上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

上海交通大学

约翰·霍普克罗夫特计算机科学中心

John Hopcroft Center for Computer Science

Online Mirror Descent

Shuai Li

John Hopcroft Center, Shanghai Jiao Tong University

https://shuaili8.github.io

Subgradients are not informative

- A subgradient does not always point in a direction where function decreases

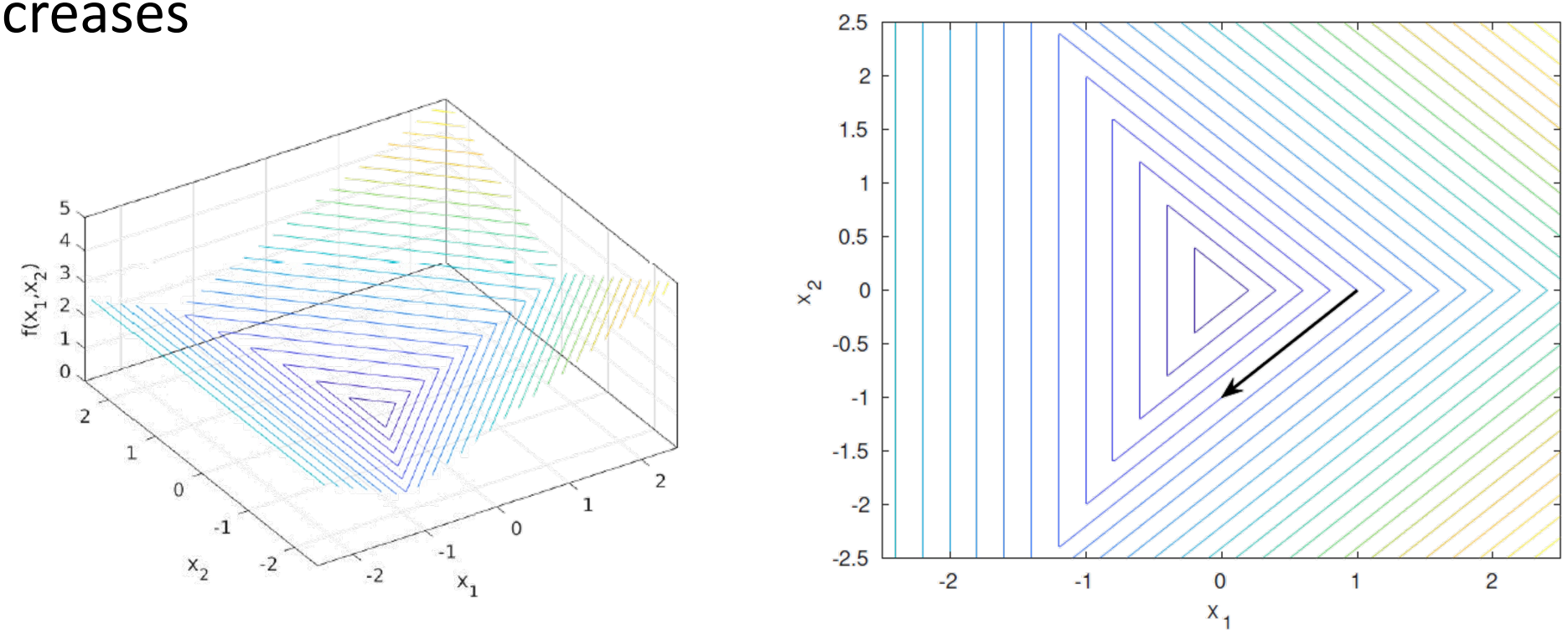


Figure 6.1: 3D plot (left) and level sets (right) of $f(\boldsymbol{x}) = \max[-x_1, x_1 - x_2, x_1 + x_2]$. A negative subgradient is indicated by the black arrow.

