TTIC 31070: Convex Optimization
Homework 3

Professor Zhiyuan Li

Spring 2026

Problem 1 (Positive-orthant entropy restricted to the simplex)**.** *In Lecture 8, multiplicative weights was presented as the mirror update on the simplex geometry. Here we derive the same update from the ambient entropy* $\Phi(x) := \sum_i x_i \log x_i$ *on* $\mathbb{R}^d_{++}$*, restricted to* $\Delta_d := \{x \in \mathbb{R}^d_+ : \sum_i x_i = 1\}$*. Let* $\Delta_d^\circ := \Delta_d \cap \mathbb{R}^d_{++}$*, and use the continuous extension of* $D_\Phi(\cdot, y)$ *to* Δ_d*, with* $0 \log 0 := 0$*. For* $y \in \Delta_d^\circ$*,* $g \in \mathbb{R}^d$*, and* $\eta > 0$*, consider*

$$x^+ \in \arg\min_{x \in \Delta_d} \{\eta \langle g, x \rangle + D_\Phi(x, y)\}.$$

The Lagrange multiplier below is the normal direction to the simplex.

(a) *Compute the ambient Bregman divergence:*

$$D_\Phi(x, y) = \sum_{i=1}^d x_i \log \frac{x_i}{y_i} - \sum_{i=1}^d x_i + \sum_{i=1}^d y_i.$$

Deduce that $D_\Phi(x, y) = \sum_i x_i \log(x_i/y_i)$ *for* $x, y \in \Delta_d^\circ$*.*

(b) *First show that every minimizer belongs to* Δ_d°*. Then write the KKT condition for the constrained minimization over* Δ_d*, and show that there exists* $\lambda \in \mathbb{R}$ *such that* $\eta g_i + \log x_i^+ - \log y_i + \lambda = 0$ *for all* i*.*

(c) *Deduce the normalized exponential update*

$$x_i^+ = \frac{y_i e^{-\eta g_i}}{\sum_{j=1}^d y_j e^{-\eta g_j}}, \qquad i = 1, \ldots, d.$$

(d) *Show that adding a constant vector to the covector does not change the update: for every* $c \in \mathbb{R}$*, replacing* g *by* $g + c\mathbf{1}$ *gives the same* x^+*. Explain this using the tangent space* $\{v \in \mathbb{R}^d : \sum_i v_i = 0\}$*.*

(e) *If* $y = (1/d, \ldots, 1/d)$*, prove* $D_\Phi(u, y) \le \log d$ *for every* $u \in \Delta_d$*.*

(f) *Now run the entropy mirror-descent update from parts* (a)–(c) *for covectors* $g_1, \ldots, g_T$*, starting from* $x_1 = (1/d, \ldots, 1/d)$*, and assume* $\|g_t\|_\infty \le G$ *for all* t*. You may use the fact that negative entropy is* 1*-strongly convex with respect to* $\|\cdot\|_1$ *on* Δ_d*. Prove that, for every* $u \in \Delta_d$*,*

$$\sum_{t=1}^T \langle g_t, x_t - u \rangle \le \frac{\log d}{\eta} + \frac{\eta G^2 T}{2}.$$

Choose η *as a function of* G, T, d *and deduce the regret bound* $\sum_{t=1}^T \langle g_t, x_t - u \rangle \le G\sqrt{2T \log d}$*.*

(g) *Compare this with Euclidean mirror descent on the same simplex, using* $\Psi(x) := \frac{1}{2}\|x\|_2^2$ *and* $x_1 = (1/d, \dots, 1/d)$. *Using the same assumption* $\|g_t\|_\infty \le G$, *prove that Euclidean mirror descent gives*

$$\sum_{t=1}^{T} \langle g_t, x_t - u\rangle \le \frac{1}{2\eta} + \frac{\eta d G^2 T}{2} \qquad \forall u \in \Delta_d.$$

Optimize η *and compare the resulting* $G\sqrt{dT}$ *dependence with the entropy bound* $G\sqrt{T\log d}$.

