TTIC 31070: Convex Optimization
Homework 3 Solutions

Professor Zhiyuan Li

Spring 2026

Problem 1

(a) Since $\nabla\Phi(y)_i = 1 + \log y_i$,

$$D_\Phi(x, y) = \sum_i x_i \log x_i - \sum_i y_i \log y_i - \sum_i (1 + \log y_i)(x_i - y_i).$$

Rearranging,

$$D_\Phi(x, y) = \sum_i x_i \log \frac{x_i}{y_i} - \sum_i x_i + \sum_i y_i.$$

If $x, y \in \Delta_d^\circ$, then $\sum_i x_i = \sum_i y_i = 1$, so

$$D_\Phi(x, y) = \sum_i x_i \log \frac{x_i}{y_i}.$$

(b) Let

$$F(x) := \eta \langle g, x \rangle + D_\Phi(x, y).$$

We first show that a minimizer cannot lie on the boundary of Δ_d. Suppose $x_i = 0$ for some i, and choose k with $x_k > 0$. For small $t > 0$, the point $x + t(e_i - e_k)$ remains in Δ_d. The directional derivative of the linear term is finite, while the entropy part contains

$$t \log \frac{t}{y_i} + (x_k - t) \log \frac{x_k - t}{y_k}.$$

Its derivative as $t \downarrow 0$ is

$$\log \frac{t}{y_i} - \log \frac{x_k - t}{y_k} \to -\infty.$$

Thus moving into the simplex strictly decreases F, so no boundary point is optimal. Hence every minimizer lies in Δ_d°.

On Δ_d°, the only active constraint is $\sum_i x_i = 1$. The Lagrangian is

$$\eta \langle g, x \rangle + D_\Phi(x, y) + \lambda \left(\sum_i x_i - 1 \right).$$

Since $\partial D_\Phi(x, y)/\partial x_i = \log(x_i/y_i)$, stationarity at x^+ gives

$$\eta g_i + \log \frac{x_i^+}{y_i} + \lambda = 0,$$

equivalently

$$\eta g_i + \log x_i^+ - \log y_i + \lambda = 0 \qquad i = 1, \ldots, d.$$

(c) From the KKT condition,

$$x_i^+ = e^{-\lambda} y_i e^{-\eta g_i}.$$

The scalar $e^{-\lambda}$ is determined by $\sum_i x_i^+ = 1$. Therefore

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

(d) Replacing g by $g + c\mathbf{1}$ multiplies every numerator $y_i e^{-\eta g_i}$ by the same factor $e^{-\eta c}$, so the normalized update is unchanged. Geometrically, feasible directions v in the simplex satisfy $\sum_i v_i = 0$, hence

$$\langle c\mathbf{1}, v \rangle = c \sum_i v_i = 0.$$

Thus the component of the covector normal to the simplex has no effect on the constrained mirror step.

(e) If $y = (1/d, \ldots, 1/d)$, then for $u \in \Delta_d$,

$$D_\Phi(u, y) = \sum_i u_i \log u_i + \log d.$$

Since $0 \le u_i \le 1$, we have $u_i \log u_i \le 0$, with the convention $0 \log 0 = 0$. Hence

$$D_\Phi(u, y) \le \log d.$$

(f) Apply the constant-stepsize version of Corollary 9.4 from the lecture notes to the negative-entropy mirror map on the centered simplex. This is legitimate here because the iterates stay in Δ_d°, while $D_\Phi(u, x_t)$ is defined for $u \in \Delta_d$ by the continuous extension in the first argument. Since Φ is 1-strongly convex with respect to $\|\cdot\|_1$ on Δ_d, and the dual norm is $\|\cdot\|_\infty$, Corollary 9.4 gives

$$\sum_{t=1}^T \langle g_t, x_t - u \rangle \le \frac{D_\Phi(u, x_1)}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \|g_t\|_\infty^2 .$$

Using $x_1 = (1/d, \ldots, 1/d)$, part (e), and $\|g_t\|_\infty \le G$, this gives

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

The right side is minimized at

$$\eta = \sqrt{\frac{2 \log d}{G^2 T}}$$

when $G > 0$. This gives

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

If $G = 0$, the regret is 0, so the same bound is trivial.