Subgradients are not informative 2

- A subgradient does not always point in a direction where function decreases

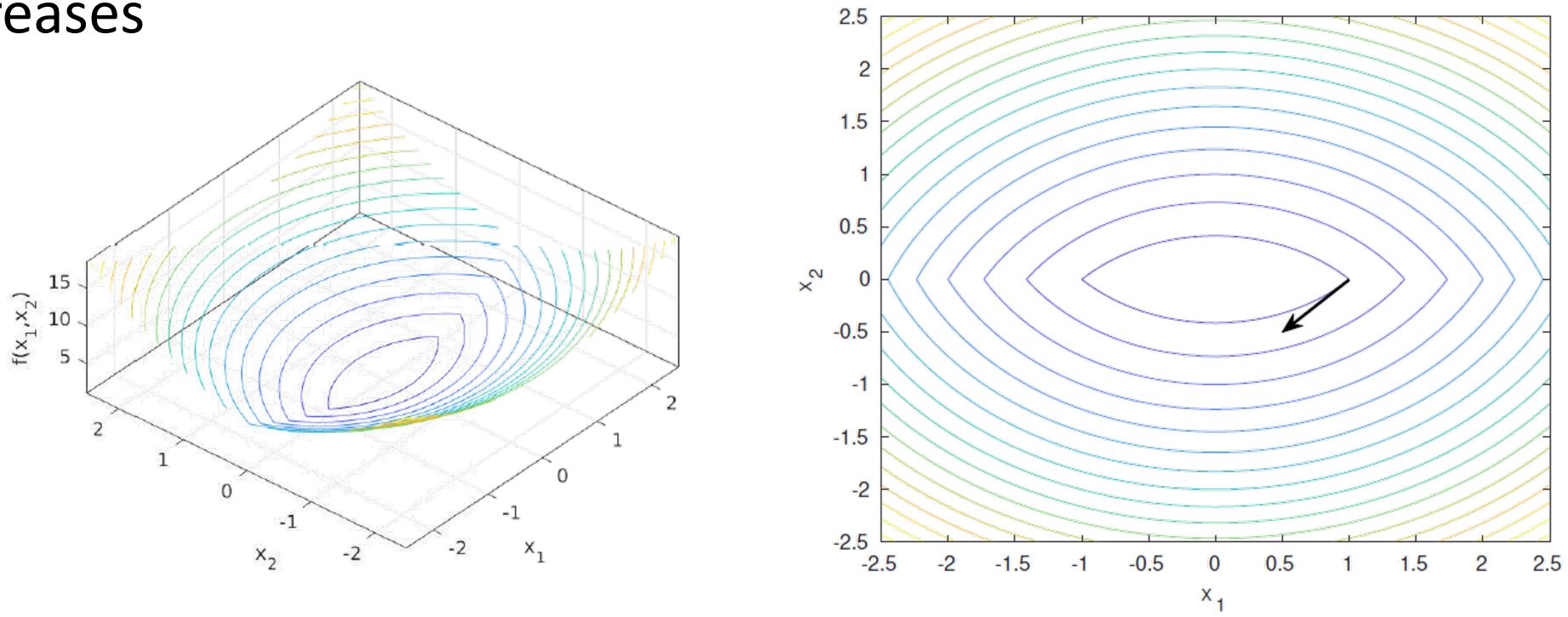


Figure 6.2: 3D plot (left) and level sets (right) of $f(\boldsymbol{x}) = \max[x_1^2 + (x_2+1)^2, x_1^2 + (x_2-1)^2]$. A negative subgradient is indicated by the black arrow.

Reinterpreting OSD

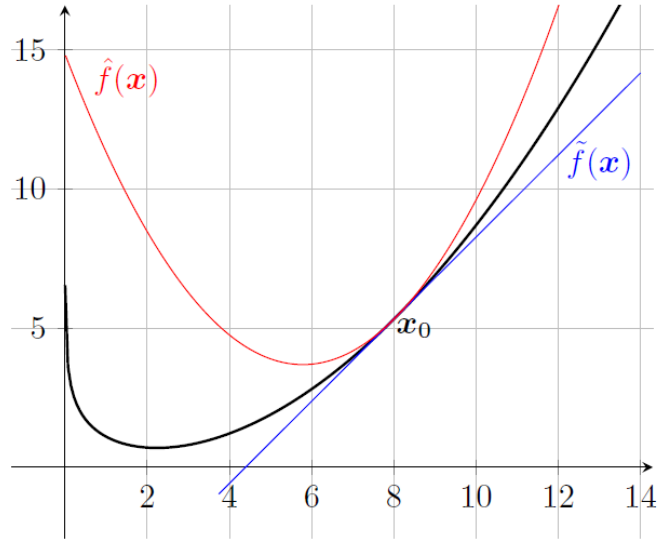
- $f(x) \geq \tilde{f}(x) \coloneqq f(x_0) + \langle g_t, x - x_0 \rangle, \forall x \in V$

  A linear lower bound to a function f

  But its minimum could be $-\infty$ over unbounded domains

- So we constrain it to a neighborhood

$$x_{t+1} = \arg\min_{x \in V} f(x_t) + \langle g_t, x - x_t \rangle$$
$$s.t. \|x - x_t\|_2^2 \leq \varepsilon$$



- Equivalently for some $\eta > 0$

$$\arg\min_{x \in V} \hat{f}(x) \coloneqq f(x_t) + \langle g_t, x - x_t \rangle + \frac{1}{2\eta} \|x - x_t\|_2^2$$
$$= \arg\min_{x \in V} \|x_t - \eta g_t - x\|_2^2$$
$$= \Pi_V(x_t - \eta g_t)$$

This is just OSD!