Problem 2 (RMSProp as diagonal AdaGrad for discounted regret)**.** *RMSProp was popularized through Geoffrey Hinton's neural-network course notes and is more common than AdaGrad in modern deep-learning practice. It is also the second-moment core of Adam, before adding first-moment momentum and bias correction. This problem studies that momentum-free core.*

Let $X \subseteq \mathbb{R}^d$ *be nonempty, closed, and convex, and assume*

$$R_\infty := \sup_{x,y\in X} \|x - y\|_\infty < \infty.$$

Fix $\eta > 0$, $\varepsilon > 0$, *and* $\beta_2 \in (0,1)$. *Let* $\widetilde{L}_1, \dots, \widetilde{L}_T : X \to \mathbb{R}$ *be convex losses. Starting from* $\widetilde{x}_1 \in X$, *define the RMSProp-style second-moment update*

$$\widetilde{g}_t \in \partial \widetilde{L}_t(\widetilde{x}_t), \qquad v_{0,i} := \varepsilon, \qquad v_{t,i} := \beta_2 v_{t-1,i} + \widetilde{g}_{t,i}^2,$$

$$\widetilde{H}_t := \frac{1}{\eta}\operatorname{diag}(\sqrt{v_{t,1}}, \dots, \sqrt{v_{t,d}}), \qquad \widetilde{x}_{t+1} \in \arg\min_{x\in X}\left\{\langle \widetilde{g}_t, x\rangle + \frac{1}{2}\|x - \widetilde{x}_t\|_{\widetilde{H}_t}^2\right\}.$$

Here $\|z\|_H^2 := z^\top H z$ *for* $H \succ 0$. *The point of the problem is that this discounted RMSProp trajectory is exactly a diagonal AdaGrad trajectory after a time-dependent rescaling of the losses.*

(a) *Show that the update is the* H_t*-metric projection*

$$\widetilde{x}_{t+1} = \arg\min_{x\in X} \frac{1}{2}\left\|x - (\widetilde{x}_t - \widetilde{H}_t^{-1}\widetilde{g}_t)\right\|_{\widetilde{H}_t}^2.$$

In particular, if $X = \mathbb{R}^d$, *prove the coordinate formula*

$$\widetilde{x}_{t+1,i} = \widetilde{x}_{t,i} - \eta\,\frac{\widetilde{g}_{t,i}}{\sqrt{v_{t,i}}}, \qquad i = 1, \dots, d.$$

This is the RMSProp update without first-moment momentum or bias correction.

(b) *Define a second trajectory as follows. Set* $x_1 := \widetilde{x}_1$. *Whenever the induction has shown* $x_t = \widetilde{x}_t$, *define rescaled losses and gradients*

$$L_t(x) := \beta_2^{-t/2}\widetilde{L}_t(x), \qquad g_t := \beta_2^{-t/2}\widetilde{g}_t \in \partial L_t(\widetilde{x}_t),$$

and diagonal AdaGrad accumulators

$$a_{0,i} := \varepsilon, \qquad a_{t,i} := a_{t-1,i} + g_{t,i}^2, \qquad A_t := \frac{1}{\eta}\operatorname{diag}(\sqrt{a_{t,1}}, \dots, \sqrt{a_{t,d}}).$$

Define the diagonal AdaGrad update by

$$x_{t+1} \in \arg\min_{x\in X}\left\{\langle g_t, x\rangle + \frac{1}{2}\|x - x_t\|_{A_t}^2\right\}.$$

Prove by induction that

$$v_{t,i} = \beta_2^t a_{t,i}, \qquad \widetilde{H}_t = \beta_2^{t/2} A_t, \qquad x_t = \widetilde{x}_t \quad \text{for every } t.$$

Hint: after using $g_t = \beta_2^{-t/2}\widetilde{g}_t$, show that the RMSProp and diagonal AdaGrad subproblems differ by multiplication by the positive scalar $\beta_2^{t/2}$, so their argmin sets are the same.