(g) Euclidean mirror descent uses $\Psi(x) = \frac{1}{2}\|x\|_2^2$, which is 1-strongly convex with respect to $\|\cdot\|_2$. Therefore

$$\sum_{t=1}^{T} \langle g_t, x_t - u\rangle \le \frac{\frac{1}{2}\|u - x_1\|_2^2}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \|g_t\|_2^2 .$$

For $u \in \Delta_d$ and $x_1 = (1/d, \dots, 1/d)$,

$$\|u - x_1\|_2^2 = \sum_i u_i^2 - \frac{1}{d} \le 1,$$

and $\|g_t\|_2^2 \le d\,\|g_t\|_\infty^2 \le dG^2$. Hence

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

The optimal choice is $\eta = 1/(G\sqrt{dT})$ when $G > 0$, yielding

$$\sum_{t=1}^{T} \langle g_t, x_t - u\rangle \le G\sqrt{dT}.$$

Thus entropy geometry gives dimension dependence $\sqrt{\log d}$, while Euclidean geometry gives $\sqrt{d}$ under the same ℓ_∞ gradient bound.

Problem 2

(a) Complete the square:

$$\langle \widetilde{g}_t, x\rangle + \frac{1}{2}\|x - \widetilde{x}_t\|_{\widetilde{H}_t}^2 = \frac{1}{2}\left\|x - (\widetilde{x}_t - \widetilde{H}_t^{-1}\widetilde{g}_t)\right\|_{\widetilde{H}_t}^2 + \text{constant}.$$

Thus

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

If $X = \mathbb{R}^d$, the minimizer is

$$\widetilde{x}_{t+1} = \widetilde{x}_t - \widetilde{H}_t^{-1}\widetilde{g}_t.$$

Since

$$\widetilde{H}_t^{-1} = \eta\,\mathrm{diag}\left(\frac{1}{\sqrt{v_{t,1}}}, \dots, \frac{1}{\sqrt{v_{t,d}}}\right),$$

we get

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

(b) We prove the three identities by induction. At $t = 0$, $v_{0,i} = a_{0,i} = \varepsilon$. If $v_{t-1,i} = \beta_2^{t-1} a_{t-1,i}$, then

$$v_{t,i} = \beta_2 v_{t-1,i} + \widetilde{g}_{t,i}^2 = \beta_2^t a_{t-1,i} + \widetilde{g}_{t,i}^2.$$

Since $g_{t,i} = \beta_2^{-t/2}\widetilde{g}_{t,i}$,

$$\beta_2^t a_{t,i} = \beta_2^t(a_{t-1,i} + g_{t,i}^2) = \beta_2^t a_{t-1,i} + \widetilde{g}_{t,i}^2.$$

Thus $v_{t,i} = \beta_2^t a_{t,i}$, and therefore

$$\widetilde{H}_t = \frac{1}{\eta}\operatorname{diag}(\sqrt{v_{t,i}}) = \beta_2^{t/2} A_t.$$

Assume now that $x_t = \widetilde{x}_t$. The diagonal AdaGrad objective is

$$\langle g_t, x\rangle + \frac{1}{2}\|x - x_t\|_{A_t}^2 = \beta_2^{-t/2}\left(\langle \widetilde{g}_t, x\rangle + \frac{1}{2}\|x - \widetilde{x}_t\|_{\widetilde{H}_t}^2\right).$$

Multiplication by the positive scalar $\beta_2^{-t/2}$ does not change the argmin set, so $x_{t+1} = \widetilde{x}_{t+1}$. Since $x_1 = \widetilde{x}_1$, the induction is complete.

(c) By convexity,

$$L_t(x_t) - L_t(u) \le \langle g_t, x_t - u\rangle.$$

It remains to bound the right side. The optimality condition for the AdaGrad step gives

$$\langle g_t + A_t(x_{t+1} - x_t), u - x_{t+1}\rangle \ge 0.$$

Hence

$$\langle g_t, x_t - u\rangle = \langle g_t, x_t - x_{t+1}\rangle + \langle g_t, x_{t+1} - u\rangle \le \langle g_t, x_t - x_{t+1}\rangle + \langle A_t(x_{t+1} - x_t), u - x_{t+1}\rangle.$$

By the Lecture 8 three-point identity, applied to $h_t(v) := \frac{1}{2}\|v\|_{A_t}^2$ with $x = x_{t+1}$, $y = x_t$, and $z = u$, we have $\nabla h_t(v) = A_t v$ and $D_{h_t}(a,b) = \frac{1}{2}\|a-b\|_{A_t}^2$, so