Bregman Divergence

- Definition 6.2. V convex. $f: V \subseteq \mathbb{R}^d \to \mathbb{R}$ is strictly convex if
  $f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y), \forall x \neq y \in V, \lambda \in (0,1)$
- Definition 6.3. $\psi: X \to \mathbb{R}$ is strictly convex and continuously differentiable on $\operatorname{int} X$. The Bregman divergence w.r.t. ψ is

$$B_\psi: X \times \operatorname{int} X \to \mathbb{R}$$
$$B_\psi(x; y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y \rangle$$

- Remark
  - $B_\psi(x; y) \geq 0$
  - $B_\psi(x; y) = 0$ if $x = y$
  - B_ψ is not symmetric
  - If ψ is μ-strongly convex w.r.t. $\|\cdot\|$, then $B_\psi(x; y) \geq \frac{\mu}{2}\|x - y\|^2$

Bregman Divergence: Properties

1-strongly convex w.r.t. $\|\cdot\|_2$

- Example 6.4. If $\psi(x) = \frac{1}{2}\|x\|_2^2$, then

$$B_\psi(x;y) = \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|y\|_2^2 - \langle y, x-y\rangle = \frac{1}{2}\|x-y\|_2^2$$

- Example 6.5. If $\psi(x) = \sum_{i=1}^d x_i \ln x_i$ and $X = \left\{x \in \mathbb{R}_+^d : \|x\|_1 = 1\right\}$, then $B_\psi(x;y) = \sum_{i=1}^d x_i \ln \frac{x_i}{y_i} = D_{\mathrm{KL}}(x \parallel y)$

(Lemma 6.15) 1-strongly convex w.r.t. $\|\cdot\|_1$

- Lemma 6.6. $\forall x, y \in \operatorname{int} X, z \in X$,

$$B_\psi(z;x) + B_\psi(x;y) - B_\psi(z;y) = \langle \nabla\psi(y) - \nabla\psi(x), z - x\rangle$$

Law of cosines

Online Mirror Descent

- Require: Closed convex set $\emptyset \neq V \subseteq X \subseteq \mathbb{R}^d$.
  ψ strictly convex and continuously differentiable on int X,
  $x_1 \in V$ such that ψ is differentiable, $\eta_1, \dots, \eta_T > 0$
- For $t = 1{:}T$ do
  - Output x_t
  - Receive $\ell_t \colon \mathbb{R}^d \to (-\infty, +\infty]$ and pay $\ell_t(x_t)$
  - Set $g_t \in \partial \ell_t(x_t)$
  - $x_{t+1} = \arg\min_{x \in V} \langle g_t, x \rangle + \frac{1}{\eta_t} B_\psi(x; x_t)$

OSD:
$\arg\min_{x \in V} \hat{f}(x) \coloneqq f(x_t) + \langle g_t, x - x_t \rangle + \frac{1}{2\eta} \|x - x_t\|_2^2$

x_{t+1} might be on the boundary of V. need $\lim_{x \to \partial X} \|\nabla \psi(x)\|_2 = +\infty$ or $V \subseteq \text{int } X$

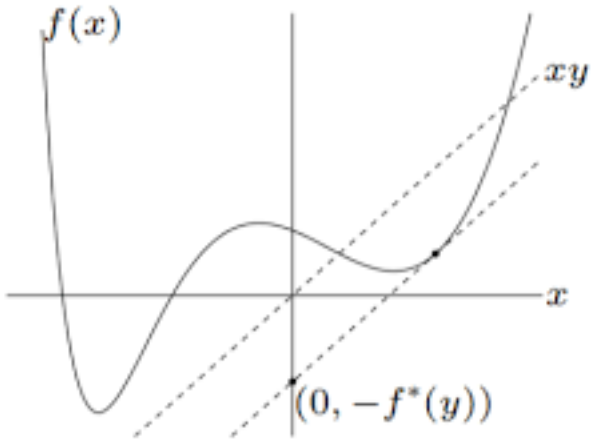
- Theorem 6.8. ψ is λ-strongly convex w.r.t. $\|\cdot\|$. $\max_{x,y \in V} B_\psi(x; y) \leq D^2$. $\|g_t\|_* \leq L$. $\eta_t \equiv \frac{D}{L\sqrt{T/\lambda}}$. Then $\forall u \in V$, $\sum_{t=1}^T (\ell_t(x_t) - \ell_t(u)) \leq DL\sqrt{T/\lambda}$.

dual norm of $\|\cdot\|$

could choose adaptive $\eta_t = \frac{D}{L\sqrt{t/\lambda}}$

$$f^*(y) = \sup_{x \in \mathbf{dom}\, f} (y^T x - f(x))$$

Fenchel Conjugate



- Definition 5.5. The Fenchel conjugate of $f: \mathbb{R}^d \to [-\infty, +\infty]$ is $f^*: \mathbb{R}^d \to [-\infty, +\infty]$ defined as
$$f^*(\theta) = \sup_{x \in \mathbb{R}^d} \langle \theta, x \rangle - f(x)$$
- Fenchel-Young's inequality: $\langle \theta, x \rangle \leq f(x) + f^*(\theta), \forall x, \theta \in \mathbb{R}^d$
- Theorem 5.6. If f is convex and closed, then $f^{**} = f$
- Theorem 5.7. $f: \mathbb{R}^d \to [-\infty, +\infty]$ is convex and proper. TFAE
  - $\theta \in \partial f(x)$
  - $\langle \theta, y \rangle - f(y)$ achieves its supremum at $y = x$
  - $f(x) + f^*(\theta) = \langle \theta, x \rangle$
  - (if f is closed) $x \in \partial f^*(\theta)$