(c) *Prove the following diagonal AdaGrad regret bound for the rescaled losses: for every $u \in X$,*

$$\sum_{t=1}^{T} \big(L_t(x_t) - L_t(u)\big) \le \left(\frac{R_\infty^2}{2\eta} + \eta\right) \sum_{i=1}^{d} \sqrt{a_{T,i}}.$$

Hint: first prove the one-step inequality

$$\langle g_t, x_t - u\rangle \le \frac{1}{2}\,\|u - x_t\|_{A_t}^2 - \frac{1}{2}\,\|u - x_{t+1}\|_{A_t}^2 + \frac{1}{2}\,\|g_t\|_{A_t^{-1}}^2,$$

then use monotonicity of A_t and the scalar inequality

$$\sum_{t=1}^{T} \frac{g_{t,i}^2}{\sqrt{a_{t,i}}} \le 2\sqrt{a_{T,i}}.$$

(d) *Convert part* (c) *back to the original losses and prove the discounted regret bound*

$$\sum_{t=1}^{T} \beta_2^{(T-t)/2} \big(\widetilde{L}_t(\widetilde{x}_t) - \widetilde{L}_t(u)\big) \le \left(\frac{R_\infty^2}{2\eta} + \eta\right) \sum_{i=1}^{d} \sqrt{\beta_2^T \varepsilon + \sum_{t=1}^{T} \beta_2^{T-t} \widetilde{g}_{t,i}^2}.$$

Problem 3 (Frank–Wolfe as mirror descent on the Fenchel dual)**.** *This problem follows the Bach–Peña viewpoint that Frank–Wolfe in the primal space can be read as mirror descent on a Fenchel dual problem.*

Let $(E, \|\cdot\|)$ be a finite-dimensional normed space, with dual norm $\|\cdot\|_*$. Let $K \subset E$ be nonempty, compact, and convex, and let $f : E \to \mathbb{R}$ be closed, convex, differentiable, and L-smooth with respect to $\|\cdot\|$. Define

$$\delta_K(x) := \begin{cases} 0, & x \in K, \\ +\infty, & x \notin K. \end{cases}$$

The primal problem is

$$p^\star := \inf_{x \in E} \{f(x) + \delta_K(x)\} = \min_{x \in K} f(x).$$

Recall from Lecture 4 that the Fenchel dual of

$$\inf_{x \in E} \{f(x) + h(x)\}$$

is

$$\sup_{u \in E^*} \{-f^*(u) - h^*(-u)\}.$$

Thus, for the constrained problem above, the dual maximization problem is

$$d^\star = \sup_{u \in E^*} \{-f^*(u) - \sigma_K(-u)\},$$

or equivalently the dual minimization problem is

$$\inf_{u \in E^*} \Psi(u), \qquad \Psi(u) := f^*(u) + \sigma_K(-u).$$

Part (a) *asks you to justify this formula by computing δ_K^*.*

(a) *Show that* $\delta_K^* = \sigma_K$, *where*

$$\sigma_K(u) := \max_{z \in K} \langle u, z \rangle .$$

Derive the displayed Fenchel dual problem from the Lecture 4 formula. For a primal point $x \in E$ *and a dual point* $u \in E^*$, *define the Fenchel primal-dual gap*

$$\mathcal{G}(x,u) := \big(f(x) + \delta_K(x)\big) - \big(-f^*(u) - \sigma_K(-u)\big).$$

Equivalently,

$$\mathcal{G}(x,u) = f(x) + \delta_K(x) + f^*(u) + \sigma_K(-u).$$

Use Fenchel–Young to show $\mathcal{G}(x,u) \geq 0$.