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

and Young's inequality in the A_t-norm gives

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

Combining these,

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

Now sum over t. Since A_t is coordinatewise nondecreasing,

$$\begin{aligned}
\sum_{t=1}^{T}\left(\frac{1}{2}\|u - x_t\|_{A_t}^2 - \frac{1}{2}\|u - x_{t+1}\|_{A_t}^2\right) &= \frac{1}{2}\|u - x_1\|_{A_1}^2 + \frac{1}{2}\sum_{t=2}^{T}\|u - x_t\|_{A_t - A_{t-1}}^2 - \frac{1}{2}\|u - x_{T+1}\|_{A_T}^2 \\
&\le \frac{1}{2}\|u - x_1\|_{A_1}^2 + \frac{1}{2}\sum_{t=2}^{T}\|u - x_t\|_{A_t - A_{t-1}}^2 \\
&\le \frac{R_\infty^2}{2}\left(\sum_{i=1}^{d}(A_1)_{ii} + \sum_{t=2}^{T}\sum_{i=1}^{d}\big((A_t)_{ii} - (A_{t-1})_{ii}\big)\right) \\
&= \frac{R_\infty^2}{2}\sum_{i=1}^{d}(A_T)_{ii} = \frac{R_\infty^2}{2\eta}\sum_{i=1}^{d}\sqrt{a_{T,i}}.
\end{aligned}$$

Also,

$$\frac{1}{2}\sum_{t=1}^{T}\|g_t\|^2_{A_t^{-1}} = \frac{\eta}{2}\sum_{i=1}^{d}\sum_{t=1}^{T}\frac{g_{t,i}^2}{\sqrt{a_{t,i}}} \le \eta\sum_{i=1}^{d}\sqrt{a_{T,i}},$$

where the last step uses

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

Indeed, since $a_{t,i} = a_{t-1,i} + g_{t,i}^2$,

$$\frac{g_{t,i}^2}{\sqrt{a_{t,i}}} = \frac{a_{t,i} - a_{t-1,i}}{\sqrt{a_{t,i}}} \le \frac{2(a_{t,i} - a_{t-1,i})}{\sqrt{a_{t,i}} + \sqrt{a_{t-1,i}}} = 2\big(\sqrt{a_{t,i}} - \sqrt{a_{t-1,i}}\big).$$

Summing over t gives

$$\sum_{t=1}^{T}\frac{g_{t,i}^2}{\sqrt{a_{t,i}}} \le 2\sum_{t=1}^{T}\big(\sqrt{a_{t,i}} - \sqrt{a_{t-1,i}}\big) = 2\big(\sqrt{a_{T,i}} - \sqrt{a_{0,i}}\big) \le 2\sqrt{a_{T,i}}.$$

Combining these estimates gives

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

(d) Since $L_t = \beta_2^{-t/2}\widetilde{L}_t$ and $x_t = \widetilde{x}_t$, part (c) gives

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

Multiplying by $\beta_2^{T/2}$,

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

Finally,

$$\beta_2^T a_{T,i} = \beta_2^T\varepsilon + \sum_{t=1}^{T}\beta_2^T g_{t,i}^2 = \beta_2^T\varepsilon + \sum_{t=1}^{T}\beta_2^{T-t}\widetilde{g}_{t,i}^2,$$

because $g_{t,i} = \beta_2^{-t/2}\widetilde{g}_{t,i}$. This is exactly the claimed discounted regret bound.

Problem 3

(a) By definition of the convex conjugate,

$$\delta_K^*(u) = \sup_{x\in E}\{\langle u, x\rangle - \delta_K(x)\} = \sup_{x\in K}\langle u, x\rangle = \sigma_K(u).$$

The supremum is a maximum because K is compact. Applying the Lecture 4 Fenchel duality formula to $h = \delta_K$, the dual of

$$\inf_x\{f(x) + \delta_K(x)\}$$

is

$$\sup_{u\in E^*}\{-f^*(u)-\delta_K^*(-u)\}=\sup_{u\in E^*}\{-f^*(u)-\sigma_K(-u)\}.$$