"$(\partial f)^{-1} = \partial f^*$"

Fenchel Conjugate: Examples

- Example 5.11. The Fenchel conjugate of $f(x) = \frac{1}{2}\|x\|^2$ is

$$f^*(\theta) = \frac{1}{2}\|\theta\|_*^2$$

- Lemma 5.12. The Fenchel conjugate of $g(x) = af(x) + b$ is

$$g^*(\theta) = af^*(\theta/a) - b$$

The “Mirror” Interpretation

- Theorem 6.11. $x_{t+1} = \arg\min_{x \in V} \langle g_t, x \rangle + \frac{1}{\eta_t} B_\psi(x; x_t)$
$= \nabla\psi_V^*(\nabla\psi(x_t) - \eta_t g_t)$

  where $\psi_V := \psi + i_V$

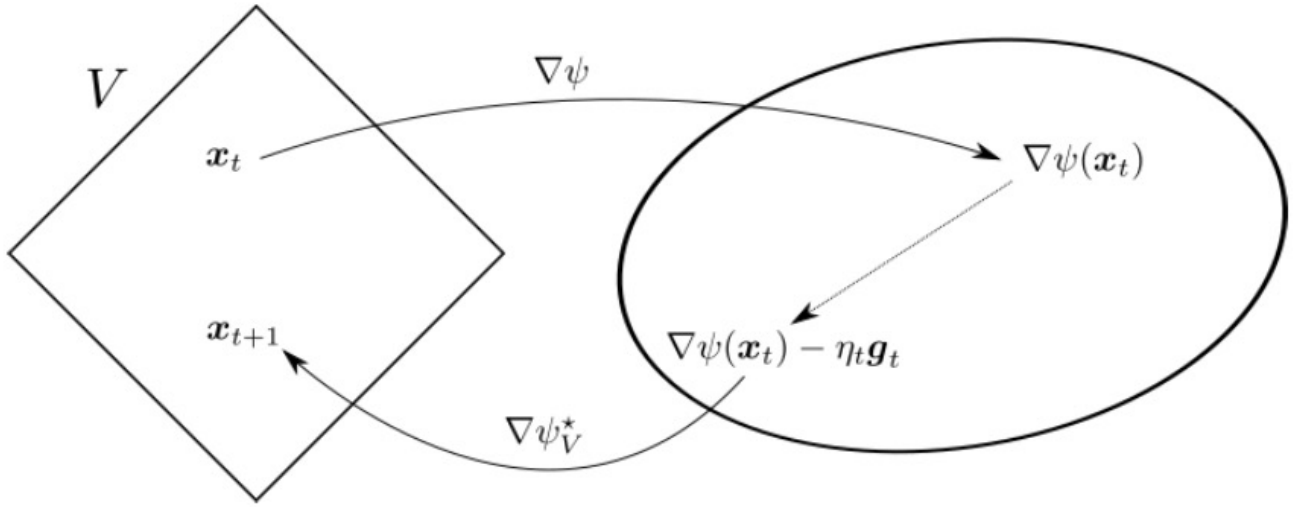


Figure 6.4: OMD update in terms of duality mappings.

Another Way to Write OMD Update

- Theorem 6.13. Under certain conditions,
  $x_{t+1} = \arg\min_{x \in V} \langle g_t, x \rangle + \frac{1}{\eta_t} B_\psi(x; x_t)$ is equivalent to

$$\tilde{x}_{t+1} = \arg\min_{x} \langle g_t, x \rangle + \frac{1}{\eta_t} B_\psi(x; x_t)$$

$$x_{t+1} = \arg\min_{x \in V} B_\psi(x; \tilde{x}_{t+1})$$

- The advantage of this update is that sometimes it gives two easier problems to solve rather than a single difficult one

Application: Learning w/ Expert Advice

- d experts to give advice on each round
- Need to decide which expert to follow on each round
- Objective: Minimize the losses we make compared to cumulative losses of the best expert
- Deterministic algorithms don't work
  - Two experts
  - The adversary always give the selected expert loss 1
- $\text{Regret}_T(e_i) = \sum_{t=1}^{T} \langle g_t, x_t \rangle - \sum_{t=1}^{T} \langle g_t, e_i \rangle, i = 1, \dots, d$
  - $x_t \in V = \{e_i : i = 1, \dots, d\}$
  - $0 \leq g_{t,i} \leq 1$

