

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}\|x\|_2^2 \quad \text{subject to} \quad A\,x = b,$$

where $A \in \mathbb{R}^{m\times d}$ and $b \in \mathbb{R}^m$.

(i) Write down the Lagrangian $\mathcal{L}(x, \lambda)$.

(ii) Derive the KKT conditions and solve for $(x^*, \lambda^*)$.

(iii) Show that $x^*$ is the orthogonal projection of 0 onto the affine subspace $\{\,x : A\,x = b\}$.

**(b) $\ell_1$-Constrained Logistic Regression.**

Suppose we have training data $(x_i, y_i)$ with $y_i \in \{-1, +1\}$. Consider the primal problem

$$\min_{w\in\mathbb{R}^d} \sum_{i=1}^{n} \log\Big(1 + \exp(-\,y_i\,\langle w,\, x_i\rangle)\Big) \quad \text{subject to} \quad \|w\|_1 \;\le\; C,$$

for some constant $C > 0$. (Hence, we treat the $\ell_1$ norm as a *hard* constraint.)

(i) Form the Lagrangian by introducing a dual variable (multiplier) $\alpha \ge 0$ for the constraint $\|w\|_1 \le C$.

(ii) State the KKT conditions for $(w, \alpha)$, and discuss how one might solve (or partially characterize) $w^*$ and $\alpha^*$.

(iii) (Optional) Briefly comment on how an $\ell_1$-constraint often induces sparsity in $w^*$ and why a dual viewpoint (or Fenchel conjugate approach) can be helpful in developing coordinate-descent or proximal methods.

**(c) Kernel Methods and the Dual.**

In many machine-learning tasks (e.g., SVM classification), one considers an *infinite-dimensional* feature mapping $\phi(\cdot)$ into a Reproducing Kernel Hilbert Space (RKHS). A generic primal form might look like:

$$\min_{w \in \mathcal{H},\, b \in \mathbb{R}} \quad \frac{1}{2}\,\|w\|_{\mathcal{H}}^2 \quad \text{subject to} \quad y_i \left\langle w,\, \phi(x_i) \right\rangle + b \;\geq\; 1, \quad i = 1, \ldots, n,$$

where $\mathcal{H}$ is a (potentially infinite-dimensional) Hilbert space, and $x_i$, $y_i$ are training data with $y_i \in \{-1, +1\}$.

(i) Write the Lagrangian for this primal problem and identify the corresponding dual variables.

(ii) Show that, in the dual problem, the dimension effectively becomes $n$, corresponding to the $n$ constraints, rather than the infinite dimensionality of $\mathcal{H}$.

(iii) Explain how the kernel function $k(x, x') = \langle \phi(x),\, \phi(x') \rangle$ enters the dual, thereby enabling one to work only with an $n \times n$ kernel matrix.

(iv) In a few sentences, discuss why this "kernel trick" is so crucial for high-dimensional or infinite-dimensional feature spaces in machine learning.

## Problem 4: Mirror Descent, Bregman Divergences, and Proximal Maps

**Setup:** Let $\psi : \mathcal{X} \to \mathbb{R}$ be a differentiable, strictly convex function on a convex domain $\mathcal{X} \subseteq \mathbb{R}^d$. We refer to $\psi$ as the *mirror map* in mirror descent. The associated *Bregman divergence* is defined by

$$D_\psi(x, y) \;=\; \psi(x) \;-\; \psi(y) \;-\; \langle \nabla\psi(y),\, x - y \rangle.$$

**(a) Properties of Bregman Divergence.**

(i) Prove that $D_\psi(x, y) \geq 0$ for all $x, y \in \mathcal{X}$, with equality if and only if $x = y$.

(ii) Show that if $\psi(x) = \frac{1}{2}\,\|x\|_2^2$, then $D_\psi(x, y) = \frac{1}{2}\,\|x - y\|_2^2$. Compare $D_\psi(\cdot, \cdot)$ to an $\ell_2$ or squared $\ell_2$ norm in terms of geometry.

**(b) Mirror Descent Update Rule.**

Consider minimizing a differentiable function $f(x)$ over $\mathcal{X}$. The *mirror descent* update is:

$$x^{(k+1)} = \arg\min_{z \in \mathcal{X}} \left\{ \langle \nabla f\big(x^{(k)}\big),\, z \rangle + \frac{1}{\eta_k}\, D_\psi\big(z,\, x^{(k)}\big) \right\}.$$

(i) Derive this rule by linearizing $f$ around $x^{(k)}$ and adding a Bregman "penalty" term $D_\psi(\cdot, x^{(k)})$.