For any x and u, Fenchel–Young gives

$$f(x)+f^*(u)\geq\langle u,x\rangle.$$

If $x\notin K$, then $\delta_K(x)=+\infty$, so the gap is nonnegative. If $x\in K$, then

$$\sigma_K(-u)\geq\langle -u,x\rangle=-\langle u,x\rangle.$$

Thus

$$\mathcal{G}(x,u)=f(x)+\delta_K(x)+f^*(u)+\sigma_K(-u)\geq 0.$$

(b) The subdifferential of the support function is the exposed face:

$$\partial\sigma_K(w)=\arg\max_{z\in K}\langle w,z\rangle.$$

Indeed, if z maximizes $\langle w,\cdot\rangle$ over K, then for every w',

$$\sigma_K(w')\geq\langle w',z\rangle=\langle w,z\rangle+\langle w'-w,z\rangle=\sigma_K(w)+\langle w'-w,z\rangle,$$

so $z\in\partial\sigma_K(w)$. Conversely, suppose $q\in\partial\sigma_K(w)$. Then

$$\sigma_K(w')\geq\sigma_K(w)+\langle w'-w,q\rangle\qquad\forall w'.$$

Taking $w'=0$ gives $\langle w,q\rangle\geq\sigma_K(w)$. Also, applying the subgradient inequality to $w+\tau a$ and using subadditivity of σ_K,

$$\sigma_K(w)+\tau\sigma_K(a)\geq\sigma_K(w+\tau a)\geq\sigma_K(w)+\tau\langle a,q\rangle\qquad\forall a,\ \tau>0.$$

Hence $\langle a,q\rangle\leq\sigma_K(a)$ for all a. By finite-dimensional separation, this implies $q\in K$. Therefore $\langle w,q\rangle\leq\sigma_K(w)$, and together with the inequality above we get $\langle w,q\rangle=\sigma_K(w)$. Thus q maximizes $\langle w,\cdot\rangle$ over K. Taking $w=-u$ gives

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

Now set $u=\nabla f(x)$ and choose $s\in\arg\min_{z\in K}\langle u,z\rangle$. Since f is differentiable, Fenchel–Young is tight at $u=\nabla f(x)$:

$$f(x)+f^*(u)=\langle u,x\rangle.$$

Also $\sigma_K(-u)=\langle -u,s\rangle=-\langle u,s\rangle$. Therefore

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

(c) Let $u_t=\nabla f(x_t)$. From part (b),

$$s_t\in\partial\sigma_K(-u_t).$$

For the composition $u\mapsto\sigma_K(-u)$,

$$\partial(\sigma_K\circ(-I))(u_t)=-\partial\sigma_K(-u_t),$$

so $-s_t \in \partial(\sigma_K \circ (-I))(u_t)$. Since f^* is differentiable at u_t and $\nabla f^*(u_t) = x_t$,

$$x_t - s_t \in \nabla f^*(u_t) + \partial(\sigma_K \circ (-I))(u_t) = \partial\Psi(u_t).$$

The dual mirror-descent subproblem is unconstrained:

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

Using

$$D_{f^*}(u, u_t) = f^*(u) - f^*(u_t) - \left\langle u - u_t, \nabla f^*(u_t) \right\rangle,$$

this means u_{t+1} minimizes

$$\gamma_t \left\langle u - u_t, x_t - s_t \right\rangle + f^*(u) - f^*(u_t) - \left\langle u - u_t, \nabla f^*(u_t) \right\rangle$$

over $u \in E^*$. Its first-order condition is

$$0 = \gamma_t(x_t - s_t) + \nabla f^*(u_{t+1}) - \nabla f^*(u_t).$$

Since $\nabla f^*(u_t) = x_t$,

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

Since $\nabla f^*(u_{t+1}) = x_{t+1}$, we have $x_{t+1} \in \partial f^*(u_{t+1})$. By conjugacy, $u_{t+1} \in \partial f(x_{t+1})$. Since f is differentiable, $\partial f(x_{t+1}) = \{\nabla f(x_{t+1})\}$. Therefore

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

Thus the dual mirror-descent update is exactly the Frank–Wolfe update written in dual coordinates.

(d) We prove the formula by induction. For $T = 1$, $\gamma_0 = 1$, so $x_1 = s_0$, and

$$s_0 = \frac{2(0+1)}{1 \cdot 2} s_0.$$

Assume

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

Using $\gamma_T = 2/(T+2)$,

$$x_{T+1} = \left(1 - \frac{2}{T+2}\right) x_T + \frac{2}{T+2} s_T = \frac{T}{T+2} x_T + \frac{2}{T+2} s_T.$$

Substituting the induction hypothesis,

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

This is exactly

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

Therefore, for every $T \geq 1$,

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