Nonconvex

Application: Learning w/ Expert Advice 2

- Randomization!
- $\text{Regret}'_T(u) = \sum_{t=1}^{T} \langle g_t, x_t \rangle - \sum_{t=1}^{T} \langle g_t, u \rangle \, , \forall u \in V'$
  - $x_t \in V' = \{x \in \mathbb{R}^d : x_i \geq 0, \|x\|_1 = 1\}$
  - $0 \leq g_{t,i} \leq 1$
- On each round, select expert i_t w/ prob. $x_{t,i}$
  - $\mathbb{E}[g_{t,i_t}] = \langle g_t, x_t \rangle$
- $\mathbb{E}[\text{Regret}_T(e_i)] = \mathbb{E}[\text{Regret}'_T(e_i)] = \mathbb{E}[\sum_{t=1}^{T} \langle g_t, x_t \rangle - \sum_{t=1}^{T} \langle g_t, e_i \rangle]$
  - Can minimize in expectation the non-convex regret w/ a randomized OCO algorithm

Convex !

Application: Learning w/ Expert Advice 3

- Require: $x_1 \in \{x \in \mathbb{R}^d : x_i \geq 0, \|x\|_1 = 1\} =: X, \eta > 0$
- For $t = 1:T$ do
  - Draw i_t according to $\mathbb{P}[i_t = i] = x_{t,i}$
  - Select expert i_t
  - Observe all experts' losses g_t and pay the loss g_{t,i_t}
  - Update $x_{t+1,j} \propto x_{t,j} \exp(-\eta g_{t,j})$ — Exponential gradient
- $\psi(x) = \sum_{i=1}^d x_i \ln x_i$ on X w/ $0 \ln 0 = 0$ and $V = X$ — 1-strongly convex w.r.t. $\|\cdot\|_1$ on X
- $\psi_V^*(\theta) = \sup_{x \in V} \langle \theta, x \rangle - \sum_{i=1}^d x_i \ln x_i = \ln\left(\sum_{i=1}^d \exp(\theta_i)\right)$
- $\Rightarrow (\nabla \psi_V^*)_j(\theta) = \frac{\exp(\theta_j)}{\sum_{i=1}^d \exp(\theta_i)}$
- $\Rightarrow x_{t+1,j} \propto x_{t,j} \exp(-\eta g_{t,j})$

Application: Learning w/ Expert Advice 4

$x_1 = \left(\frac{1}{d}, \dots, \frac{1}{d}\right)$

- $\Rightarrow \sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) \le \frac{\ln d}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|g_t\|_\infty^2 = O\left(L_\infty\sqrt{T \ln d}\right)$

$\eta = \sqrt{\frac{2 \ln d}{L_\infty^2 T}}$

- But couldn't use time-varying η!
  - Need to bound $\max_{1 \le t \le T} B_\psi(u; x_t)$
  - But the KL-divergence $B_\psi(u; x) = \sum_{i=1}^{d} u_i \ln \frac{u_i}{x_i}$ is unbounded
  - FTRL could overcome this issue using a time-varying regularizer
- OSD:

$$\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) \le \frac{\|u - x_1\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|g_t\|_2^2$$