(b) *For* $x \in K$, *set* $u = \nabla f(x)$, *and choose a Frank–Wolfe atom*

$$s \in \arg\min_{z \in K} \langle u, z \rangle .$$

Prove that

$$\partial \sigma_K(-u) = \arg\min_{z \in K} \langle u, z \rangle ,$$

and then prove

$$\mathcal{G}(x, \nabla f(x)) = \langle \nabla f(x), x - s \rangle .$$

Thus the Frank–Wolfe gap is exactly the Fenchel primal-dual gap evaluated at the dual point $u = \nabla f(x)$.

Comment. *Part* (b) *says that the Frank–Wolfe step can equivalently be written as*

$$u_t = \nabla f(x_t), \qquad s_t \in \partial \sigma_K(-u_t), \qquad x_{t+1} = (1 - \gamma_t) x_t + \gamma_t s_t.$$

This is only a change of notation for the same Frank–Wolfe trajectory once the same initial point, stepsizes, and set-valued atom choices are fixed.

(c) *Now look at the same dynamics from the dual space, namely as mirror descent on the dual minimization objective* Ψ. *Assume in this part additionally that* f *is* μ*-strongly convex for some* $\mu > 0$. *You may use the standard conjugacy facts that* f^* *is differentiable, that* $\nabla f^*(\nabla f(x_t)) = x_t$ *at the relevant dual iterates, and that* f^* *is* $1/L$*-strongly convex with respect to* $\|\cdot\|_*$. *Since*

$$\Psi - f^* = \sigma_K(-u)$$

is convex, Ψ *is* 1*-strongly convex relative to the mirror map* f^*.

Given a Frank–Wolfe trajectory, set $u_t := \nabla f(x_t)$, *and choose*

$$s_t \in \arg\min_{z \in K} \langle u_t, z \rangle .$$

Prove that

$$x_t - s_t \in \partial \Psi(u_t),$$

where the subgradient is taken for the dual-space objective $\Psi : E^* \to \mathbb{R}$. *Then prove that the Frank–Wolfe update*

$$x_{t+1} = (1 - \gamma_t) x_t + \gamma_t s_t$$

is equivalent to the dual-space mirror-descent update

$$u_{t+1} \in \arg\min_{u \in E^*} \{\gamma_t \langle u - u_t, x_t - s_t \rangle + D_{f^*}(u, u_t)\},$$

where

$$D_{f^*}(u, v) := f^*(u) - f^*(v) - \langle u - v, \nabla f^*(v) \rangle .$$

More explicitly, show that the first-order condition for the displayed mirror-descent subproblem is

$$\nabla f^*(u_{t+1}) = \nabla f^*(u_t) - \gamma_t(x_t - s_t) = (1 - \gamma_t)x_t + \gamma_t s_t = x_{t+1}.$$

Thus, along the coupled trajectory,

$$u_{t+1} = \nabla f(x_{t+1}).$$

(d) *Now take the standard Frank–Wolfe stepsizes*

$$\gamma_t = \frac{2}{t+2}, \qquad t = 0, 1, \ldots .$$

Observe that

$$\gamma_t = \frac{t+1}{1 + 2 + \cdots + (t+1)}.$$

Thus, after the reindexing $k = t+1$, this is exactly the strongly-convex mirror-descent stepsize pattern

$$\eta_k = \frac{\lambda_k}{\Lambda_k}, \qquad \lambda_k = k, \qquad \Lambda_k = \sum_{r=1}^{k} r.$$

Assume $x_0 \in K$ and

$$x_{t+1} = (1 - \gamma_t)x_t + \gamma_t s_t.$$

Prove by induction that, for every $T \geq 1$,

$$x_T = \sum_{t=0}^{T-1} \frac{2(t+1)}{T(T+1)}\, s_t.$$

In particular, Frank–Wolfe with $\gamma_t = 2/(t+2)$ maintains the same linearly weighted average pattern as the weighted averages used for strongly convex nonsmooth mirror descent.