(ii) Show that if $\psi(x) = \frac{1}{2}\,\|x\|_2^2$, then the update reduces to standard gradient descent,

$$x^{(k+1)} = x^{(k)} \;-\; \eta_k\, \nabla f\big(x^{(k)}\big).$$

**(c) Constrained Example (Probability Simplex).**

Let $\mathcal{X} = \Delta^d = \{\, x \in \mathbb{R}^d : x_i \geq 0,\ \sum_{i=1}^d x_i = 1 \}$ be the probability simplex, and define $\psi(x) = \sum_{i=1}^d x_i \log x_i$ (the negative-entropy mirror map).

(i) Show how the mirror descent step yields the *multiplicative (exponential) weights* update:

$$x_i^{(k+1)} \;\propto\; x_i^{(k)} \exp\Big(-\eta_k \left[\nabla f\big(x^{(k)}\big)\right]_i\Big).$$

(ii) Provide a short example (e.g., a small online learning or 2D scenario) to illustrate how $x^{(k)}$ evolves within the simplex.

**(d) Additional Examples.**
Choose *two* of the following (or propose a similar example):

- **Box Constraint**: $\mathcal{X} = [0,1]^d$ with a *log-barrier* mirror map $\psi(x) = -\sum_{i=1}^d \log\big(x_i(1-x_i)\big)$ (assuming $0 < x_i < 1$).
- **$p$-norm Ball**: $\mathcal{X} = \{\, x : \|x\|_p \le 1\}$ for $p > 1$, with $\psi(x) = \frac{1}{p}\,\|x\|_p^p$ (if differentiable at 0).
- **Mahalanobis Geometry**: $\mathcal{X} = \mathbb{R}^d$, $\psi(x) = \frac{1}{2}\,x^\top Q x$ where $Q \succ 0$ (a positive-definite matrix). This is often used in preconditioning or second-order methods.
- **Exponential Family / Negative Log-likelihood**: In a statistical setting, $\psi$ might be related to a log-likelihood or cumulant generating function. For instance, if $\mathcal{X}$ is unconstrained, one can set $\psi(x)$ to be the negative log-likelihood for a particular exponential family, yielding a geometry that matches the Fisher information.

(i) Show that $\psi$ is strictly convex and differentiable on $\mathrm{int}(\mathcal{X})$.

(ii) Outline the mirror-descent update for a function $f(x)$ with step size $\eta_k$.

(iii) (Optional) Explain how the choice of mirror map helps to maintain constraints implicitly (e.g., a log barrier ensures $x_i \in (0,1)$).

**(e) Basic Convergence Bound (Outline).**
Here we examine how mirror descent can achieve $O\big(\frac{1}{\sqrt{k}}\big)$ convergence for general convex $f$, or $O\left(\frac{1}{k}\right)$ when $f$ is also $L$-smooth.

(i) State a typical mirror descent convergence theorem, for instance,

$$f\left(\frac{1}{k}\sum_{j=1}^{k} x^{(j)}\right) - f\big(x^*\big) \;\le\; O\left(\frac{1}{\sqrt{k}}\right), \quad (\text{or } O\left(\frac{1}{k}\right) \text{ under smoothness}).$$

(ii) Outline the main proof ingredients: (1) Expand using $D_\psi$ and local linearization, (2) Use a telescoping sum in terms of $D_\psi$, (3) Bound the total "Bregman diameter" of $\mathcal{X}$, etc.

(iii) (Optional) Complete full details of the proof outlined above.

**(f) Proximal Map Interpretation.**
In the Euclidean setting, the *proximal map* of a convex function $g$ is

$$\mathrm{prox}_g(v) = \arg\min_{x\in\mathbb{R}^d}\Big\{g(x) + \frac{1}{2}\,\|x - v\|_2^2\Big\}.$$

(i) Describe how the mirror-descent update can be viewed as a generalized "prox" step, where the squared $\ell_2$ norm is replaced by $D_\psi(\cdot,\cdot)$ and $f$ is linearized rather than added in full.

(ii) If $f(x) = 0$ (pure feasibility), show that the mirror-descent step is the *Bregman projection* of a point $v$ onto $\mathcal{X}$ under $\psi$.

(iii) (Optional) Give a short example of a Bregman projection in negative-entropy geometry and interpret the result.

# References

[B+15] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[BV04] Stephen P Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[N+18] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.