$$\le \frac{2}{\eta} + \frac{\eta}{2} T d L_\infty^2 = O(L_\infty\sqrt{Td})$$

Application: Learning w/ Expert Advice 5

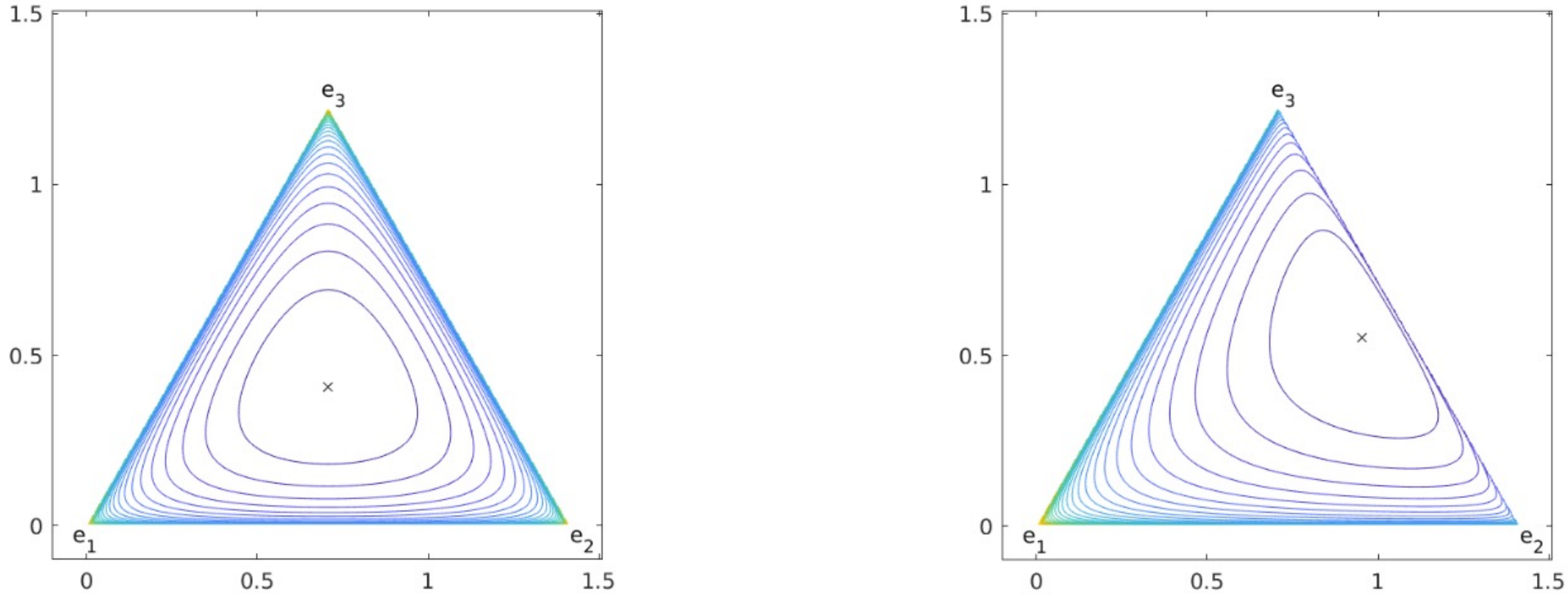


Figure 10.1: Contour plots of the KL divergence in 3-dimensions when $\boldsymbol{x}_t = [1/3, 1/3, 1/3]$ (left) and when $\boldsymbol{x}_t = [0.1, 0.45, 0.45]$ (right).

Follow-the-Regularized-Leader

Follow-the-Regularized-Leader

- $\emptyset \neq V \subseteq \mathbb{R}^d, \psi_1, \dots, \psi_T : \mathbb{R}^d \to (-\infty, +\infty]$
- For $t = 1:T$ do
  - Output $x_t \in \arg\min_{x \in V} \psi_t(x) + \sum_{i=1}^{t-1} \ell_i(x) =: F_t(x)$
  - Receive $\ell_t : \mathbb{R}^d \to (-\infty, +\infty]$ and pay $\ell_t(x_t)$

- OMD only remembers x_t, not the iterate before projection
- FTRL keeps in memory entire history, thus can recover every iterate before projection

⇒ FTRL is more computationally and memory expensive?
Generally YES!
But in some cases ≈ OMD and more informative

- Theorem 7.9. $\emptyset \neq V \subseteq \mathbb{R}^d$ closed convex. $\psi : V \to \mathbb{R}$ is μ-strongly convex w.r.t. $\|\cdot\|$. $\psi_t(x) := \frac{1}{\eta_{t-1}} \left(\psi(x) - \min_z \psi(z) \right)$. ℓ_t are L-Lipschitz continuous. $\eta_{t-1} = \frac{\alpha \sqrt{\mu}}{L \sqrt{t}}$. Then

$$\sum_{t=1}^{T} (\ell_t(x_t) - \ell_t(u)) \leq \left(\frac{\psi(u) - \min_x \psi(x)}{\alpha} + \alpha \right) \frac{L\sqrt{T}}{\sqrt{\mu}}$$

Key Lemmas

- Lemma 7.1.

$$\sum_{t=1}^{T}(\ell_t(x_t)-\ell_t(u)) = \psi_{T+1}(u)-\min_x \psi_1(x) + \sum_{t=1}^{T}\left(F_t(x_t)-F_{t+1}(x_{t+1})+\ell_t(x_t)\right) + F_{T+1}(x_{T+1})-F_{T+1}(u)$$

$F_1(x_1)$

$\psi_t \equiv \psi \Rightarrow = F_{t+1}(x_t)-F_{t+1}(x_{t+1})$ is small if $x_t \approx x_{t+1}$

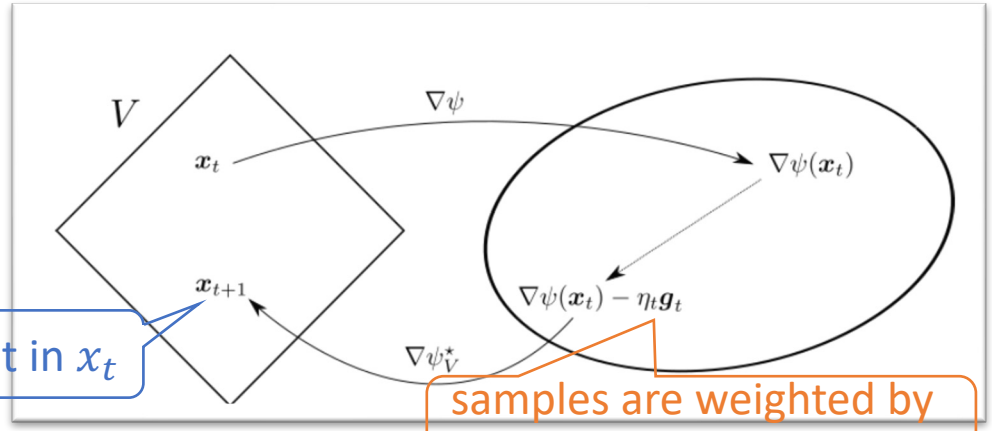
- Lemma 7.6. $f:\mathbb{R}^d \to (-\infty,+\infty]$ is μ-strongly convex w.r.t. $\|\cdot\|$. $g \in \partial f(x), g' \in \partial f(y)$. Then

$$f(x)-f(y) \le \langle g', x-y\rangle + \frac{1}{2\mu}\|g-g'\|_*^2$$

- Lemma 7.8. $\forall g_t \in \partial \ell_t(x_t)$, ψ_t is μ_t-strongly convex. Then

$$F_t(x_t)-F_{t+1}(x_{t+1})+\ell_t(x_t) \le \frac{\|g_t\|_*^2}{2\mu_t} + \psi_t(x_{t+1})-\psi_{t+1}(x_{t+1})$$

FTRL w/ Linearized Losses

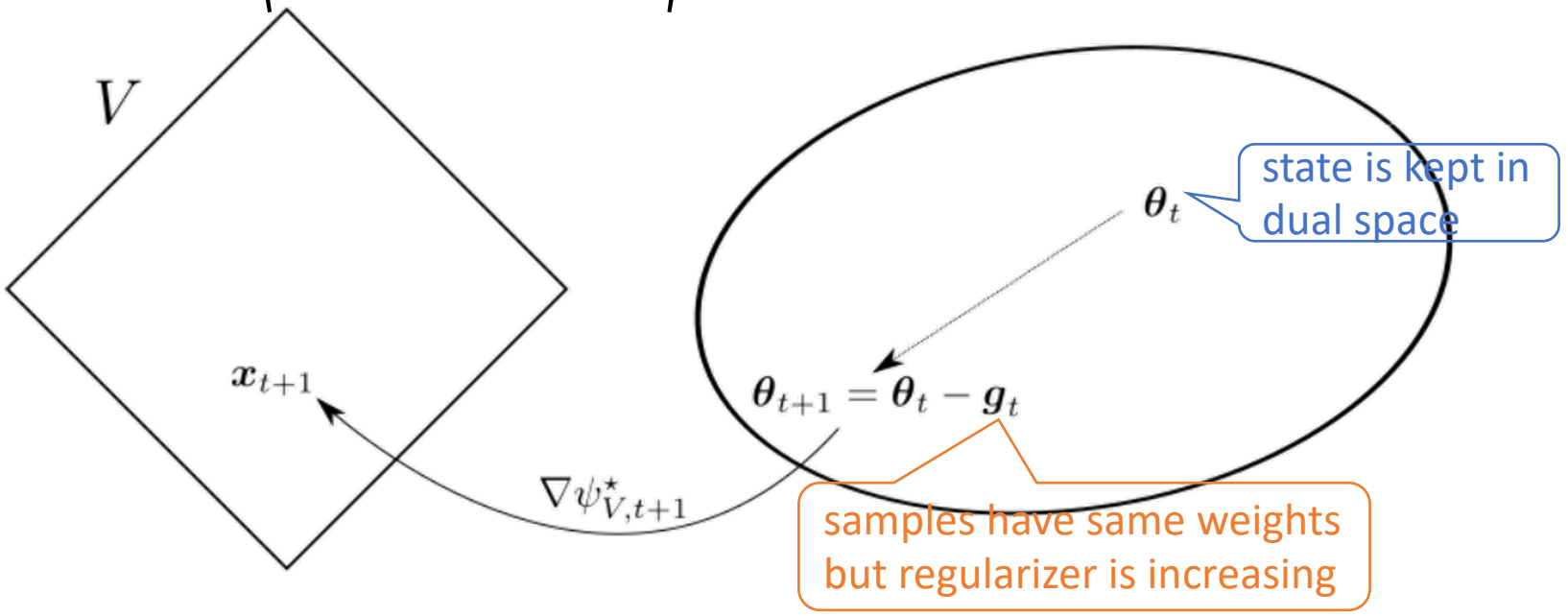


state is kept in x_t

samples are weighted by decreasing learning rates

- $x_{t+1} \in \arg\min_{x \in V} \psi_{t+1}(x) + \sum_{i=1}^{t} \langle g_i, x \rangle$

$$= \arg\max_x \left\langle -\sum_{i=1}^{t} g_i, x \right\rangle - \psi_{t+1}(x) - i_V(x) = \nabla\psi^*_{V,t+1}\left(-\sum_{i=1}^{t} g_i\right)$$



state is kept in dual space

samples have same weights but regularizer is increasing

Figure 7.1: Dual mapping for FTRL with linear losses.

FTRL w/ Linearized Losses 2

- $V = X = \operatorname{dom} \psi$
- OMD:
  - $x_{t+1} = \arg\min_x \langle \eta g_t, x \rangle + B_\psi(x; x_t)$
  - Assume $x_{t+1} \in \operatorname{int} \operatorname{dom} \psi$, $\forall t$
  - $\Rightarrow \eta g_t + \nabla\psi(x_{t+1}) - \nabla\psi(x_t) = 0$
  - Assume $x_1 = \arg\min_x \psi(x)$
  - $\Rightarrow \nabla\psi(x_{t+1}) = -\eta \sum_{i=1}^t g_i$

- FTRL w/ $\psi_t = \frac{1}{\eta}\psi$
  - $x_{t+1} = \arg\min_x \frac{1}{\eta}\psi(x) + \sum_{i=1}^t \langle g_i, x \rangle$
  - Assume $x_{t+1} \in \operatorname{int} \operatorname{dom} \psi$, $\forall t$
  - $\Rightarrow \nabla\psi(x_{t+1}) = -\eta \sum_{i=1}^t g_i$
- The regularizer of FTRL can vary over time, not only a scaled version of a fixed regularizer

FTRL: Exponential Gradient w.o. Knowing T

- $V = \{x \in \mathbb{R}^d : x_i \geq 0, \|x\|_1 = 1\}$
- $\ell_t : V \to \mathbb{R}$ convex, L_∞-Lipschitz w.r.t. $\|\cdot\|_\infty$
- $\psi : V \to \mathbb{R}$, $\psi(x) = \sum_{i=1}^d x_i \ln x_i$, $0 \ln 0 = 0$
- $\psi_t(x) = \alpha L_\infty \sqrt{t} \psi(x)$ is $\alpha L_\infty \sqrt{t}$-strongly convex w.r.t. $\|\cdot\|_1$
- $x_t = \nabla \psi_{V,t}^* \left(-\sum_{i=1}^{t-1} g_i \right)$
- $x_{t,j} \propto \exp\left(-\frac{1}{\alpha L_\infty \sqrt{t}} \sum_{i=1}^{t-1} g_{i,j} \right)$
- $\sum_{t=1}^T (\ell_t(x_t) - \ell_t(u)) \leq \left(\alpha \left(\sum_{i=1}^d u_i \ln u_i + \ln d \right) + \frac{1}{\alpha} \right) L_\infty \sqrt{T}$
  $\leq \left(\alpha \ln d + \frac{1}{\alpha} \right) L_\infty \sqrt{T} \leq 2 L_\infty \sqrt{T \ln d}$

$\alpha = \sqrt{\ln d}$

The Fenchel conjugate of $g(x) = af(x) + b$ is $g^*(\theta) = af^*(\theta/a) - b$

$(\nabla \psi_V^*)_j(\theta) = \frac{\exp(\theta_j)}{\sum_{i=1}^d \exp(\theta_i)}$

Composite Losses

- $\ell_t(x) = \tilde{\ell}_t(x) + \lambda\|x\|_1$
- Using linearization
  - might just take the subgradient of ℓ_t
  - but will lose the ability of $\|\cdot\|_1$-regularization to produce sparse solutions
- Better way:
  - Run FTRL with $\psi_t(x) = L\sqrt{t}\left(\psi(x) - \min_y \psi(y)\right) + \lambda t\|x\|_1$ and loss $\tilde{\ell}_t$
  - ψ is 1-strongly convex and $\tilde{\ell}_t$ is L-Lipschitz
  - $\Rightarrow \sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) \le L\sqrt{T}\left(\psi(u) - \min_x \psi(x) + 1\right)$
- Example 7.20. $\psi(x) = \frac{1}{2}\|x\|_2^2$, $x_t = \arg\min_x \frac{L\sqrt{t}}{2}\|x\|_2^2 + \lambda t\|x\|_1 + \sum_{i=1}^{t-1}\langle g_i, x\rangle$
  - $\theta_t = \sum_{i=1}^{t-1} g_i$, $x_{t,i} = -\frac{\operatorname{sign}(\theta_{t,i})\max(|\theta_{t,i}| - \lambda t, 0)}{L\sqrt{t}}$

produce sparse solutions

Summary

- Subgradients are not informative
- Reinterpreting OSD, Bregman Divergence
- Online Mirror Descent
  - Fenchel Conjugate and the “Mirror” Interpretation
  - Application: Learning w/ Expert Advice
- Follow-the-Regularized-Leader
  - FTRL w/ Linearized Losses v.s. OMD
  - Exponential Gradient: no need to know T
  - Composite Losses

Shuai Li

https://shuaili8.github.io

Questions?